Journal of Machine Learning Research 24 (2023) 1-21 Submitted 1/23; Published 5/23

From Classification Accuracy to Proper Scoring Rules: Elicitability of Probabilistic Top List Predictions

Johannes Resin JOHANNES.RESIN@AWI.UNI-HEIDELBERG.DE
Alfred Weber Institute for Economics
Heidelberg University
Bergheimer Str. 58, 69115 Heidelberg, Germany

Computational Statistics Group
Heidelberg Institute for Theoretical Studies

Editor: Sayan Mukherjee

Abstract

In the face of uncertainty, the need for probabilistic assessments has long been recognized in the literature on forecasting. In classification, however, comparative evaluation of classifiers often focuses on predictions specifying a single class through the use of simple accuracy measures, which disregard any probabilistic uncertainty quantification. I propose probabilistic top lists as a novel type of prediction in classification, which bridges the gap between single-class predictions and predictive distributions. The probabilistic top list functional is elicitable through the use of strictly consistent evaluation metrics. The proposed evaluation metrics are based on symmetric proper scoring rules and admit comparison of various types of predictions ranging from single-class point predictions to fully specified predictive distributions. The Brier score yields a metric that is particularly well suited for this kind of comparison.

Keywords: Brier score, consistent scoring functions, evaluation metrics, probabilistic multi-class classification, symmetric proper scoring rules

1. Introduction

In the face of uncertainty, predictions ought to quantify their level of confidence (Gneiting and Katzfuss, 2014). This idea has been recognized for decades in the literature on weather forecasting (Brier, 1950; Murphy, 1977) and probabilistic forecasting (Dawid, 1984; Gneiting and Raftery, 2007). Ideally, a prediction specifies a probability distribution over potential outcomes. Such predictions are evaluated and compared by means of proper scoring rules, which quantify their value in a way that rewards truthful prediction (Gneiting and Raftery, 2007). In statistical classification and machine learning, the need for reliable uncertainty quantification has not gone unnoticed, as exemplified by the growing interest in the calibration of probabilistic classifiers (Guo et al., 2017; Vaicenavicius et al., 2019). However, classifier evaluation often focuses on the most likely class (i.e., the mode of the predictive distribution) through the use of classification accuracy and related metrics derived from the confusion matrix (Tharwat, 2020; Hui and Belkin, 2021).

Probabilistic classification separates the prediction task from decision making. This enables informed decisions that account for diverse cost-loss structures, for which decisions

based simply on the most likely class may lead to adverse outcomes (Elkan, 2001; Gneiting, 2017). Probabilistic classification is a viable alternative to classification with reject option, where classifiers may refuse to predict a class if their confidence in a single class is not sufficient (Herbei and Wegkamp, 2006; Ni et al., 2019).

In this paper, I propose *probabilistic top lists* as a way of producing probabilistic classifications in settings where specifying entire predictive distributions may be undesirable, impractical, or even impossible. While multi-label classification serves as a key example of such a setting, the theory presented here applies to classification in general. I envision the probabilistic top list approach to be particularly useful in settings eluding traditional probabilistic forecasting, where the specification of probability distributions on the full set of classes is hindered by a large number of classes and missing (total) order. Consistent evaluation is achieved through the use of proper scoring rules.

Whereas in traditional classification an instance is associated with a single class (e.g., *cat* or *dog*), multi-label classification problems (as reviewed by Tsoumakas and Katakis, 2007; Zhang and Zhou, 2014; Tarekegn et al., 2021) admit multiple labels for an instance (e.g., *cat* or *dog* or *cat and dog*).[1] Applications of multi-label classification include text categorization (Zhang and Zhou, 2006), image recognition (Chen et al., 2019), and functional genomics (Barutcuoglu et al., 2006; Zhang and Zhou, 2006). Multi-label classification methods often output confidence scores for each label independently, and the final label set prediction is determined by a simple cut-off (Zhang and Zhou, 2014). As this approach does not account for label correlations, computing label set probabilities in a postprocessing step can improve predictions and probability estimates (Li et al., 2020) over simply multiplying probabilities to obtain label set probabilities. Probabilistic top lists offer a flexible approach to multi-label classification, which embraces the value of probabilistic information. In fact, the *BR-rerank* method introduced by Li et al. (2020) produces top list predictions. Yet, comparative performance evaluation focuses on (set) accuracy and the improper instance F1 score. This discrepancy has been a key motivation for this research.

In probabilistic forecasting, a scoring rule assigns a numerical score to a predictive distribution based on the true outcome (Gneiting and Raftery, 2007). It is proper if the expected score is optimized by the true distribution of the outcome of interest. Popular examples in classification are the Brier (or quadratic) score and the logarithmic (or cross-entropy) loss (Gneiting and Raftery, 2007; Hui and Belkin, 2021). When one is not interested in full predictive distributions, simple point predictions are frequently preferred. A meaningful point prediction admits interpretation in terms of a statistical functional (Gneiting, 2011). Point predictions are evaluated by means of consistent scoring or loss functions. Similar to proper scoring rules, a scoring function is consistent for a functional if the expected score is optimized by the true functional value of the underlying distribution. For example, accuracy (or, equivalently, misclassification or zero-one loss) is consistent for the mode in classification (Gneiting, 2017).

Probabilistic top lists bridge the gap between mode forecasts and full predictive distributions in classification. In this paper, I define a probabilistic top-k list as a collection of k classes deemed most likely together with confidence scores quantifying the predictive probability associated with each of the k classes. The key question tackled in this work is how

1. Multi-label classification is a special case of classification if classes are (re-)defined as subsets of labels.

to evaluate such top list predictions in a consistent manner. To this end, I propose what I call *padded symmetric scores*, which are based on proper symmetric scoring rules. I show that the proposed padded symmetric scores are consistent for the probabilistic top-k list functional. The padded symmetric score of a probabilistic top list prediction is obtained from a symmetric proper scoring rule by padding the top list to obtain a fully specified distribution. The padded distribution divides the probability mass not accounted for by the top list's confidence scores equally among the classes that are not included in the list. Padded symmetric scores exhibit an interesting property, which allows for balanced comparison of top lists of different length as well as single-class point predictions and predictive distributions. Notably, the expected score of a correctly specified top list only depends on the top list itself and is invariant to other aspects of the true distribution. Comparability of top lists of differing length is ensured as the expected score does not deteriorate upon increasing the length of the predicted top list. Nonetheless, if the scoring function is based on the Brier score, there is little incentive to provide unreasonably large top lists. In the case of a single-class prediction, the padded version of the Brier score reduces to twice the misclassification loss. Hence, the padded Brier score essentially generalizes classification accuracy.

The remainder of the paper proceeds as follows. Section 2 recalls the traditional multi-class classification problem with a focus on probabilistic classification and suitable evaluation metrics. A short introduction to the multi-label classification problem is also provided. Section 3 introduces probabilistic top lists, and related notation and terminology used throughout this work. Section 4 introduces some preliminary results on symmetric proper scoring rules and some results relating to the theory of majorization. These results are used in Section 5 to show that the padded symmetric scores yield consistent scoring functions for the top list functionals. Section 6 discusses the comparison of various types of predictions using the padded Brier and logarithmic scores. A theoretical argument as well as numerical examples illustrate that the padded Brier score is well suited for this task. Section 7 concludes the paper.

2. Statistical Classification

The top list functionals and the proposed scoring functions are motivated by multi-label classification, but they apply to other classification problems as well. Here, I give a short formal introduction to the general classification problem and related evaluation metrics from the perspective of probabilistic forecasting. In what follows, the symbol $\mathcal{L}$ refers to the law or distribution of a given random variable.

2.1 Traditional Multi-Class Classification

In the classical (multi-class) classification problem, one tries to predict the distinct class Y of an instance characterized by a vector of features $\mathbf{X}$. Formally, the outcome Y is a random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ taking values in the set of classes $\mathcal{Y}$ of cardinality $m \in \mathbb{N}$, and the feature vector $\mathbf{X}$ is a random vector taking values in some feature space $\mathcal{X} \subseteq \mathbb{R}^d$. Ideally, one learns the entire conditional distribution $p(\mathbf{X}) = \mathcal{L}(Y \mid \mathbf{X})$ of Y given $\mathbf{X}$ through a *probabilistic* classifier $c\colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ mapping the features of a given instance to a probability distribution from the set of probability distributions $\mathcal{P}(\mathcal{Y})$ on $\mathcal{Y}$. The set

$\mathcal{P}(\mathcal{Y})$ of probability distributions is typically identified with the probability simplex

$$\Delta_{m-1} = \{p \in [0,1]^m \mid p_1 + \cdots + p_m = 1\}$$

by (arbitrarily) labeling the classes as $1, \ldots, m$, and probability distributions are represented by vectors $p \in \Delta_{m-1}$, where the i-th entry p_i is the probability assigned to class i for $i = 1, \ldots, m$. To ease notation in what follows, vectors in Δ_{m-1} are indexed directly by the classes in $\mathcal{Y}$ without explicit mention of any (re-)labeling.

