Copyright
by
Zachary James Smith
2019

The Dissertation Committee for Zachary James Smith
certifies that this is the approved version of the following dissertation:

Proper Scoring Rules: Properties and Applications

Committee:

J. Eric Bickel, Supervisor

Sheldon Landsberger

Benjamin D. Leibowicz

John J. Hasenbein

Proper Scoring Rules: Properties and Applications

by

Zachary James Smith

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2019

For my parents, Dan and Ruth

Acknowledgments

I would like to begin by thanking my advisor, J. Eric Bickel, who has supported my research since my arrival at the University of Texas. I am proud of the work we have done together. My research benefited greatly from your insightful input and our occasional (highly constructive) debates.

I would also like to thank the members of my committee. Dr. Hasenbein, I have learned more mathematics from you than from anybody else. Your enthusiastic teaching style, sense of humor, and ability to make difficult concepts simple made each of the many classes I took with you a pleasure. Dr. Landsberger, I appreciate your help in finding funding during my education, and for providing me with interesting external projects. Dr. Leibowicz, thank you for taking the time to help me with my research on this committee. I know that you have a bright future (and present) as a researcher and teacher in our field.

To my friends, at UT and elsewhere, as with everything else in my life, I would not be reaching this milestone if it were not for your support. A special thank you to Areesh, for all the time spent talking through the math, editing papers, spotting me at the gym, and being the best and most brilliant friend anyone could ask for. Andrew, Dan, Murat, and Sudesh: you are all terrific people, hard workers, and constant sources of joy. Truly, getting to know you during this process is far more important to me than the letters that will follow my name on a business card. Matt, Reiss, Ben, Eddie, Cullen, Eric, Erik and John, you are the greatest group of lifelong friends anyone could possibly ask

for. I do not know how I got so lucky.

To my fiance(!) Caitlyn, I love you. You always have my back. You raise me up. I couldn't ask for anything else. To my sisters, Natalie and Allison, you guys are amazing and I know I can count on you. That being said, hopefully obtaining this degree puts to bed the question of who is the smartest. Finally, to my parents. Any success I have is directly attributable to you. My love of learning, my determination, and my character have been built out of the care you have given me, the lessons you have imparted, and the examples you have set.

Proper Scoring Rules: Properties and Applications

Publication No. _____

Zachary James Smith, Ph.D.
The University of Texas at Austin, 2019

Supervisor: J. Eric Bickel

Proper and strictly proper scoring rules provide a rigorous method for evaluating the accuracy of a probabilistic forecast while encouraging honesty. In this dissertation, we develop new proper and strictly proper scoring rules. We introduce additive and strongly additive scoring rules that can be used to reward a sequence of probabilistic forecasts. We construct new tailored scoring rules and demonstrate a general economic interpretation for all weighted proper scores. We also present a matrix-based construction method for scoring forecasts that can be represented as affine transformations of an underlying distribution.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Proper scoring rules are used to evaluate the accuracy of a probabilistic forecast while incentivizing honest reporting [76]. The primary forecasting task that we consider in this dissertation is as follows. Let Ω be the discrete sample space of an uncertain experiment with $1 < n < \infty$ mutually exclusive and collectively exhaustive possible realizations $\omega_1, \omega, \ldots, \omega_n \in \Omega$. A scoring rule is used to reward a distributional forecast $\mathbf{r} = [r_1, r_2, \ldots, r_n]$, where r_i is the probability of observing ω_i. In many applications, particularly in decision analysis, such a forecast will be elicited from a subject matter expert. In this case, $\mathbf{r}$ is the expert's declared subjective probability distribution.

Under the scoring rule S, after observing a realization $\omega \in \Omega$, the forecaster receives a score of $S(\omega, \mathbf{r})$. When the function S is a proper or strictly proper scoring rule, the forecaster is incentivized to reveal a distribution $\mathbf{r}$ that corresponds with his privately held subjective beliefs in the following manner. Prior to an observation of the uncertainty, the forecaster's expected score is $E_{\mathbf{p}}[S(\omega, \mathbf{r})]$. A scoring rule is called *proper* if the expected score is maximized by setting $\mathbf{r} = \mathbf{p}$ and *strictly proper* if $\mathbf{p}$ is the unique maximizer. In other words, a forecaster maximizes his expected score by reporting truthfully.

Thus, scoring rules provide a useful tool in eliciting accurate forecasts from subject matter experts who may otherwise choose to report a forecast that does not align with their true beliefs. For example, in weather fore-

casting, where proper scoring rules have often been used to rank predictions, external incentives cause meteorologists to systematically avoid reporting certain probabilities [17]. In the context of decision analysis, proper scoring rules are particularly useful, as decision analysts frequently elicit distributional information for relevant uncertainties directly from a subject matter expert. In this context, a scoring rule is used to reward the expert for the accuracy of the prediction. When multiple forecasts are obtained for the same event, the scores can be used to rank the predictions.

Table 1.1 presents the Logarithmic, Quadratic and Spherical scoring rules. Each of these scoring rules is strictly proper and has been well studied in the literature and applied in practice. We will frequently reference these rules throughout the dissertation.

For a proper scoring rule S, the function

$$H(\mathbf{p}) := E_{\mathbf{p}}[S(\omega, \mathbf{p})] \tag{1.1}$$

is known as the *optimal expected score function*, the *sharpness*, or the (negative) *generalized entropy* function associated with S [44]. We view $-H(\mathbf{p})$ as a measure of the uncertainty in the distribution $\mathbf{p}$. From this perspective, a distribution contains more uncertainty if it yields a lower optimal expected score. The Shannon Entropy [73] is an important special case corresponding with the Logarithmic scoring rule.

In this dissertation, we identify and construct scoring rules that are particularly suited for use in decision analysis applications. In decision analysis, the forecast $\mathbf{r}$ obtained from an expert is typically a joint distribution over multiple uncertainties (see [23] for an example of a decision analysis problem with this structure). Due to the complexity of directly eliciting joint probabilities,

Table 1.1: Examples of strictly proper scoring rules

Rule	$S(\omega, \mathbf{r})$	Range	$H(\mathbf{p})$
Logarithmic	$\ln(r_\omega)$	$[-\infty, 0]$	$\sum p_\omega \ln(p_\omega)$
Quadratic	$2r_\omega - \|\mathbf{r}\|_2^2$	$[-1, 1]$	$\|\|\mathbf{p}\|\|_2^2$
Spherical	$\frac{r_\omega}{\|\mathbf{r}\|_2}$	$[0, 1]$	$\|\|\mathbf{p}\|\|_2$

the forecasting task to obtain $\mathbf{r}$ may be decomposed into forecasting exercises involving the elicitation of marginal or conditional distributions. The full joint distribution $\mathbf{r}$ might never be fully specified, either because knowledge of all joint probabilities is unnecessary, or because the expert cannot specify all the necessary information with confidence. For example, marginal forecasts might be simple to obtain, but the full dependence structure between the random variables might not be well understood. After elicitation, the forecast $\mathbf{r}$ may be subsequently used to solve a specific decision problem.