Proper scoring rules quantify the value of a probabilistic classification and facilitate the comparison of multiple probabilistic classifiers (Gneiting and Raftery, 2007). A *scoring rule* is a mapping $\mathrm{S}\colon \mathcal{P}(\mathcal{Y}) \times \mathcal{Y} \to \overline{\mathbb{R}}$, which assigns a, possibly infinite, score $\mathrm{S}(p, y)$ from the extended real numbers $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ to a predictive distribution p if the true class is y. Typically, scores are negatively oriented in that lower scores are preferred. A scoring rule S is called *proper* if the true distribution $p = \mathcal{L}(Y)$ of Y minimizes the expected score, i.e.,

$$\mathbb{E}[\mathrm{S}(p, Y)] \leq \mathbb{E}[\mathrm{S}(q, Y)] \quad \text{for } Y \sim p \text{ and } p, q \in \mathcal{P}(\mathcal{Y}). \tag{1}$$

It is *strictly proper* if inequality (1) is strict unless $p = q$. Prominent examples are the logarithmic score

$$\mathrm{S}_{\log}(p, y) = -\log p_y \tag{2}$$

and the Brier score

$$\mathrm{S}_{\mathrm{B}}(p, y) = (1 - p_y)^2 + \sum_{z \neq y} p_z^2 = 1 - 2p_y + \sum_{z \in \mathcal{Y}} p_z^2. \tag{3}$$

Frequently, current practice does not focus on learning the full conditional distribution but, rather, on simply predicting the most likely class, i.e., the mode of the conditional distribution $p(\mathbf{X})$. This practice is formalized by a *hard* classifier $c\colon \mathcal{X} \to \mathcal{Y}$ aspiring to satisfy the functional relationship $c(\mathbf{X}) \in \mathrm{Mode}(p(\mathbf{X}))$, where the *mode functional* is given by

$$\mathrm{Mode}(p) = \operatorname*{arg\,max}_{y \in \mathcal{Y}} p_y = \{z \in \mathcal{Y} \mid p_z = \max_{y \in \mathcal{Y}} p_y\} \tag{4}$$

for $p \in \Delta_{m-1}$. Other functionals may be learned as well. When it comes to point forecasts of real-valued outcomes, popular choices are the mean or a quantile, see for example Gneiting and Resin (2021). Formally, a *statistical functional* $\mathrm{T}\colon \mathcal{P}(\mathcal{Y}) \to 2^{\mathcal{T}}$ reduces probability measures to certain facets in some space $\mathcal{T}$. Note that the functional T maps a distribution to a subset in the power set $2^{\mathcal{T}}$ of $\mathcal{T}$ owing to the fact that the functional value may not be uniquely determined. For example, the mode (4) of a distribution is not unique if multiple classes are assigned the maximum probability. The probabilistic top lists introduced in Section 3 are a nonstandard example of a statistical functional, which lies at the heart of this work.

Similar to the evaluation of probabilistic classifiers through the use of proper scoring rules, predictions aimed at a statistical functional are evaluated by means of consistent scoring functions. Given a functional T, a *scoring function* is a mapping $\mathrm{S}\colon \mathcal{T} \times \mathcal{Y} \to \overline{\mathbb{R}}$, which assigns a score $\mathrm{S}(t, y)$ to a predicted facet t if the true class is y. A scoring function S

is *consistent* for the functional T if the expected score is minimized by any prediction that is related to the true distribution of Y by the functional, i.e.,

$$\mathbb{E}[\mathrm{S}(t, Y)] \leq \mathbb{E}[\mathrm{S}(s, Y)] \quad \text{for } Y \sim p, t \in \mathrm{T}(p), p \in \mathcal{P}(\mathcal{Y}), \text{ and } s \in \mathcal{T}. \tag{5}$$

It is *strictly consistent* for T if inequality (5) is strict unless $s \in \mathrm{T}(p)$. A functional T is called *elicitable* if a strictly consistent scoring function for T exists.[2] For example, the mode (4) is elicited by the zero-one scoring function or misclassification loss (Gneiting, 2017)

$$\mathrm{S}(x, y) = \mathbb{1}\{x \neq y\} = \begin{cases} 1, & \text{if } x \neq y, \\ 0, & \text{if } x = y, \end{cases},$$

which is simply a negatively oriented version of the ubiquitous classification accuracy. As discussed by Gneiting (2017) and references therein, decisions based on the mode are suboptimal if the losses invoked by different misclassifications are not uniform, which is frequently the case.

(Strictly) Proper scoring rules arise as a special case of (strictly) consistent scoring functions if T is the identity on $\mathcal{P}(\mathcal{Y})$. Furthermore, any consistent scoring function yields a proper scoring rule if predictive distributions are reduced by means of the respective functional first (Gneiting, 2011, Theorem 3). On the other hand, a point prediction $x \in \mathcal{Y}$ can be assessed by means of a scoring rule as the classes can be embedded in the probability simplex by identifying a class $y \in \mathcal{Y}$ with the point mass $\delta_y \in \mathcal{P}(\mathcal{Y})$ in y. For example, applying the Brier score to a class prediction in this way yields twice the misclassification loss, $\mathrm{S_B}(x, y) = \mathrm{S_B}(\delta_x, y) = 2 \cdot \mathbb{1}\{x \neq y\}$.

Naturally, the true conditional distributions are unknown in practice, and expected scores are estimated by the mean score attained across all instances available for evaluation purposes.

2.2 Multi-Label Classification

In multi-label classification problems, an instance may be assigned multiple (class) labels. Here, I frame this problem as a special case of multi-class classification instead of an entirely different problem.

Let L be the set of labels and $\mathcal{Y} \subseteq 2^L$ be the set of label sets, i.e., classes are subsets of labels. In this setting, it may be difficult to specify a sensible predictive distribution on $\mathcal{Y}$, even for moderately sized sets of labels L, since the number of classes may grow exponentially with the number of labels. Extant comparative evaluation practices in multi-label classification focus mainly on hard classifiers ignoring the need for uncertainty quantification through probabilistic assessments (e.g., Tsoumakas and Katakis, 2007; Zhang and Zhou, 2014; Li et al., 2020; Tarekegn et al., 2021) with the exception of Read et al. (2011), who also consider a sum of binary logarithmic losses to evaluate the confidence scores associated with individual labels.

2. The notion of elicitability used in this work is termed "selective elicitability" by Fissler et al. (2021). In contrast, they call a functional "exhaustively elicitable" if a strictly consistent scoring function $2^{\mathcal{T}} \times \mathcal{Y} \to \overline{\mathbb{R}}$ for set-valued predictions exists. The functionals discussed in this work are selectively elicitable, which precludes them from being exhaustively elicitable (Fissler et al., 2021, Theorem 3.9).

Classification accuracy is typically referred to as (sub-)set accuracy in multi-label classification. Other popular evaluation metrics typically quantify the overlap between the predicted label set and the true label set. For example, the comparative evaluation by Li et al. (2020) reports instance F1 scores in addition to set accuracy, where instance F1 of a single instance is defined as

$$\mathrm{S}_{\mathrm{F1}}(x, y) = \frac{2 \sum_{\ell \in L} \mathbb{1}\{\ell \in x\} \mathbb{1}\{\ell \in y\}}{\sum_{\ell \in L} \mathbb{1}\{\ell \in x\} + \sum_{\ell \in L} \mathbb{1}\{\ell \in y\}}.$$

(and the overall score is simply the average across all instances as usual). Note that this metric is positively oriented, i.e., higher instance F1 scores are preferred. Caution is advised as the instance F1 score is not consistent for the mode, as illustrated by the following example. Hence, evaluating the same predictions using set accuracy and instance F1 seems to be a questionable practice.

Example 1 *Let the label set $L = \{1, 2, 3, 4, 5\}$ consist of five labels and the set of classes $\mathcal{Y} = 2^L$ be the power set of the label set L. Consider the distribution $p \in \mathcal{P}(\mathcal{Y})$ that assigns all probability mass to four label sets as follows:*

$$p_{\{1,2\}} = 0.28, \quad p_{\{1,3\}} = 0.24, \quad p_{\{1,4\}} = 0.24, \quad p_{\{1,5\}} = 0.24.$$

Then the expected instance F1 score of the most likely label set $\{1, 2\}$,

$$\mathbb{E}[\mathrm{S}_{\mathrm{F1}}(\{1, 2\}, Y)] = 0.64,$$

given $Y \sim p$ is surpassed by predicting only the single label $\{1\}$,

$$\mathbb{E}[\mathrm{S}_{\mathrm{F1}}(\{1\}, Y)] = \tfrac{2}{3}.$$

3. Probabilistic Top Lists

In what follows, I develop a theory informing principled evaluation of top list predictions based on proper scoring rules. To this end, a concise mathematical definition of probabilistic top lists is fundamental.

Let $k \in \{0, \ldots, m\}$ be fixed. A *(probabilistic) top-k list* is a collection $t = (\widehat{Y}, \widehat{t})$ of a set $\widehat{Y} \subset \mathcal{Y}$ of $k = |\widehat{Y}|$ classes together with a vector $\widehat{t} = (\widehat{t}_y)_{y \in \widehat{Y}} \in [0, 1]^k$ of *confidence scores* (or predicted probabilities) indexed by the set $\widehat{Y}$ whose sum does not exceed one, i.e., $\sum_{y \in \widehat{Y}} \widehat{t}_y \leq 1$, and equals one if $k = m$. Let $\mathcal{T}_k$ denote the set of probabilistic top-k lists. On the one hand, the above definition includes the empty top-0 list $t_\emptyset = (\emptyset, ())$ for technical reasons. At the other extreme, top-m lists specify entire probability distributions on $\mathcal{Y}$, i.e., $\mathcal{T}_m \equiv \mathcal{P}(\mathcal{Y})$. The *proxy probability*

$$\pi(t) := \frac{1 - \sum_{y \in \widehat{Y}} \widehat{t}_y}{m - k}$$

associated with a top-k list $t = (\widehat{Y}, \widehat{t}) \in \mathcal{T}_k$ of size $k < m$ is the probability mass not accounted for by the top list t divided by the number of classes not listed. For a top-m

list $t \in \mathcal{T}_m$, the proxy probability $\pi(t) \equiv 0$ is defined to be zero. The *padded probability distribution* $\widetilde{t} = (\widetilde{t}_y)_{y\in\mathcal{Y}} \in \Delta_{m-1}$ associated with a probabilistic top-k list $t = (\widehat{Y}, \widehat{t}) \in \mathcal{T}_k$ assigns the proxy probability $\pi(t)$ to all classes not in $\widehat{Y}$, i.e.,

$$\widetilde{t}_y = \begin{cases} \widehat{t}_y, & \text{if } y \in \widehat{Y}, \\ \pi(t), & \text{if } y \notin \widehat{Y} \end{cases} \tag{6}$$

for $y \in \mathcal{Y}$.

A top-k list $t = (\widehat{Y}, \widehat{t})$ is *calibrated* relative to a distribution $p = (p_y)_{y\in\mathcal{Y}} \in \Delta_{m-1}$ if the confidence score $\widehat{t}_y$ of class y matches the true class probability p_y for all $y \in \widehat{Y}$. A top-k list $t = (\widehat{Y}, \widehat{t})$ is *true* relative to a distribution $p \in \mathcal{P}(\mathcal{Y})$ if it is calibrated relative to p and $\widehat{Y}$ consists of k most likely classes. There may be multiple true top-k lists for a given $k \in \mathbb{N}$ if the class probabilities are not pairwise distinct (i.e., if some classes have the same probability). References to the true distribution of the outcome Y are usually omitted in what follows. For example, a calibrated top list is understood to be calibrated relative to the distribution $\mathcal{L}(Y)$ of Y. The *(probabilistic) top-k list functional* $\mathrm{T}_k \colon \mathcal{P}(\mathcal{Y}) \to \mathcal{T}_k$ maps any probability distribution $p \in \mathcal{P}(\mathcal{Y})$ to the set

$$\mathrm{T}_k(p) = \left\{ (\widehat{Y}, (p_y)_{y\in\widehat{Y}}) \in \mathcal{T}_k \;\middle|\; \widehat{Y} \in \operatorname*{arg\,max}_{S\subset\mathcal{Y}:|S|=k} \sum_{y\in S} p_y \right\}$$

of top-k lists that are true relative to p. The top-m list functional T_m identifies $\mathcal{P}(\mathcal{Y})$ with $\mathcal{T}_m$. A top-k list $t \in \mathcal{T}_k$ is *valid* if it is true relative to some probability distribution, i.e., there exists a distribution $p \in \mathcal{P}(\mathcal{Y})$ such that $t \in \mathrm{T}_k(p)$. Equivalently, a top-k list $t = (\widehat{Y}, \widehat{t})$ is valid if the associated proxy probability does not exceed the least confidence score, i.e., $\min_{y\in\widehat{Y}} \widehat{t}_y \geq \pi(t)$. Hence, the proxy probability associated with a valid top-k list is simply the mean confidence score of the bottom $(m-k)$ classes. Let $\widetilde{\mathcal{T}}_k \subset \mathcal{T}_k$ denote the set of valid top-k lists. The following is a simple example illustrating the previous definitions.

Example 2 *Let $k = 2$, $m = 4$, $\mathcal{Y} = \{1, 2, 3, 4\}$, and $Y \sim p = (0.5, 0.2, 0.2, 0.1)$, i.e., $\mathbb{P}(Y = y) = p_y$. There are two true top-2 lists, namely, $\mathrm{T}_2(p) = \{(\{1, 2\}, (0.5, 0.2)), (\{1, 3\}, (0.5, 0.2))\}$. The list $s = (\{1, 4\}, (0.5, 0.1))$ is calibrated (relative to p) but fails to be valid because it cannot be true relative to a probability distribution on $\mathcal{Y}$. On the other hand, the list $r = (\{1, 4\}, (0.5, 0.2))$ is valid as it is true relative to $q = (0.5, 0.2, 0.1, 0.2)$ but fails to be calibrated.*

An invalid top-k list $t = (\widehat{Y}, \widehat{t})$ contains a *largest valid sublist* $t' = (\widehat{Y}', (\widehat{t}_y)_{y\in\widehat{Y}'})$. The largest valid sublist is uniquely determined by recursively removing the class $z \in \arg\min_{y\in\widehat{Y}} \widehat{t}_y$ with the lowest confidence score from the invalid list until a valid list remains. Removing a class $x \in \widehat{Y}$ with $\pi(t) > \widehat{t}_x$ cannot result in a valid top list $t' = (\widehat{Y} \setminus \{x\}, (\widehat{t}_y)_{y\in\widehat{Y}\setminus\{x\}})$ as long as there is another class z such that $\widehat{t}_x \geq \widehat{t}_z$ because $\pi(t) > \pi(t') > \widehat{t}_x \geq \widehat{t}_z$. Similarly, removing a class $x \in \widehat{Y}$ with $\pi(t) \leq \widehat{t}_x$ cannot prevent the removal of a class z if $\pi(t) > \widehat{t}_z$, because it does not decrease the proxy probability, $\pi(t') \geq \pi(t)$. Hence, *no* sublist containing a class with minimal confidence score in the original list is

valid, and removal results in a superlist of the largest valid sublist. Notably, the largest valid sublist may be the empty top-0 list $t_\emptyset$.

In what follows, I show how to construct consistent scoring functions for the top-k list functional using proper scoring rules. Recall from Section 2.1 that a scoring function $\mathrm{S}\colon \mathcal{T}_k \times \mathcal{Y} \to \overline{\mathbb{R}}$ is *consistent* for the top list functional T_k if the expected score under any probability distribution $p \in \mathcal{P}(\mathcal{Y})$ is minimized by any true top-k list $t \in \mathrm{T}_k(p)$, i.e.,

$$\mathbb{E}[\mathrm{S}(t, Y)] \leq \mathbb{E}[\mathrm{S}(s, Y)]$$

holds for $Y \sim p$ and any $s \in \mathcal{T}_k$. It is *strictly consistent* if the expected score is minimized only by the true top-k lists $t \in \mathrm{T}_k(p)$, i.e., the inequality is strict for $s \notin \mathrm{T}_k(p)$. The functional T_k is *elicitable* if a strictly consistent scoring function for T_k exists. In what follows, such a scoring function is constructed, giving rise to the following theorem.

Theorem 1 *The top-k list functional* T_k *is elicitable.*

Proof The theorem is an immediate consequence of either Theorem 9 or 10. ■

As the image of T_k is $\widetilde{\mathcal{T}}_k$ by definition, invalid top-k lists may be ruled out a priori, and the domain of S may be restricted to $\widetilde{\mathcal{T}}_k \times \mathcal{Y}$ in the above definitions. On the other hand, the use of a consistent scoring function on the larger domain $\mathcal{T}_k \times \mathcal{Y}$ merely encourages valid predictions, but it does not preclude invalid predictions. Any scoring function that is consistent for valid top list predictions can be extended by assigning an infinite score to any invalid top list regardless of the observation. This extension effectively precludes invalid predictions as an invalid prediction cannot outperform any arbitrary valid prediction, thereby disqualifying it in comparison. In what follows, I focus on the construction of consistent scoring functions for valid top lists at first and then propose a way of extending such scoring functions to invalid top lists that is less daunting than simply assigning an infinite score.