Each part of this dissertation develops practical methodology for proper scoring in the setting described above. First, we describe additive scoring rules, which are particularly well suited for scoring the forecasts for multiple, possibly dependent random variables, or, alternatively, for directly scoring a corresponding joint distribution. In particular, the additive scoring rules we construct have a useful information-theoretic consistency property: the score for a joint forecast is equal to a sum of scores for marginal and/or conditional forecasts. We further demonstrate that commonly used rules that are not additive can give nonsensical ex-post scores when applied over multiple uncertainties.

Second, we construct general classes of tailored weighted scoring rules.

Tailored scoring rules [52] arise naturally in decision analysis applications and align the information-gathering incentives of a decision maker with those of the forecaster. This connection is established by choosing a scoring rule whose optimal expected score function matches the optimal rewards available to the decision maker, so that improving the score directly implies greater expected returns for the decision maker.

Weighted scoring rules directly incorporate a known baseline forecast into the score function. Generalizing the results in [54], where weighted scoring rules are first defined, we demonstrate that every weighted proper scoring rule can be interpreted as a tailored scoring rule. In particular, each proper weighted scoring rule is tailored to a decision problem involving the optimization of a convex risk measure [35] by a risk averse investor.

Finally, we give a matrix-based construction method for proper scoring rules that can be applied to linear forecasts. A linear forecast is information representable as an affine transformation of the forecaster's underlying probability distribution. We demonstrate that linear scoring rules generalize typical distributional scoring rules. We also introduce tailored linear scoring rules, which align incentives between a forecaster and a decision maker who solves a distributionally robust optimization problem.

Prior to describing our contributions in detail, we provide a high-level review of the scoring rules literature.

1.1 Literature Review

The literature on scoring rules is extensive, with applications in weather forecasting, sports betting, education, economics, machine learning, and theo-

retical statistics. The previous work most relevant to ours will be expounded upon in subsequent sections of the dissertation. The literature on scoring rules extends at least back to Good [42], who first considered the Logarithmic scoring rule presented in Table 1.1. Many other scoring rules have since been proposed and studied. Detailed discussions of the Quadratic and Spherical scoring rules (Table 1.1) can be found in [72] and [53], respectively.

The fundamental mathematical characterization of proper scoring rules was introduced by McCarthy [60] and later proved formally by Savage [71] and Hendrickson and Buehler [46]. These authors establish a link between scoring rules and convex functions by identifying a correspondence between a proper scoring rule S and its associated generalized entropy function H. These results have been further generalized within a modern framework in [29, 40, 62]. We will apply these theorems, which are presented formally in Chapter 2, throughout the dissertation.

Although our focus is on the discrete setting described above, proper scoring rules can be extended in many directions. For a general, measure-theoretic characterization of scoring rules, see Gneitting and Raftery [40]. Using the basic construction described therein, scoring rules have been created to reward forecasts for probability density functions (PDFs) and cumulative distribution functions (CDFs) [59]. In addition, scoring rules exist for evaluating forecasts for other distributional information, such as moments and quantiles [55]. For an overview of the many variants of proper scoring rules that appear in the literature, see [76, 40, 21, 32].

Recent work, starting with Lambert [57], has led to the development of proper scores for so-called elicitable properties of a distribution. An elicitable property is a real number or vector of real numbers that can be represented

as a function $F(\mathbf{p})$ of a distribution $\mathbf{p}$. Examples include the mean, variance, quantiles, or the distribution itself. A property is elicitable if there exists a scoring rule for which the expected score with respect to the distribution $\mathbf{p}$ is maximized at $F(\mathbf{p})$. The scoring rules for distributional forecasts, described in the beginning of the introduction, are proper for the specific property $F(\mathbf{p}) := \mathbf{p}$.

Scoring rules and their generalizations have also received attention in the statistics community. Proper scoring rules can be applied in the context of parameter estimation; finding the distribution within a parametric family that minimizes the score divergence with respect to the empirical distribution yields an unbiased estimating equation [30]. Furthermore, as established in [44], the function H defined in Equation (1.1) that is associated with a strictly proper scoring rule can be interpreted as a generalized measure of entropy. This connection between information theory and proper scoring rules has been further examined in [29, 54, 63].

In the context of decision making, efforts have been directed towards designing scoring rules that align the information-gathering incentives of the forecaster with those of a decision maker [51, 54, 34, 52]. As pointed out in [44], scoring rules arise naturally in the context of generic optimization problems under uncertainty, although the resulting reward function might not have a closed form, complicating the practicality of this observation. We revisit and expand upon the connection between scoring rules and optimization problems in Chapter 4. The applicability of scoring rules under risk aversion [14, 22] and competition between forecasters [58] has also been analyzed.

Applications for proper scoring rules are wide-ranging. Examples include weather forecasting [20], psychology [13], statistics [33, 47], sports betting (see

[25] and the references therein), education [15], and machine learning [49]. More comprehensive overviews of the literature surrounding strictly proper scoring rules, and more detail on application areas, are provided in [76, 40, 21, 32].

1.2 Organization of the Dissertation

The material comprising this dissertation will be divided into five chapters.

Chapter 2 provides the necessary notation and mathematical background. All material presented in Chapter 2 is already established in the literature.

Chapter 3 defines additive and strongly additive scoring rules, explains why such rules may be desirable in practical applications, and gives methods for constructing these rules. We also prove new theorems connecting the additive properties of scoring rules with well-known properties of entropy measures, deepening the established connections between the theory of proper scoring rules and information theory.

Chapter 4 extends the connection between the weighted scoring rules introduced in Jose et al. [54] and generic utility maximization problems. In the process, we introduce new families of weighted scoring rules and establish connections between these rules and convex risk measures.