4. Mathematical Preliminaries

This section introduces some preliminary results, which are used heavily in the next section.

4.1 Symmetric Scoring Rules

The proposed scoring functions are based on symmetric proper scoring rules. Recall from Gneiting and Raftery (2007) that (subject to mild regularity conditions) any proper scoring rule $\mathrm{S}\colon \mathcal{P}(\mathcal{Y}) \to \overline{\mathbb{R}}$ admits a *Savage representation*,

$$\mathrm{S}(p, y) = G(p) - \langle G'(p), p \rangle + G'_y(p), \tag{7}$$

in terms of a concave function $G\colon \Delta_{m-1} \to \mathbb{R}$ and a supergradient $G'\colon \Delta_{m-1} \to \mathbb{R}^m$ of G, i.e., a function satisfying the *supergradient inequality*

$$G(q) \leq G(p) + \langle G'(p), q - p \rangle \tag{8}$$

for all $p, q \in \Delta_{m-1}$. Conversely, any function of the form (7) is a proper scoring rule. The function G is strictly concave if, and only if, S is strictly proper. It is called the *entropy (function)* of S, and it is simply the expected score $G(p) = \mathbb{E}[\mathrm{S}(p, Y)]$ under the posited distribution, $Y \sim p$. The supergradient inequality (8) is strict if G is strictly concave and $p \neq q$ (Jungnickel, 2015, Satz 5.1.12).

Let $\mathrm{Sym}(\mathcal{Y})$ denote the symmetric group on $\mathcal{Y}$, i.e., the set of all permutations of $\mathcal{Y}$. A scoring rule is called *symmetric* if scores are invariant under permutation of classes, i.e.,

$$\mathrm{S}((p_y), y) = \mathrm{S}((p_{\tau^{-1}(y)}), \tau(y))$$

holds for any permutation $\tau \in \mathrm{Sym}(\mathcal{Y})$ and all $y \in \mathcal{Y}, p \in \mathcal{P}(\mathcal{Y})$. Clearly, the entropy function G of a symmetric scoring rule is also symmetric, i.e., invariant to permutation in the sense that $G(p) = G((p_{\tau(y)}))$ holds for any permutation $\tau \in \mathrm{Sym}(\mathcal{Y})$ and any distribution $p \in \mathcal{P}(\mathcal{Y})$. Vice versa, any symmetric entropy function admits a symmetric proper scoring rule.

Proposition 2 *Let $G\colon \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{Y})$ be a concave symmetric function. Then there exists a supergradient $G'\colon \mathcal{P}(\mathcal{Y}) \to \mathbb{R}^m$ such that the Savage representation* (7) *yields a symmetric proper scoring rule.*

Proof Let $\bar{G}'$ be a supergradient of G. Using the shorthand $v_\tau = (v_{\tau^{-1}(y)})_{y \in \mathcal{Y}}$ for vectors $v = (v_y)_{y \in \mathcal{Y}} \in \mathbb{R}^m$ indexed by $\mathcal{Y}$ and permutations $\tau \in \mathrm{Sym}(\mathcal{Y})$, define G' by

$$G'(p) = \frac{1}{|\operatorname{Sym}(\mathcal{Y})|} \sum_{\tau \in \mathrm{Sym}(\mathcal{Y})} \bar{G}'_{\tau^{-1}}(p_\tau)$$

for $p \in \mathcal{P}(\mathcal{Y})$. By symmetry of G and the supergradient inequality,

$$G(q) = G(q_\tau) \leq G(p_\tau) + \langle \bar{G}'(p_\tau), q_\tau - p_\tau \rangle = G(p) + \langle \bar{G}'_{\tau^{-1}}(p_\tau), q - p \rangle$$

holds for all $p, q \in \mathcal{P}(\mathcal{Y})$ and $\tau \in \mathrm{Sym}(\mathcal{Y})$. Summation over all $\tau \in \mathrm{Sym}(\mathcal{Y})$ and division by the cardinality of the symmetric group $\mathrm{Sym}(\mathcal{Y})$ yields

$$G(q) \leq \frac{1}{|\operatorname{Sym}(\mathcal{Y})|} \sum_{\tau \in \mathrm{Sym}(\mathcal{Y})} (G(p) + \langle \bar{G}'_{\tau^{-1}}(p_\tau), q - p \rangle) = G(p) + \langle G'(p), q - p \rangle$$

for any $p, q \in \mathcal{P}(\mathcal{Y})$. Therefore, G' is a supergradient, and the Savage representation (7) yields a symmetric scoring rule since

$$\begin{aligned} G'(p) &= \frac{1}{|\operatorname{Sym}(\mathcal{Y})|} \sum_{\tau \in \mathrm{Sym}(\mathcal{Y})} \bar{G}'_{\tau^{-1}}(p_\tau) = \frac{1}{|\operatorname{Sym}(\mathcal{Y})|} \sum_{\tau \in \mathrm{Sym}(\mathcal{Y})} \bar{G}'_{(\tau \circ \rho)^{-1}}(p_{\tau \circ \rho}) \\ &= \frac{1}{|\operatorname{Sym}(\mathcal{Y})|} \sum_{\tau \in \mathrm{Sym}(\mathcal{Y})} \bar{G}'_{\rho^{-1} \circ \tau^{-1}}(p_{\tau \circ \rho}) = \frac{1}{|\operatorname{Sym}(\mathcal{Y})|} \sum_{\tau \in \mathrm{Sym}(\mathcal{Y})} (\bar{G}'_{\tau^{-1}}(p_{\tau \circ \rho}))_{\rho^{-1}} \\ &= \left(\frac{1}{|\operatorname{Sym}(\mathcal{Y})|} \sum_{\tau \in \mathrm{Sym}(\mathcal{Y})} \bar{G}'_{\tau^{-1}}((p_\rho)_\tau) \right)_{\rho^{-1}} = G'_{\rho^{-1}}(p_\rho) \end{aligned}$$

and

$$\langle G'(p), p\rangle = \langle G'_{\rho^{-1}}(p_\rho), p\rangle = \langle G'(p_\rho), p_\rho\rangle$$

holds for any permutation $\rho \in \mathrm{Sym}(\mathcal{Y})$ and all $p \in \mathcal{P}(\mathcal{Y})$. ■

On the other hand, not all proper scoring rules with symmetric entropy function are symmetric. The following result provides a necessary condition satisfied by supergradients of symmetric proper scoring rules.

Lemma 3 *Let* S *be a symmetric proper scoring rule. If* $p \in \Delta_{m-1}$ *satisfies* $p_x = p_z$ *for* $x, z \in \mathcal{Y}$*, then the supergradient* $G'(p) = (G'_y(p))_{y\in\mathcal{Y}}$ *at* p *in the Savage representation* (7) *satisfies* $G'_x(p) = G'_z(p)$*.*

Proof Let $\tau = (x\ z)$ be the permutation swapping x and z while keeping all other classes fixed. Using notation as in the proof of Proposition 2, the equality $\mathrm{S}(p, x) = \mathrm{S}(p_\tau, \tau(x))$ holds by symmetry of S. Since $p = p_\tau$, the Savage representation (7) yields $G'_x(p) = G'_{\tau(x)}(p) = G'_z(p)$. ■

The Brier score (3) and the logarithmic score (2) are both symmetric scoring rules. The entropy function of the Brier score is given by

$$G(p) = 1 - \sum_{y\in\mathcal{Y}} p_y^2, \tag{9}$$

whereas the entropy of the logarithmic score is given by

$$G(p) = -\sum_{y\in\mathcal{Y}} p_y \log(p_y)$$

(see Gneiting and Raftery, 2007).

4.2 Majorization and Schur-Concavity

In this section, I adopt some definitions and results on majorization and Schur-concavity from Marshall et al. (2011). The theory of majorization is essentially a theory of inequalities, which covers many classical results and a plethora of mathematical applications not only in stochastics.

For a vector $v \in \mathbb{R}^m$, the vector $v_{[\,]} := (v_{[i]})_{i=1}^m$, where

$$v_{[1]} \geq \cdots \geq v_{[m]}$$

denote the components of v in decreasing order, is called the *decreasing rearrangement* of v. A vector $w \in \mathbb{R}^m$ is a permutation of $v \in \mathbb{R}^m$ (i.e., w is obtained by permuting the entries of v) precisely if $v_{[\,]} = w_{[\,]}$. For vectors $v, w \in \mathbb{R}^m$ with equal sum of components, $\sum_i v_i = \sum_i w_i$, the vector v is said to *majorize* w, or $v \succ w$ for short, if the inequality

$$\sum_{i=1}^k v_{[i]} \geq \sum_{i=1}^k w_{[i]}$$

holds for all $k = 1, \ldots, m-1$.

Let $D \subseteq \mathbb{R}^m$. A function $f\colon D \to \mathbb{R}$ is *Schur-concave on* D if $v \succ w$ implies $f(v) \leq f(w)$ for all $v, w \in D$. A Schur-concave function f is *strictly Schur-concave* if $f(v) < f(w)$ holds whenever $v \succ w$ and $v_{[]} \neq w_{[]}$. In particular, any symmetric concave function is Schur-concave and strictly Schur-concave if it is strictly concave (Marshall et al., 2011, Chapter 3, Proposition C.2 and C.2.c). Hence, the following lemma holds.

Lemma 4 *The entropy function of any symmetric proper scoring rule is Schur-concave. It is strictly Schur-concave if the scoring rule is strictly proper.*

A set $D \subset \mathbb{R}^m$ is called *symmetric* if $v \in D$ implies $w \in D$ for all vectors $w \in \mathbb{R}^m$ such that $v_{[]} = w_{[]}$. By the Schur-Ostrowski criterion (Marshall et al., 2011, Chapter 3, Theorem A.4 and A.4.a) a continuously differentiable function $f\colon D \to \mathbb{R}$ on a symmetric convex set D with non-empty interior is Schur-concave if, and only if, f is symmetric and the partial derivatives $f_{(i)}(v) = \frac{\partial}{\partial v_i} f(v)$ increase as the components v_i of v decrease, i.e., $f_{(i)}(v) \leq f_{(j)}(v)$ if (and only if) $v_i \geq v_j$. Unfortunately, supergradients of concave functions do not share this property. The following is a slightly weaker condition, which applies to supergradients of symmetric concave functions.

Lemma 5 (Schur-Ostrowski condition for concave functions) *Let $f\colon D \to \mathbb{R}$ be a symmetric concave function on a symmetric convex set D, $v \in D$ and $f'(v) = (f'_1(v), \ldots, f'_m(v))$ be a supergradient of f at v, i.e., a vector satisfying the supergradient inequality*

$$f(w) \leq f(v) + \langle f'(v), w - v \rangle \tag{10}$$

for all $w \in D$. Then $v_i > v_j$ implies $f'_i(v) \leq f'_j(v)$.

Proof For $i = 1, \ldots, m$, let $e_i = (\mathbb{1}\{i = j\})_{j=1}^m$ denote the i-th vector of the standard basis of $\mathbb{R}^m$. Let $v \in D$ be such that $v_i > v_j$ for some indices i, j and let $0 < \varepsilon \leq v_i - v_j$. Define $w = v - \varepsilon e_i + \varepsilon e_j$. Then $v \succ w$ (by Marshall et al., 2011, Chapter 2, Theorem B.6) because w is obtained from v through a so-called 'T-transformation' (see Marshall et al., 2011, p. 32), i.e., $w_i = \lambda v_i + (1-\lambda) v_j$ and $w_j = \lambda v_j + (1-\lambda) v_i$ with $\lambda = \frac{v_i - v_j - \varepsilon}{v_i - v_j}$. Therefore, Schur-concavity of f implies $f(v) \leq f(w)$, and the supergradient inequality (10) yields

$$\varepsilon(f'_j(v) - f'_i(v)) = \langle f'(v), w - v \rangle \geq f(w) - f(v) \geq 0.$$

Hence, the inequality $f'_j(v) \geq f'_i(v)$ holds. ■

With this result, there is no need to restrict attention to differentiable entropy functions when applying the Schur-Ostrowski condition in what follows. Furthermore, true top-k lists can be characterized using majorization.

Lemma 6 *Let $Y \sim p$ be distributed according to $p \in \mathcal{P}(\mathcal{Y})$. The padded distribution $\widetilde{t}$ associated with a true top-k list $t \in \mathrm{T}_k(p)$ majorizes the padded distribution $\widetilde{s}$ associated with any calibrated top-k list $s \in \mathcal{T}_k$.*

Proof The sum of confidence scores $\sum_{i=1}^{k} \widetilde{t}_{[i]} = \sum_{i=1}^{k} p_{[i]} \geq \sum_{i=1}^{k} \widetilde{s}_{[i]}$ of a true top-k list is maximal among calibrated top-k lists by definition. Hence, the confidence score $\widehat{t}_{[i]} = \widetilde{t}_{[i]}$ of the true top-k list $t = (\widehat{Y}, \widehat{t}\,)$ matches the i-th largest class probability $p_{[i]}$ for $i = 1, \ldots, k$. Therefore, the partial sums $\sum_{i=1}^{\ell} \widetilde{t}_{[i]} = \sum_{i=1}^{\ell} p_{[i]} \geq \sum_{i=1}^{\ell} \widetilde{s}_{[i]}$ across the largest confidence scores are also maximal for $\ell = 1, \ldots, k-1$. Furthermore, the proxy probability $\pi(t) = \frac{1-\sum_{i=1}^{k} \widetilde{t}_{[i]}}{m-k}$ associated with a true top-k list is minimal among calibrated top-k lists. Hence, the partial sums

$$\sum_{i=1}^{\ell} \widetilde{t}_{[i]} = 1 - (m-\ell)\pi(t) \geq 1 - (m-\ell)\pi(s) = \sum_{i=1}^{\ell} \widetilde{s}_{[i]}$$

are maximal for $\ell > k$. ■

5. Consistent Top List Scores

Having reviewed the necessary preliminaries, this section shows that the proposed padded symmetric scores constitute a family of consistent scoring functions for the probabilistic top list functionals. The padded symmetric scores are defined for valid top lists and can be extended to invalid top lists by scoring the largest valid sublist, which yields a consistent scoring function. Strict consistency is preserved by adding an additional penalty term to the score of an invalid prediction.

5.1 Padded Symmetric Scores

From now on, let $\mathrm{S}\colon \mathcal{P}(\mathcal{Y}) \to \overline{\mathbb{R}}$ be a proper symmetric scoring rule with entropy function G. The scoring rule S is extended to valid top-k lists for $k = 0, 1, \ldots, m-1$ by setting

$$\mathrm{S}(t, y) := \mathrm{S}(\widetilde{t}, y)$$

for $y \in \mathcal{Y}, t \in \widetilde{\mathcal{T}}_k$, where $\widetilde{t} \in \Delta_{m-1}$ is the padded distribution (6) associated with the top-k list t. I call the resulting score $\mathrm{S}\colon \bigcup_{k=0}^{m} \widetilde{\mathcal{T}}_k \times \mathcal{Y} \to \overline{\mathbb{R}}$ a *padded symmetric score*. For example, the logarithmic score (2) yields the *padded logarithmic score*

$$\mathrm{S}_{\log}((\widehat{Y}, \widehat{t}\,), y) = \begin{cases} -\log(\widehat{t}_y), & \text{if } y \in \widehat{Y}, \\ \log(m-k) - \log(1 - \sum_{z \in \widehat{Y}} \widehat{t}_z), & \text{otherwise,} \end{cases}$$

whereas the Brier score (3) yields the *padded Brier score*

$$\mathrm{S}_{\mathrm{B}}((\widehat{Y}, \widehat{t}\,), y) = 1 + \sum_{z \in \widehat{Y}} \widehat{t}_z^{\,2} + \frac{(1 - \sum_{z \in \widehat{Y}} \widehat{t}_z)^2}{m-k} - 2 \cdot \begin{cases} \widehat{t}_y, & \text{if } y \in \widehat{Y}, \\ \frac{1-\sum_{z \in \widehat{Y}} \widehat{t}_z}{m-k}, & \text{otherwise.} \end{cases} \tag{11}$$

The following example shows that padded symmetric scores should not be applied to invalid top lists without further considerations.

Example 3 *If a padded symmetric score based on a strictly proper scoring rule is used to evaluate the invalid top-2 list s in Example 2, it attains a lower expected score than a true top list $t \in \mathrm{T}_2(p)$ because $\widetilde{s} = p$, whereas $\widetilde{t} \neq p$. Hence, the score would fail to be consistent.*

The following lemma shows that the expected score of a calibrated top list is fully determined by the top list itself and does not depend on (further aspects of) the underlying distribution.