Chapter 5 examines scoring rules that can be used to incentivize honesty when a forecast is provided that is an affine transformation of the forecaster's underlying subjective probability distribution. As we show, many existing families of proper scoring rules can be cast in this form. After establishing a construction method for linear scoring rules, we present numerous practical applications where these scoring functions can be used. We also introduce

tailored scoring rules for a family of distributionally robust decision problems.

Finally, Chapter 6 provides concluding remarks and describes potential avenues for future research.

Chapter 2

Mathematical Background and Notation

2.1 Introduction

This section introduces the mathematical background and notation that will be used throughout the dissertation. Chapter-specific notation and background will be introduced as needed. We adopt the framework of Hendrickson and Buehler [46], which characterizes proper scoring rules as subgradients of convex functions. Their central theorem is stated formally and further explained in Section 2.3. We also introduce the probabilistic notation that will be used.

2.2 Defining Proper and Strictly Proper Scoring Rules

We use Ω to denote a discrete sample space with $1 < |\Omega| = n < \infty$ outcomes $\omega_1, \omega_2, \ldots, \omega_n$. We use the bold letters $\mathbf{p}, \mathbf{q}$ and $\mathbf{r}$ to denote probability distributions over Ω. For the distribution $\mathbf{p} = [p_{\omega_1}, p_{\omega_2}, \ldots, p_{\omega_n}]$, a component p_ω represents the probability of the generic outcome $\omega \in \Omega$. The set of all possible probability distributions over Ω is given by

$$\mathcal{P} := \left\{ \mathbf{p} : \mathbf{p} \in \mathbb{R}^n, \mathbf{p} \geq 0, \sum_{\omega \in \Omega} p_\omega = 1 \right\}.$$

Let $\mathcal{A}$ be the cone defined as:

$$\mathcal{A} := \{\lambda \mathbf{p} : \mathbf{p} \in \mathcal{P}, \lambda > 0\}. \tag{2.1}$$

We interpret $\mathcal{A}$ as the set of all denormalized probability distributions over $\mathcal{X}$.

The notation $\mathcal{P}^+$ and $\mathcal{A}^+$ will denote the relative interiors of $\mathcal{P}$ and $\mathcal{A}$, respectively. For real-valued vectors $\mathbf{q}, \mathbf{p}$ of length n, $\mathbf{q}^\top \mathbf{p}$ will denote the standard inner product $\sum_{i=1}^n q_i p_i$. We denote by $\mathbf{1} \in \mathbb{R}^n$ the vector with all components equal to 1. Similarly, for $\omega \in \Omega$, $\boldsymbol{\delta}_\omega$ will be taken as a vector with $\delta_\omega = 1$ and all other components 0. We now give a formal definition for proper and strictly proper scoring rules.

A *scoring rule* is a function $S : \Omega \times \mathcal{A} \to \mathbb{R}$ that is 0-homogeneous in its second argument, meaning that $S(\omega, \lambda \mathbf{r}) = S(\omega, \mathbf{r})$ for all scalars $\lambda > 0$ [46, 62]. Defining scoring rules to be 0-homogeneous encodes the fact that the score does not depend on positive scaling of the forecasted distribution. For example, the distributions $[.5, .5]$ and $[1, 1]$ will be scored equivalently. A scoring rule is *proper* if $E_{\mathbf{p}}[S(\omega, \mathbf{r})] \leq E_{\mathbf{p}}[S(\omega, \mathbf{p})]$ for all $\mathbf{p}, \mathbf{r} \in \mathcal{P}$, and *strictly proper* if the inequality is strict when $\mathbf{r} \neq \mathbf{p}$.

The definition given above extends the domain of scoring rules from $\mathcal{P}$ to a full-dimensional subset of $\mathbb{R}^n$. Among other advantages, this extension allows clear definition of $\nabla_{\mathbf{p}} S(x, \mathbf{p})$, the gradient of the scoring rule with respect to its second argument. Any proper or strictly proper scoring rule S defined on $\mathcal{P}$, such as the Logarithmic, Quadratic and Spherical rules presented in Table 1.1, can easily be extended to allow for scoring of denormalized distributions by using $\hat{S}(\omega, \mathbf{r}) = S(\omega, \mathbf{r}/\mathbf{1}^\top \mathbf{r})$.

2.3 Characterizations of Proper Scoring Rules

The primary characterization result of this section, due to Hendrickson and Buehler [46], demonstrates that proper scoring rules are nothing but sub-

gradients of convex functions. We will appeal to this theorem frequently, and thus devote this section to a full explanation of this result.

A function $H : \mathcal{A} \to \mathbb{R}$ is *convex* if:

$$H(\theta\mathbf{p} + (1-\theta)\mathbf{r}) \leq \theta H(\mathbf{p}) + (1-\theta)H(\mathbf{r})$$

for all $\mathbf{p}, \mathbf{r} \in \mathcal{A}$ and $0 \leq \theta \leq 1$. H is said to be *homogeneous* of order k if $H(\lambda\mathbf{p}) = \lambda^k H(\mathbf{p})$ for every $\mathbf{p} \in \mathcal{A}$ and $\lambda > 0$. In the context of proper scoring rules, we will encounter 1-homogenenous and 0-homogeneous functions.

For a convex function H, a vector $\mathbf{v}(\mathbf{r})$ satisfying

$$H(\mathbf{r}) + \mathbf{v}(\mathbf{r})^\top(\mathbf{p} - \mathbf{r}) \leq H(\mathbf{p}) \tag{2.2}$$

for all $\mathbf{p} \in \mathcal{A}$ is called a *subgradient* of H at $\mathbf{r} \in \mathcal{A}$. For a 1-homogeneous convex function, it can be shown that $\mathbf{v}(\mathbf{p})^\top\mathbf{p} = H(\mathbf{p})$. In this case, it follows from Equation (2.2) that $\mathbf{v}(\mathbf{r})$ is a subgradient of H at $\mathbf{r}$ if

$$\mathbf{v}(\mathbf{r})^\top\mathbf{p} \leq H(\mathbf{p}) \tag{2.3}$$

for all $\mathbf{p} \in \mathcal{A}$, with $\mathbf{v}(\mathbf{p})^\top\mathbf{p} = H(\mathbf{p})$. A subgradient, as defined in (2.3), is said to be *strict* if $\mathbf{v}(\mathbf{r})^\top\mathbf{p} < H(\mathbf{p})$ whenever $\mathbf{p}$ is not equivalent to $\mathbf{r}$ up to multiplication by a positive scalar. When H is 1-homogeneous, the subgradient map, as a function of $\mathbf{q}$, can be chosen to be 0-homogeneous [46]. Indeed, for $\mathbf{r} \in \mathcal{P}$, we have

$$\mathbf{v}(\mathbf{r})^\top\lambda\mathbf{r} = \lambda H(\mathbf{r}) = H(\lambda\mathbf{r}).$$

The subdifferential of H at $\mathbf{r}$ is given by

$$\partial H(\mathbf{r}) := \{\mathbf{v} : \mathbf{v} \text{ is a subgradient of } H \text{ at } \mathbf{r}\}. \tag{2.4}$$

If H is differentiable, then $\partial H(\mathbf{q}) = \{\nabla H(\mathbf{q})\}$, where $\nabla H(\mathbf{q})$ denotes the gradient of H at $\mathbf{q}$.