Lemma 7 *Let* S *be a padded symmetric score. If $p \in \mathcal{P}(\mathcal{Y})$ is the true distribution of $Y \sim p$ and t is a calibrated valid top list, then the expected score of the top list t matches the entropy of the padded distribution $\widetilde{t}$,*

$$\mathbb{E}[\mathrm{S}(t, Y)] = G(\widetilde{t}).$$

Proof Let $t = (\widehat{Y}, \widehat{t}) \in \widetilde{\mathcal{T}}_k(p)$. Assume w.l.o.g. $k < m$ (the claim is trivial if $k = m$), and let $z \in \mathcal{Y} \setminus \widehat{Y}$. By Lemma 3 the supergradient at $\widetilde{t}$ satisfies $G'_y(\widetilde{t}) = G'_z(\widetilde{t})$ for all $y \notin \widehat{Y}$. Hence, the Savage representation (7) of the underlying scoring rule yields

$$\begin{aligned}\mathbb{E}[\mathrm{S}(t, Y)] &= G(\widetilde{t}) - \langle G'(\widetilde{t}), \widetilde{t} \rangle + \sum_{y \in \mathcal{Y}} p_y G'_y(\widetilde{t}) \\ &= G(\widetilde{t}) - \sum_{y \in \widehat{Y}} (p_y - \widehat{t}_y) G'_y(\widetilde{t}) - \left(\sum_{y \notin \widehat{Y}} p_y - (m-k)\pi(t) \right) G'_z(\widetilde{t}) = G(\widetilde{t})\end{aligned}$$

because t is calibrated. ∎

Padded symmetric scores exhibit an interesting property that admits balanced comparison of top list predictions of varying length. A top list score $\mathrm{S}\colon \bigcup_{k=0}^{m} \widetilde{\mathcal{T}}_k \times \mathcal{Y} \to \overline{\mathbb{R}}$ exhibits the *comparability property* if the expected score does not deteriorate upon extending a true top list, i.e., for $k = 0, 1, \ldots, m-1$ and any distribution $p \in \mathcal{P}(\mathcal{Y})$ of $Y \sim p$,

$$\mathbb{E}[\mathrm{S}(t_{k+1}, Y)] \leq \mathbb{E}[\mathrm{S}(t_k, Y)] \tag{12}$$

holds for $t_k \in \mathrm{T}_k(p)$ and $t_{k+1} \in \mathrm{T}_{k+1}(p)$. The following theorem shows that padded symmetric scores in fact exhibit the comparability property. I use the comparability property to show consistency of the individual padded symmetric top-k list scores $\mathrm{S}\,|_{\widetilde{\mathcal{T}}_k \times \mathcal{Y}}$ and to extend these scores to invalid top lists. Section 6 provides further discussion and some numerical insights.

Theorem 8 *Padded symmetric scores exhibit the comparability property.*

Proof Let S be a padded symmetric score and G be the concave entropy function of the underlying proper scoring rule. Let $Y \sim p$ be distributed according to some distribution $p \in \mathcal{P}(\mathcal{Y})$, and let $t_k = (\widehat{Y}_k, (p_y)_{y \in \widehat{Y}_k})$ be a calibrated valid top-k list for some $k = 0, 1, \ldots, m-1$, which is extended to a calibrated valid top-$(k+1)$ list $t_{k+1} = (\widehat{Y}_{k+1}, (p_y)_{y \in \widehat{Y}_{k+1}})$ in the sense that $\widehat{Y}_{k+1} = \widehat{Y}_k \cup \{z\}$ for some $z \in \mathcal{Y}$. It is easy to verify that $\widetilde{t}_{k+1} \succ \widetilde{t}_k$ since $p_z \geq \pi(t_k) \geq \pi(t_{k+1})$. Hence, the inequality $G(\widetilde{t}_{k+1}) \leq G(\widetilde{t}_k)$ holds by Schur-concavity of G (Lemma 4), which yields the desired inequality of expected scores by Lemma 7.

Clearly, there exists a true top-$(k+1)$ list $t_{k+1} \in \mathrm{T}_{k+1}(p)$ extending a true top-k list $t_k \in \mathrm{T}_k(p)$ in the above sense. By a symmetry argument, all true top lists of a given length

have the same expected score, and hence S exhibits the comparability property. ■

Note that the proof of Theorem 8 shows that (12) holds for any calibrated valid extension t_{k+1} of a calibrated valid top list t_k and not only for true top lists. I proceed to show that padded symmetric scores restricted to valid top-k lists are consistent for the top-k list functional.

Theorem 9 *Let $k \in \{0, 1, \ldots, m\}$ be fixed and $\mathrm{S}\colon \bigcup_{\ell=0}^{m} \widetilde{\mathcal{T}}_\ell \times \mathcal{Y} \to \overline{\mathbb{R}}$ be a padded symmetric score. Then the restriction $\mathrm{S}\,|_{\widetilde{\mathcal{T}}_k \times \mathcal{Y}}$ of the score S to the set of valid top-k lists $\widetilde{\mathcal{T}}_k$ is consistent for the top-k list functional T_k. It is strictly consistent if the underlying scoring rule $\mathrm{S}\,|_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}$ is strictly proper.*

Proof Let $p = (p_y)_{y \in \mathcal{Y}} \in \mathcal{P}(\mathcal{Y})$ be the true probability distribution of $Y \sim p$. Clearly, all true top-k lists in $\mathrm{T}_k(p)$ attain the same expected score by symmetry of the underlying scoring rule. Let $t = (\widehat{Y}, (p_y)_{y \in \widehat{Y}}) \in \mathrm{T}_k(p)$ be a true top-k list and $s = (\widehat{Z}, (\widehat{s}_y)_{y \in \widehat{Z}}) \in \widetilde{\mathcal{T}}_k$ be an arbitrary valid top-k list. To show consistency of $\mathrm{S}\,|_{\widetilde{\mathcal{T}}_k \times \mathcal{Y}}$, it suffices to show that the valid top-k list s does not attain a lower (i.e., better) expected score than the true top-k list t. Strict consistency follows if the expected score of any $s \notin \mathrm{T}_k(p)$ is higher than that of the true top-k list t.

First, consider $s \notin \mathrm{T}_k(p)$ to be a calibrated top-k list, i.e., $\widehat{s}_y = p_y$ for all $y \in \widehat{Z}$. Since $\widetilde{t}$ majorizes $\widetilde{s}$ by Lemma 6, the inequality

$$\mathbb{E}[\mathrm{S}(t, Y)] = G(\widetilde{t}\,) \leq G(\widetilde{s}\,) = \mathbb{E}[\mathrm{S}(s, Y)]$$

holds by Schur-concavity of the entropy function G (Lemma 4) and Lemma 7. If the underlying scoring rule is strictly proper, the entropy function is strictly (Schur-)concave, and hence the inequality is strict.

Now, consider s to be an uncalibrated top-k list, and let $r = (\widehat{Z}, (p_y)_{y \in \widehat{Z}})$ be the respective calibrated top-k list on the same classes. The calibrated top-k list r may not be valid and cannot be scored if it is invalid. However, its largest valid sublist $r' = (\widehat{Z}', (p_y)_{y \in \widehat{Z}'})$ with $\widehat{Z}' \subseteq \widehat{Z}$ can be scored. Let $z \in \mathcal{Y} \setminus \widehat{Z}$. The difference in expected scores

$$\begin{aligned}
&\mathbb{E}[\mathrm{S}(s, Y)] - \mathbb{E}[\mathrm{S}(r', Y)] \\
&= G(\widetilde{s}\,) - G(\widetilde{r}\,') - \langle G'(\widetilde{s}\,), \widetilde{s}\,\rangle + \langle G'(\widetilde{r}\,'), \widetilde{r}\,'\rangle + \sum_{y \in \mathcal{Y}} p_y (G'_y(\widetilde{s}\,) - G'_y(\widetilde{r}\,')) \\
&\qquad\qquad \text{(by the Savage representation (7))} \\
&\geq \langle G'(\widetilde{r}\,') - G'(\widetilde{s}\,), \widetilde{r}\,'\rangle + \sum_{y \in \mathcal{Y}} p_y (G'_y(\widetilde{s}\,) - G'_y(\widetilde{r}\,')) \quad \text{(by the supergradient inequality (8))} \\
&= \sum_{y \in \widehat{Z} \setminus \widehat{Z}'} (p_y - \pi(r'))(G'_y(\widetilde{s}\,) - G'_z(\widetilde{r}\,')) + \sum_{y \in \mathcal{Y} \setminus \widehat{Z}} (p_y - \pi(r'))(G'_z(\widetilde{s}\,) - G'_z(\widetilde{r}\,')) \\
&\qquad\qquad \text{(by Lemma 3)} \\
&= \sum_{y \in \widehat{Z} \setminus \widehat{Z}'} (p_y - \pi(r'))(G'_y(\widetilde{s}\,) - G'_z(\widetilde{s}\,)) \text{ (as } \textstyle\sum_{y \in \mathcal{Y} \setminus \widehat{Z}} (p_y - \pi(r')) = -\sum_{y \in \widehat{Z} \setminus \widehat{Z}'} (p_y - \pi(r'))\text{)}
\end{aligned}$$

is nonnegative by the fact that $(p_y - \pi(r')) \leq 0$ for $y \in \widehat{Z} \setminus \widehat{Z}'$ (since r' is the largest valid sublist) and Lemma 5 (and Lemma 3 if $\widehat{s}_y = \pi(s)$ for some $y \in \widehat{Z} \setminus \widehat{Z}'$).