For a proper scoring rule S, let

$$S(\mathbf{p}) := [S(\omega_1, \mathbf{p}), S(\omega_2, \mathbf{p}), \ldots, S(\omega_n, \mathbf{p})]$$

be the n-dimensional vector of possible score realizations given a report $\mathbf{p}$. The optimal score function associated with a proper scoring rule S, $H(\mathbf{p}) = \mathbf{p}^\top S(\mathbf{p})$, is known as the generalized *entropy measure* associated with S [44]. The optimal expected score function can be viewed as a generalized measure of entropy because Shannon entropy is a special case. Furthermore, as shown in [44], score entropies satisfy a max-min property shared by the Shannon entropy. The term "entropy" is often used in the literature to describe the optimal expected function associated with a proper score and we will adopt that convention. The generalized entropy measures corresponding with a proper scoring rule satisfies the following definition.

Definition 2.1. *Entropy Measure*

A function $H : \mathcal{A} \to \mathbb{R}$ is an entropy measure (or entropy function) if H is convex and 1-homogeneous on $\mathcal{A}$, and strictly convex on $\mathcal{P}$.

The primary Theorem of this section of the dissertation, stated below, provides the fundamental relationship between proper scoring rules and entropy measures.

Theorem 2.1. *Hendrickson and Buehler [46]*

The function $S : \Omega \times \mathcal{A} \to \mathbb{R}$ is a strictly proper scoring rule if, and only if, for each $\mathbf{r}$ in $\mathcal{A}$, $S(\mathbf{r}) \in \partial H(\mathbf{r})$ for some entropy measure H. The rule S

will be proper if the entropy measure H is convex (but not strictly convex) on $\mathcal{P}$.

In particular, given a strictly proper scoring rule S, it follows from propriety that S is a strict subgradient of the 1-homogeneous convex function $H(\mathbf{p}) = \mathbf{p}^\top S(\mathbf{p})$. Conversely, for a 1-homogeneous convex function H, any scoring rule generated from H (by forming subgradients) will have $H(\mathbf{p})$ as its associated entropy function on $\mathcal{P}$. These subgradients will be strict when H is strictly convex on $\mathbf{p} \in \mathcal{P}$, and $S(x, \mathbf{r}) := \boldsymbol{\delta}_x^\top \mathbf{v}(\mathbf{r})$ will be a strictly proper scoring rule for any choice of $\mathbf{v}(\mathbf{r}) \in \partial H(\mathbf{r})$. If the convex and 1-homogeneous function H is differentiable on $\mathcal{A}$, then the scoring rule S associated with H is uniquely determined (at least on the interior of $\mathcal{A}$) by $S(\mathbf{r}) = \nabla_{\mathbf{r}} H(\mathbf{r})$.

A parallel result to Theorem 2.1 was stated by McCarthy [60] and expounded upon by Savage [71]. Take a convex function $H : \mathcal{P} \to \mathbb{R}$, and let $\mathbf{v}(\mathbf{r})$ be a subgradient of H at $\mathbf{r} \in \mathcal{P}$ in the sense of (2.2). (H is no longer assumed to be 1-homogeneous and H has been limited to have domain $\mathcal{P}$.) Then it can be shown that

$$S(\omega, \mathbf{r}) := H(\mathbf{r}) + \boldsymbol{\delta}_\omega^\top \mathbf{v}(\mathbf{r}) - \mathbf{r}^\top \mathbf{v}(\mathbf{r}) \tag{2.5}$$

is a proper scoring rule (strictly proper when H is strictly convex). When H has a unique gradient at every point $\mathbf{p} \in \mathcal{P}$, then Equation (2.5) becomes

$$S(\omega, \mathbf{r}) = H(\mathbf{r}) + \boldsymbol{\delta}_\omega^\top \nabla H(\mathbf{r}) - \mathbf{r}^\top \nabla H(\mathbf{r}). \tag{2.6}$$

This scoring rule is uniquely determined by H, and satisfies $E_{\mathbf{p}}[S(X, \mathbf{p})] = H(\mathbf{p})$. It follows from Equation (2.6) that

$$H(\mathbf{p}) - E_{\mathbf{p}}[S(X, \mathbf{r})] = H(\mathbf{p}) - H(\mathbf{r}) - \nabla H(\mathbf{p})^\top (\mathbf{p} - \mathbf{r}), \tag{2.7}$$

where the term on the right hand side is the *Bregman divergence* [31] associated with the convex function H.

The Bregman divergence associated with a proper scoring rule can be viewed as a generalized "distance" measure between probability distributions. It is minimized at 0 when $\mathbf{p} = \mathbf{r}$, although it need not be symmetric. Equation (2.7) implies that a forecaster maximizes his expected score by minimizing the associated Bregman divergence. Conversely, the Bregman divergence governs how the score function measures ex-post accuracy of a forecast. In particular, a higher score, after an observation ω, corresponds with a lower Bregman divergence between the empirical distribution and the forecasted distribution. When multiple realizations of the underlying uncertainty are observed, and a sequence of scores is added, the same interpretation holds.

Theorem 2.1 implies the characterization (2.5). In particular, given any convex function $H(\mathbf{p})$ with domain $\mathcal{P}$, H can be extended as a 1-homogeneous convex function on $\mathcal{A}$ by instead considering $H' : \mathcal{A} \to \mathbb{R}$ with $H'(\mathbf{p}) := (\mathbf{1}^\top \mathbf{p}) H(\mathbf{p}/\mathbf{1}^\top \mathbf{p})$. Equation (2.5) then follows by differentiation. The following example illustrates the results presented in this section.