Let $k' = |\widehat{Z}'|$. Then, r' scores no better than a true top-k' list $t_{k'} \in \mathrm{T}_{k'}(p)$, which in turn scores no better than t by the comparability property. Therefore,

$$\mathbb{E}[\mathrm{S}(s, Y)] \geq \mathbb{E}[\mathrm{S}(r', Y)] \geq \mathbb{E}[\mathrm{S}(t_{k'}, Y)] \geq \mathbb{E}[\mathrm{S}(t, Y)]$$

holds. If the underlying scoring function is strictly proper, the difference in expected scores $\mathbb{E}[\mathrm{S}(s, Y)] - \mathbb{E}[\mathrm{S}(r', Y)]$ above is strictly positive by strictness of the supergradient inequality (Jungnickel, 2015, Satz 5.1.12), and hence $\mathbb{E}[\mathrm{S}(t, Y)] < \mathbb{E}[\mathrm{S}(s, Y)]$ holds in this case, which concludes the proof. ■

5.2 Penalized Extensions of Padded Symmetric Scores

The comparability property can be used to extend a padded symmetric score S to invalid top lists in a consistent manner. To this end, recall that t' denotes the largest valid sublist of a top list $t = (\widehat{Y}, \widehat{t}) \in \mathcal{T}_k$. Assigning the score of the largest valid sublist to an invalid top-k list yields a consistent score by the comparability property. Strict consistency of the padded symmetric score S is preserved by adding a positive penalty term $c_{\text{invalid}} > 0$ to the score of the largest valid sublist in the case of an invalid top list prediction. I call the resulting score extension $\mathrm{S}\colon \bigcup_{k=0}^{m} \mathcal{T}_k \times \mathcal{Y} \to \overline{\mathbb{R}}$, which assigns the score

$$\mathrm{S}(t, y) = \mathrm{S}(t', y) + c_{\text{invalid}} \tag{13}$$

to an invalid top list $t \in \mathcal{T}_k \setminus \widetilde{\mathcal{T}}_k$ for $k = 1, 2, \ldots, m-1$, a *penalized extension* of a padded symmetric score. The following example illustrates that the positive penalty is necessary to obtain a strictly consistent scoring function.

Example 4 *Consider a setting similar to that of Example 2 with* $Y \sim p = (0.4, 0.2, 0.2, 0.2)$. *The padded distribution associated with the largest valid sublist* $t' = (\{1\}, (0.4))$ *of the invalid list* $t = (\{1, 2\}, (0.4, 0.1))$ *matches the true distribution,* $\widetilde{t'} = p$, *and hence the expected score of* t *in* (13) *exceeds the expected score of any true top-2 list in* $\mathrm{T}(p)$ *only if* $c_{\text{invalid}} > 0$.

The following theorem summarizes the properties of the proposed score extension.

Theorem 10 *Let* $k \in \{0, 1, \ldots, m\}$ *be fixed and* $\mathrm{S}\colon \bigcup_{\ell=0}^{m} \mathcal{T}_\ell \times \mathcal{Y} \to \overline{\mathbb{R}}$ *be a penalized extension* (13) *of a padded symmetric score with penalty term* $c_{\text{invalid}} \geq 0$. *Then the restriction* $\mathrm{S}\,|_{\mathcal{T}_k \times \mathcal{Y}}$ *of the score* S *to the set of top-k lists* $\mathcal{T}_k$ *is consistent for the top-k list functional* T_k. *It is strictly consistent if the underlying scoring rule* $\mathrm{S}\,|_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}$ *is strictly proper and the penalty term* c_{invalid} *is nonzero.*

Proof In light of Theorem 9, it remains to show that an invalid top-k list attains a worse expected score than a true top-k list $t \in \mathrm{T}_k(p)$ under the true distribution $p \in \mathcal{P}(\mathcal{Y})$ of $Y \sim p$. To this end, let $s \in \mathcal{T}_k$ be invalid. By construction of the penalized extension, the top list s is assigned the score of its largest valid sublist s' plus the additional penalty c_{invalid}. By consistency of the padded symmetric score and the comparability property, the expected score of s' cannot fall short of the expected score of t. Hence, $\mathrm{S}\,|_{\mathcal{T}_k \times \mathcal{Y}}$ is consistent

for the top-k list functional. If a positive penalty $c_{\text{invalid}} > 0$ is added, the score extension is strictly consistent given a strictly consistent padded symmetric score. ■

6. Comparability

The comparability property (12) ensures that additional information provided by an extended true top list does not adversely influence the expected score. The information gain is quantified by a reduction in entropy, which depends on the underlying scoring rule. Ideally, a top list score encourages the prediction of classes that account for a substantial portion of probability mass, while offering little incentive to provide unreasonably large top lists. In what follows, I argue that the padded Brier score satisfies this requirement.

Let S be a padded symmetric score with entropy function G (of the underlying proper scoring rule). Furthermore, let $1 \leq k < m$ and $t = (\widehat{Y}, (\widehat{t}_y)_{y \in \widehat{Y}})$ be a top-k list that accounts for most of the probability mass. In particular, assume that the unaccounted probability $\alpha = \alpha(t) = 1 - \sum_{y \in \widehat{Y}} \widehat{t}_y$ is less than the least confidence score but nonzero, i.e.,

$$0 < \alpha < \min_{y \in \widehat{Y}} \widehat{t}_y. \tag{14}$$

Let $Q = Q(t) = \{p \in \mathcal{P}(\mathcal{Y}) \mid t \in \mathrm{T}_k(p)\}$ be the set of all probability measures relative to which t is a true top-k list. Let $p \in Q$ assign the remaining probability mass α to a single class. Then p majorizes any $q \in Q$, and the distribution p has the lowest entropy, i.e., $G(p) = \min_{q \in Q} G(q)$, by Schur-concavity of the entropy function (Lemma 4). As the expected score of the top list t is invariant under distributions in Q by Lemma 7, the relative difference in expected scores between the true top list t and the true distribution $q \in Q$ is bounded by the relative difference in expected scores between t and p,

$$\frac{G(\widetilde{t}) - G(q)}{G(q)} \leq \frac{G(\widetilde{t}) - G(p)}{G(p)}.$$

The upper bound can be simplified by bounding the entropy of p from below as $G(p) \geq G((1-\alpha, \alpha, 0, \ldots, 0))$ by Schur-concavity of G.

If $\mathrm{S} = \mathrm{S}_{\mathrm{B}}$ is the padded Brier score (11) with entropy (9), the lower bound reduces to $G(p) \geq G((1-\alpha, \alpha, 0, \ldots, 0)) = 2(\alpha - \alpha^2) > \alpha$ since $\alpha < 0.5$ by assumption (14) and hence $2\alpha^2 < \alpha$. Therefore, the relative difference in expected scores has a simple upper bound,

$$\frac{G(\widetilde{t}) - G(p)}{G(p)} \leq \frac{\alpha^2 - \alpha\pi(t)}{2(\alpha - \alpha^2)} < \frac{\alpha^2}{\alpha} = \alpha.$$

For the padded logarithmic score, no such bound exists, and the deviation of the expected top list score from the optimal score can be severe, as illustrated in the following numerical example. The example sheds some light on the behavior of the (expected) padded symmetric scores and demonstrates that top lists of length $k > 1$ may provide valuable additional information over a simple mode prediction.

Example 5 *Suppose there are $m = 5$ classes labeled $1, 2, \ldots, 5$ and the true (conditional) distribution $p = p(\mathbf{x}) = \mathcal{L}(Y \mid \mathbf{X} = \mathbf{x})$ of Y (given a feature vector $\mathbf{x} \in \mathcal{X}$) is known.*

Table 1: Expected padded Brier scores and expected padded logarithmic scores of various types of true predictions and multiple distributions discussed in Example 5. Relative score differences (in percent) with respect to the optimal scores are in brackets.

			$\mathbb{E}[\mathrm{S}(\cdot, Y)]$		
p	S	Mode(p)	$\mathrm{T}_1(p)$	$\mathrm{T}_2(p)$	p
$p^{(\mathrm{h})}$	$\mathrm{S_B}$	0.02 (1.01%)	0.0199 (0.38%)	0.0198 (0%)	0.0198
$p^{(\mathrm{m})}$	$\mathrm{S_B}$	1 (80.18%)	0.6875 (23.87%)	0.5552 (0.04%)	0.5550
$p^{(\mathrm{l})}$	$\mathrm{S_B}$	1.5 (88.87%)	0.7969 (0.34%)	0.7955 (0.16%)	0.7942
$p^{(\mathrm{h})}$	$\mathrm{S_{log}}$	∞	0.0699 (24.75%)	0.0560 (0%)	0.0560
$p^{(\mathrm{m})}$	$\mathrm{S_{log}}$	∞	1.3863 (47.9%)	0.9425 (0.56%)	0.9373
$p^{(\mathrm{l})}$	$\mathrm{S_{log}}$	∞	1.6021 (0.45%)	1.5984 (0.23%)	1.5948

Table 1 features expected padded Brier and logarithmic scores of various types of truthful predictions under several distributions, as well as relative differences with respect to the optimal score. The considered distributions

$$p^{(\mathrm{h})} = (0.99, 0.01, 0, 0, 0), \quad p^{(\mathrm{m})} = (0.5, 0.44, 0.03, 0.02, 0.01),$$
$$p^{(\mathrm{l})} = (0.25, 0.22, 0.2, 0.18, 0.15).$$

exhibit varying degrees of predictability. Distribution $p^{(\mathrm{h})}$ exhibits high predictability in the sense that a single class can be predicted with high confidence. Distribution $p^{(\mathrm{m})}$ exhibits moderate predictability in that it is possible to narrow predictions down to a small subset of classes with high confidence, but getting the class exactly right is a matter of luck. Distribution $p^{(\mathrm{l})}$ exhibits low predictability in the sense that all classes may well realize. Predictions are of increasing information content. The first prediction is the true mode, i.e., a hard classifier without uncertainty quantification that predicts class 1 under all considered distributions. The hard mode is interpreted as assigning all probability mass to the predicted class. Scores are obtained by embedding the predicted class in the probability simplex or, equivalent, by scoring the top-1 list $(\{1\}, 1)$. The second prediction is the true top-1 list $(\{1\}, p_1)$, i.e., the mode with uncertainty quantification. The third prediction is the true top-2 list $(\{1, 2\}, (p_1, p_2))$, and the final prediction is the true distribution p itself.

By consistency of the padded symmetric scores, the true top-1 lists score better in expectation than the mode predictions, and, by the comparability property, the true top-2 lists score better than the top-1 lists, while the true distributions attain the optimal scores. The mode predictions perform significantly worse than the probabilistic predictions, which highlights the importance of truthful uncertainty quantification. Note that the log score assigns an infinite score in cases where the true outcome is predicted as having zero probability, and hence the mode prediction is assigned an infinite score with positive probability.

The expected padded Brier score of the probabilistic top-1 list under the highly predictable distribution $p^{(\mathrm{h})}$ is not far from optimal, whereas the respective logarithmic score is inflated

by the discrepancies between the padded and true distributions, even though the top list accounts for most of the probability mass ($\alpha = 0.01$). Deviations from the optimal scores are more pronounced under the logarithmic score in all considered cases.

Under the distribution exhibiting moderate predictability, the top-2 list prediction is much more informative than the top-1 list prediction, which results in a significantly improved score that is not far from optimal. Under the distribution exhibiting low predictability, all probabilistic predictions perform well as there is little information to be gained.

Estimation of small probabilities is frequently hindered by finite sample size. The specification of top list predictions in conjunction with the padded Brier score circumvents this issue as the Brier score is driven by absolute differences in probabilities, whereas the logarithmic score emphasizes relative differences in probabilities. In other words, the padded distribution is deemed a good approximation of the true distribution if the true top list accounts for most of the probability mass by the Brier score.

In light of these considerations, I conclude that the padded Brier score is suitable for the comparison of top list predictions of varying length.

7. Concluding Remarks

In this paper, I argued for the use of evaluation metrics rewarding truthful probabilistic assessments in classification. To this end, I introduced the probabilistic top list functionals, which offer a flexible probabilistic framework for the general classification problem. Padded symmetric scores yield consistent scoring functions, which admit comparison of various types of predictions. The padded Brier score appears particularly suitable as top lists accounting for most of the probability mass obtain an expected padded Brier score that is close to optimal.

The entropy of a distribution is a measure of uncertainty or information content. Majorization provides a relation characterizing common decreases in entropy shared by all symmetric proper scoring rules. In particular, for two distributions $p \in \mathcal{P}(\mathcal{Y})$ and $q \in \mathcal{P}(\mathcal{Y})$, the entropy of the distribution p does not exceed the entropy of q, i.e., $G(p) \leq G(q)$, if p majorizes q. The inequality is strict if the scoring rule is strictly proper and q is not a permutation of p.

Similar to probabilistic top-k lists, a probabilistic top-β list with $\beta \in (0, 1)$ may be defined as a minimal top list accounting for a probability mass of at least β. However, the padded symmetric scores proposed in this paper are not consistent for the top-β list functional, and the question whether this functional is elicitable constitutes an open problem for future research.

As a simple alternative to the symmetric padded scores proposed in this paper, top-k error (Yang and Koyejo, 2020) is also a consistent scoring function for the top-k list functional, however, it is not strictly consistent, as it does not evaluate the confidence scores. On a related note, strictly proper scoring rules are essentially top-k consistent surrogate losses in the sense of Yang and Koyejo (2020). The idea of a consistent surrogate loss is to find a loss function that is easier to optimize than the target accuracy measure such that the optimal confidence scores also optimize accuracy. However, confidence scores need not represent probabilities. In contrast, strictly proper scoring rules elicit probabilities.

Essentially, strictly proper scoring rules are consistent surrogates for any loss or scoring function that is consistent for a statistical functional.

Typically, classes cannot simply be averaged. Therefore, combining multiple class predictions may be difficult as majority voting may result in a tie, while learning individual voting weights or a meta-learner requires training data (see Kotsiantis et al., 2006, Section 8.3 for a review of classifier combination techniques). Probabilistic top lists facilitate the combination of multiple predictions as confidence scores can simply be averaged, which may be an easy way to improve the prediction.

The prediction of probabilistic top lists appears particularly useful in problems, where classification accuracy is not particularly high, as is frequently the case in multi-label classification. Probabilistic predictions are an informative alternative to classification with reject option. Furthermore, if it is possible to predict top lists of arbitrary length, the empty top-0 list may be seen as a reject option. Shifting focus towards probabilistic predictions may well increase prediction quality and usefulness in various decision problems, where misclassification losses are not uniform. The padded symmetric scores serve as general purpose evaluation metrics that account for the additional value provided by probabilistic assessments. Applying the proposed scores in a study with real predictions (e.g., the study conducted by Li et al. (2020)) is left as a topic for future work.

Acknowledgments

This work has been supported by the Klaus Tschira Foundation. The author gratefully acknowledges financial support from the German Research Foundation (DFG) through grant number 502572912. The author would like to thank Timo Dimitriadis, Tobias Fissler, Tilmann Gneiting, Norbert Henze, Alexander Jordan, Sebastian Lerch and Fabian Ruoff, as well as one anonymous reviewer, for helpful comments and discussion.

References

Zafer Barutcuoglu, Robert E. Schapire, and Olga G. Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22:830–836, 2006.

Glen W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5177–5186, 2019.

A. Philip Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society Series A*, 147:278–292, 1984.

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

Tobias Fissler, Rafael Frongillo, Jana Hlavinová, and Birgit Rudloff. Forecast evaluation of quantiles, prediction intervals, and other set-valued functionals. *Eletronic Journal of Statistics*, 15:1034–1084, 2021.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762, 2011.

Tilmann Gneiting. When is the mode functional the Bayes classifier? *Stat*, 6:204–206, 2017.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

Tilmann Gneiting and Johannes Resin. Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams and coefficient of determination, 2021. Preprint, arXiv:2108.03210v3.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

Radu Herbei and Marten H. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics*, 34:709–721, 2006.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*, 2021.

Dieter Jungnickel. *Optimierungsmethoden: Eine Einführung*. Springer Spektrum, Berlin, Heidelberg, 2015.

Sotiris B. Kotsiantis, Ioannis D. Zaharakis, and Panayiotis E. Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26: 159–190, 2006.

Cheng Li, Virgil Pavlu, Javed Aslam, Bingyu Wang, and Kechen Qin. Learning to calibrate and rerank multi-label predictions. In *Machine Learning and Knowledge Discovery in Databases*, pages 220–236, 2020.

Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer Series in Statistics. Springer, New York, second edition, 2011.

Allan H. Murphy. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, 105:803–816, 1977.

Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. In *Advances in Neural Information Processing Systems*, 2019.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85, 2011.

Adane N. Tarekegn, Mario Giacobini, and Krzysztof Michalak. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965, 2021.

Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 17: 168–192, 2020.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3:1–13, 2007.

Juozas Vaicenavicius, David Widmann, Andersson Carl, Fredrik Lindsten, Jacob Roll, and Thomas B. Schön. Evaluating model calibration in classification. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 3459–3467, 2019.

Forest Yang and Sanmi Koyejo. On the consistency of top-k surrogate losses. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10727–10735, 2020.

Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18:1338–1351, 2006.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26:1819–1837, 2014.