Example 2.1. *Logarithmic Scoring Rule*

Choose $H(\mathbf{p}) = \sum p_\omega \ln(p_\omega/\mathbf{1}^\top \mathbf{p})$ *for* $\mathbf{p} \in \mathcal{A}$. H *is* 1*-homogeneous and strictly convex on* $\mathcal{P}$. *Its associated* 0*-homogeneous scoring rule, found by differentiation, is* $S(\omega, \mathbf{r}) = \ln(r_\omega/\mathbf{1}^\top \mathbf{r})$. *Restricted to* $\mathcal{P}$, *the entropy* H *and score function* S *correspond exactly with the Shannon entropy and the logarithmic scoring rule, respectively. The Bregman divergence measure associated with the Logarithmic rule is the Kullback-Leibler divergence [14].*

This section concludes by defining the notion of equivalence [36] for proper

scoring rules. Given a proper scoring rule S, the reward function

$$S'(\omega, \mathbf{p}) := aS(\omega, \mathbf{p}) + b, \ a > 0, b \in \mathbb{R} \tag{2.8}$$

is also a proper scoring rule. Scoring rules S and S' are said to be *equivalent* if one is a positive affine transformation of the other. In general, we will not distinguish between scoring rules, up to equivalence.

2.4 Probabilistic Notation

We now define our basic notation for expressing probabalistic events, marginal distributions, conditional distributions, and joint distributions. These concepts will be particularly relevant in Chapters 3 and 5, which assume that forecasts are provided for the distributions (or joint distribution) governing multiple uncertainties.

In typical forecasting applications, we will be interested in obtaining information regarding a joint distribution over $J < \infty$ uncertainties. In this case, the individual sample spaces are denoted by Ω^j, $j = 1, \ldots, J$, with $|\Omega^j| = n^j$. The joint sample space is then the cartesian product $\Omega = \Omega^1 \times \Omega^2 \times \ldots \times \Omega^J$ with $n = \prod_{j=1}^{J} n^j$ joint outcomes of the form $\omega = (\omega^1, \omega^2, \ldots, \omega^J)$. X and Y (Y^j) will denote discrete random variables defined on Ω (Ω^j). The set of all (denormalized) probability distributions over Ω^j will be denoted by ($\mathcal{A}^j$) $\mathcal{P}^j$, and the joint sample space by $\mathcal{P}$. $[J]$ will be shorthand for the index set $\{1, 2, \ldots, J\}$.

Let the vector $\chi(A) \in \mathbb{R}^n$ be such that the ith component is 1 if $\omega_i \in A$ for $\omega_i \in \Omega$ and 0 otherwise. This notation can be used to represent many functions of the underlying probability distribution using vector inner products. For

example, given a joint distribution $\mathbf{p} \in \mathcal{P}$, the marginal probability of the ith outcome of the jth uncertainty can be computed as

$$p^j_{\omega^j_i} = \chi\left(\{(\omega^1, \ldots, \omega^j, \ldots, \omega^J) \in \Omega : \ \omega^j = \omega^j_i\}\right)^\top \mathbf{p},$$

or cumulative probabilities (in the one-dimensional case) as

$$P(Y \leq y) = \chi(\{\omega : Y(\omega) \leq z\})^\top \mathbf{p}.$$

Conditional probabilities, expectations, etc. can be represented similarly.

A given joint distribution $\mathbf{p} \in \mathcal{P}$ has associated marginal distributions $\mathbf{p}^j \in \mathcal{P}^j$, for $j \in [J]$. The associated conditional probability distributions will be denoted by

$$\mathbf{p}^{j|\omega_{i_1},\ldots,\omega_{i_k}} \in \mathcal{P}^j$$

for a given $j \in [J]$. In particular, $\mathbf{p}^{j|\omega_{i_1},\ldots,\omega_{i_k}}$ is the conditional distribution for the jth uncertainty given the k observations $\omega^{i_1}, \ldots, \omega^{i_k}$, where $i_k \neq j$ for each k. Prior to observing the realizations $\omega^{i_1}, \ldots, \omega^{i_k}$, the distribution $\mathbf{p}^{j|\omega_{i_1},\ldots,\omega_{i_k}}$ is a random vector. Whether we are interpreting $\mathbf{p}^{j|\omega_{i_1},\ldots,\omega_{i_k}}$ as random or as a fixed distribution will be clear from the context.

Chapter 3

Additive Scoring Rules

3.1 Chapter Summary

This chapter develops strictly proper scoring rules that may be used to evaluate the accuracy of a sequence of probabilistic forecasts[1]. When forecasts are submitted for multiple uncertainties, competing forecasts are typically ranked by their cumulative or average score. One could also score the implied joint distributions, if available. We demonstrate that these measures of forecast accuracy disagree under some commonly used rules. Furthermore, we show that forecast rankings can depend upon the selected scoring procedure so that the relative ranking of probabilistic forecasts does not depend solely upon the information content of those forecasts and the observed outcome. Instead, the relative ranking of forecasts is a function of the *process* by which those forecasts are evaluated.

In light of these issues, we describe additive and strongly additive strictly proper scoring rules, which have the property that the score for the joint distribution is equal to a sum of scores for the associated marginal and conditional

[1]This chapter is based on the author's work found in the accepted publication: *Additive Scoring Rules for Discrete Sample Spaces, Smith Z, Bickel JE, in print at Decision Analysis.* The second author (Bickel) is the dissertation author's (Smith) faculty advisor. The dissertation author was responsible for producing the mathematical content in the paper under guidance and with feedback from the second author. The dissertation author was also the primary writer and produced all tables and figures.

distributions. We give methods for constructing additive rules and demonstrate that the Logarithmic score is the only strongly additive rule. Finally, we connect the additive properties of scoring rules with analogous properties for a general class of entropy measures.

3.2 Introduction

This section develops proper scoring rules that work well when a sequence of forecasts for $J < \infty$ uncertainties is obtained. In many applications, a series of possibly dependent forecasts are provided by a single forecaster. For example, a meteorologist may provide daily forecasts regarding the chance of precipitation [77, 16]. Or, a student may provide a series of assessments on an exam that is scored using a strictly proper rule [14]. In other situations, assessments may be given for a series of marginal and conditional distributions that together specify another uncertainty. For example, a decision analyst may face a problem that depends on a joint distribution over multiple uncertainties [23, 18]. From a forecasting perspective, it may be simpler to factor this distribution into multiple, possibly dependent distributions.

In these examples, the marginal and conditional distributions are typically scored individually and added (or averaged) to obtain a cumulative measure of accuracy. Over the series of J forecasting tasks, after a realization $\omega^j \in \Omega^j$ for each uncertainty is observed, the forecaster receives a score (or a reward) of

$$\sum_{j=2}^{J} S^j \left(\omega^j, \mathbf{r}^{j|\omega^{j-1},\ldots,\omega^1}\right) + S^1(\omega^1, \mathbf{r}^1), \tag{3.1}$$

where $\mathbf{r}^{j|\omega^{j-1},\ldots,\omega^1}$ is the forecast for the conditional distribution of the jth uncertainty given realizations ω^k for $k \in \{j-1,\ldots,1\}$, and each S^j is a

strictly proper scoring rule. The functional form of the scoring rule is allowed to depend on j.

As a simple example of this procedure, consider a weather forecaster who provides the following probabilistic assessments: $P(\text{Rain Today}) = .80$, $P(\text{Rain Tomorrow}|\text{Rain Today}) = .20$, $P(\text{Rain Tomorrow}|\text{No Rain Today}) = .60$. If it does not rain today but rains tomorrow, the weather forecaster's cumulative score is given by:

$$S^2(\text{Rain Tomorrow}, [.60, .40]) + S^1(\text{No Rain Today}, [.80, .20]).$$

If several forecasters submit predictions in this manner, each collection of forecasts may be ranked according to its cumulative or average scores [76, 40]. This scoring procedure is used in weather forecasting [77, 70, 20, 40], sports betting [25], financial market predictions [75], and student testing [15]. Although adding scores is sometimes taken to be axiomatic [12], this procedure's validity has been questioned. For example, the recent paper [43] shows that a scoring rule tailored to a stylized betting problem should be *multiplied* over independent assessments.

In statistical settings where a scoring rule is used to fit a distribution to data, a sum or average score is computed over the set of observations [30]. Given a forecast $\mathbf{r}$ and a sequence of N samples of the same uncertainty, the measure

$$N^{-1} \sum_{i=1}^{N} S(\omega(i), \mathbf{r})$$

is used to evaluate the forecast. This cumulative score increases when the Bregman Divergence associated with S between the empirical distribution and the forecasted distribution decreases. Summing the scores over distinct uncertainties, as in Equation (3.1), mirrors this procedure.

There are other theoretical and practical justifications for adding the scores for a sequence of forecasts. From an ex ante perspective, the sequential score given in Equation 3.1 is itself a strictly proper scoring rule with respect to the forecaster's subjective joint distribution [64]. From an ex post perspective, sequentially adding scores is natural, particularly when the forecasts are made over time and a reward is assigned after each forecasting exercise. However, in applications where a joint distribution is elicited, as in many decision analysis applications [23, 18], a logical alternative would be to reward the forecaster by applying a scoring rule directly to the reported joint probability distribution. Scoring the joint distribution could be considered desirable, as this forecast fully captures the expert's knowledge.

The following section presents a series of motivating examples demonstrating that both adding scores and directly scoring the joint distribution can be problematic, depending on the scoring rule chosen. We show that certain rules, such as the popular Quadratic and Spherical scores, provide illogical measures of accuracy for joint forecasts. Under these rules, a *more accurate* marginal forecast can result in a *lower* joint score, even when events are independent. Conversely, when conditional forecasts are submitted for a sequence of dependent uncertainties, we demonstrate that the rankings produced by adding scores fundamentally depend upon arbitrary decisions regarding the order in which forecasts are scored.

Taken together, a set of marginal and conditional assessments and their corresponding joint distribution contain exactly the same probabilistic information. One may therefore expect the *relative ranking* of forecasts produced by a scoring procedure to be invariant to the order of scoring or how the scoring procedure is factored into component distributions, even though the

absolute score assigned to forecasts may differ. For example, it may seem reasonable for a scoring procedure to yield the same ranking of forecasts whether the marginal/conditional distributions are scored or the joint distribution is scored. More generally, it might be considered desirable, or even necessary, for the relative ranking of probabilistic forecasts to depend solely upon the information content of those forecasts and not upon the process by which those forecasts are evaluated.

We develop a set of additive and strongly additive scoring rules for which the distinction between sequential and joint scoring is immaterial. Under these rules, the relative ranking of forecasts does not depend upon the process by which the scoring rule is implemented.

Our primary contributions are as follows.

First, in Section 3.3, we present a series of examples that illustrate the problems that arise when using certain common rules to score multiple forecasts.

In Section 3.4, we introduce and define additive and strongly additive scoring rules. We present methods for constructing additive strictly proper scoring rules, for which the joint score is equal to a sum of marginal scores when events are independent. Our additive rules are well-suited for scoring joint distributions when the underlying dependence structure is known. In particular, the additive structure of these scoring rules guarantees that for independent events, a joint forecast receives a higher reward for more accurate marginal forecasts – a logical consistency property not shared by other common rules. We further demonstrate that the Logarithmic score is the only strongly additive strictly proper scoring rule. The strong additivity of the Logarithmic

score guarantees that sequentially adding logarithmic scores produces the same ranking of forecasts irrespective of how the forecasts are ordered.

Finally, in Section 3.5, we show that additive scoring rules have corresponding generalized entropy measures possessing additive consistency properties considered in the information-theory literature [28]. These entropy functions are invariant to rearrangement and decomposition of the underlying probabilistic information. In addition, we give the necessary conditions under which an additive entropy measure can be used to construct an additive proper scoring rule.

3.3 Problems with Current Rules for Scoring Multiple Assessments

The following examples demonstrate that, when scoring multiple assessments, different scoring procedures can produce divergent and sometimes illogical forecast rankings. Although this is illustrated using the Quadratic and Spherical scoring rules, other scoring rules are not immune. Our examples are divided into two categories. The first involves scoring the forecasts for a collection of mutually independent uncertainties. The second involves scoring forecasts for uncertainties that are probabilistically dependent.

3.3.1 Scoring for Independent Experiments

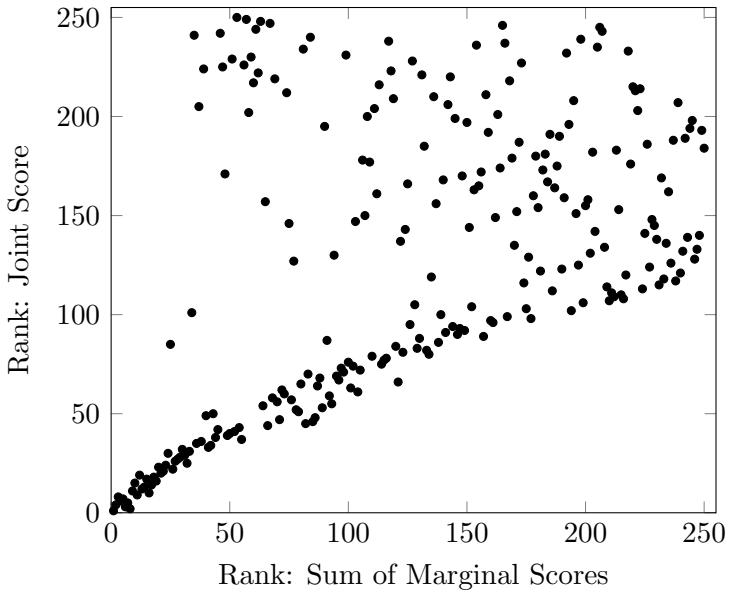
In the case of independence, the sequential scoring procedure in Equation (3.1) corresponds to adding scores for the forecasted marginal distributions $\mathbf{r}^j$. Alternatively, a scoring rule could be applied directly to the implied joint distribution. The rules in Table 1.1 accept probability distributions of varying dimensions and do not discern whether the forecast corresponds to a joint

distribution or a marginal distribution. We compare these two scoring procedures using the forecasts provided by 250 students as part of a decision analysis midterm exam. The test consisted of multiple-choice questions, each with four possible answers. For each question, unlike a traditional multiple-choice exam, the students submitted probability distributions that reflected their beliefs regarding the likelihood of each answer being the correct one. Full details of the exam procedure are given in [15]. We assume that students viewed the questions as probabilistically independent.

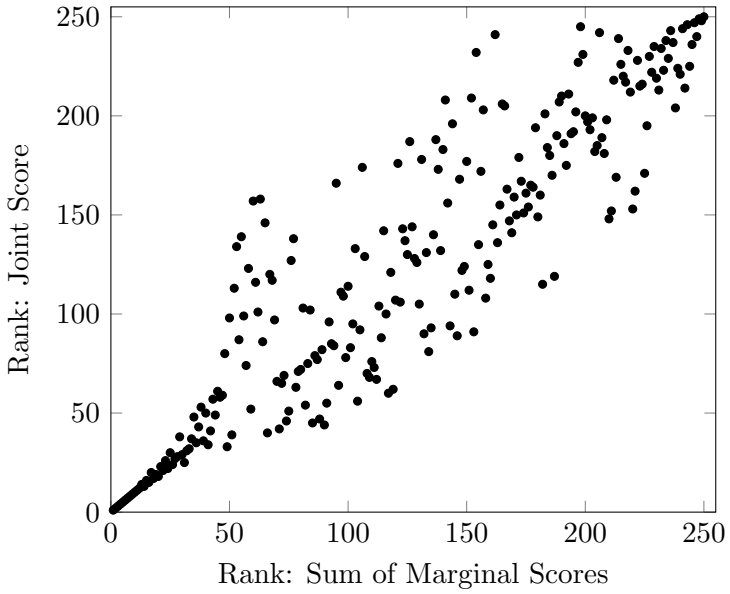
Figure 3.1 plots the student rankings under the marginal scoring procedure and the rankings under the joint scoring procedure using the Quadratic and Spherical scoring rules. A lower numerical ranking corresponds with better performance: a student ranked 1 scored higher than a student ranked 10.

Despite applying the same scoring function, the rankings depend heavily upon the scoring procedure. For example, under the Quadratic scoring rule, in the most extreme case of ranking discrepancy, a student received the 35th highest score (rank) based on his/her marginal forecasts, and the 241st highest score (rank) based on his/her joint forecast. Under Spherical scoring, a different student was ranked 97 places higher using marginal scoring than with joint scoring.

Marginal versus joint scoring clearly can produce very different relative rankings. Although a divergence of rankings does not necessarily imply that one scoring rule is better than another, we contend that in the case of independent uncertainties, scoring the marginal distributions is preferable. Consider the following example. Suppose that a gold coin and a silver coin, both possibly biased, are to be flipped. Each coin can land on heads (H) or tails (T). Assume further that the coin flips are probabilistically independent. Suppose

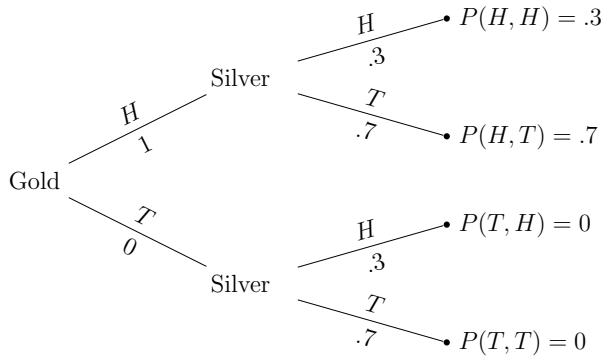


(a) Quadratic Marginal vs. Joint Score Ranking

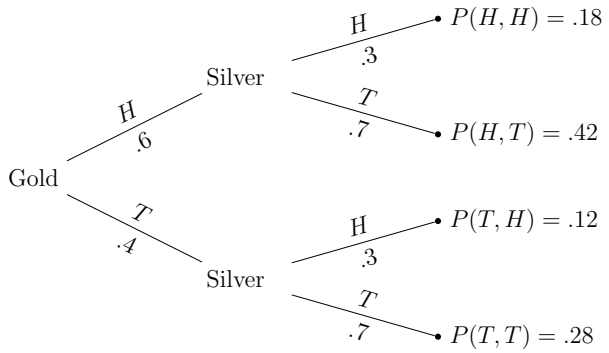


(b) Spherical Marginal vs. Joint Score Ranking

Figure 3.1: Marginal vs. Joint Score Student Rankings



(a) Forecast 1



(b) Forecast 2

Figure 3.2: Probabilities for Independent Coins

that two forecasts are provided for the probability that each coin will land H. These forecasts are presented in Figures 3.2a and 3.2b, respectively.

After flipping, both coins land on H and we score the forecasts using the Quadratic rule. As in the student testing example, we compute the sum of the Quadratic scores for the marginal assessments as well as the Quadratic score for the implied joint assessments. For example, for Forecast 1 the cumulative marginal score is $S_Q(H, [1, 0]) + S_Q(H, [.3, .7]) = 1.02$, and the joint score is $S_Q((H, H), [.3, .7, 0, 0]) = .02$. The scores are presented in Table 3.1.

Table 3.1: Marginal vs. Joint Quadratic Scores for Independent Coin Flip Forecasts

Forecast	Marginal	Joint
Forecast 1	1.02	.02
Forecast 2	.70	.06
Best	Forecast 1	Forecast 2

Using the joint Quadratic score places Forecast 2 ahead of Forecast 1, whereas the the sum of marginal Quadratic scores yields the reverse. In this case, however, only one rank ordering is reasonable. To see why, first observe that the marginal forecasts differ only with respect to the probability that the gold coin will land H, whereas each forecast assigns the same probabilities to the silver coin. Because we observed H upon flipping the gold coin, and given that the experiments are independent, it is intuitive that *any measurement of forecast quality should favor Forecast 1.*

In our example, Forecast 1's joint probability for the realized outcome (H, H) exceeds the corresponding probability in Forecast 2. Nevertheless, Forecast 2 is awarded a higher joint score because the Quadratic scoring rule penalizes Forecast 1's concentration of probability on a single incorrect answer: (H, T). However, Forecast 1's concentration of probability on this event is a direct result of a *more accurate* marginal forecast.

Thus, the Quadratic scoring rule, when applied directly to a joint forecast, has an illogical property: for independent events, a better marginal forecast does not imply a higher joint score. As illustrated in Figure 3.3, although the

cumulative marginal score is monotonic in the probability assigned to the gold coin landing H (holding fixed the silver coin probabilities at the values given in our example), the joint Quadratic score is not. This characteristic of the joint Quadratic score seems problematic. For independent binary events, the score should increasing in the marginal probability placed on the observed marginal outcome, holding all other values fixed. Any rule failing this criterion will appear unfair to the forecaster and provide a poor signal of forecast accuracy given a fixed observation.

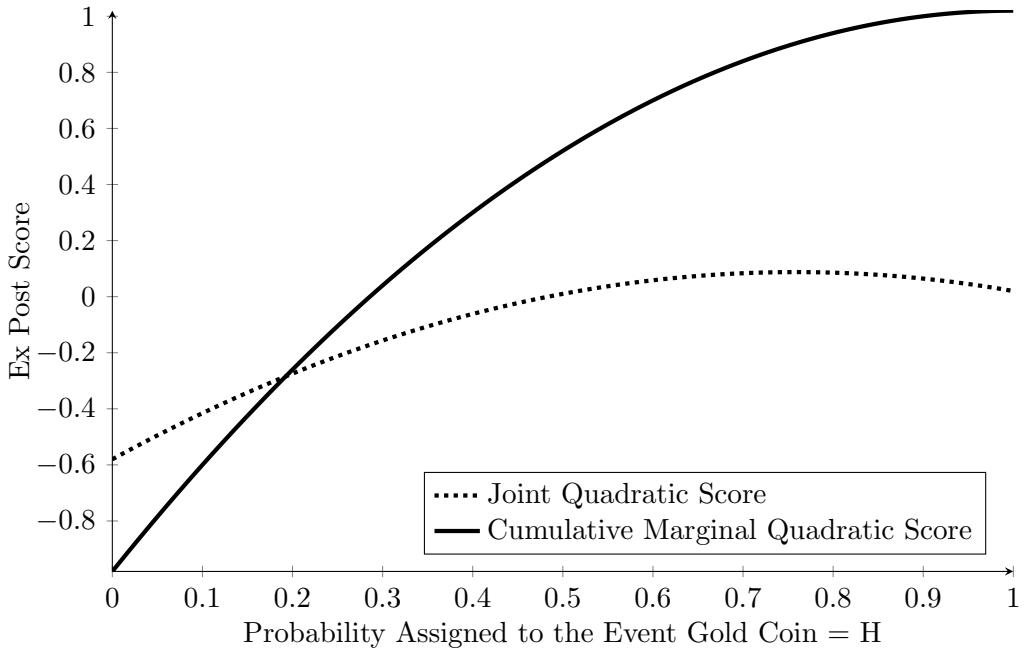


Figure 3.3: Monotonicity of Joint Quadratic Scores and Cumulative Marginal Quadratic Scores

Up to this point in our coin flipping example, we have considered forecast rankings after only a single observation. Since the obtained score itself is a random variable, determination of the more accurate forecast requires re-

peated sampling from the "true" distribution. One would then hope, under a reasonable score function, that the long-run average score would ultimately indicate who made a forecast closer to the truth.

A counter-argument is that in a subjective forecasting exercise, illogical ranking under one sample suffices to make the scoring procedure appear unfair. This assertion is especially salient because many forecasting applications do not allow for multiple observations of the experiment. For example, in weather forecasting, sports betting, election prediction, etc., there will necessarily only be a single observation under which a forecast's accuracy can be measured. Moreover, in applications such as student testing, there is no notion of a non-degenerate true sampling distribution. Putting aside these arguments, as we now show, the *average* joint Quadratic score taken over repeated samples from a non-degenerate empirical distribution may still rank the forecasters in an illogical order.

Returning to our coin flip example, suppose that, in truth, the probability of the gold coin landing H is 1 (i.e., Forecast 1's marginal prediction is exactly correct) and the corresponding true probability for the silver coin is .9, so that the true joint sampling distribution is $\mathbf{p}^* = [.9, .1, 0, 0]$. Clearly, Forecast 1 more accurately describes the true distribution: the marginal prediction for the gold coin corresponds exactly with the truth whereas Forecast 2's does not. However, applying the strong law of large numbers, the average joint score for Forecast 1 converges to $E_{\mathbf{p}^*}[S_Q(\omega, \mathbf{r})] = .1$, whereas the average score for Forecast 2 converges to .106. Since Forecast 2 still receives the higher score, this demonstrates that the joint Quadratic score does not provide a reasonable measure of closeness between the true joint distribution and the reported joint forecasts. In other words, the expected score function $E_{\mathbf{p}^*}[S_Q(\omega, \mathbf{r})]$ has inher-

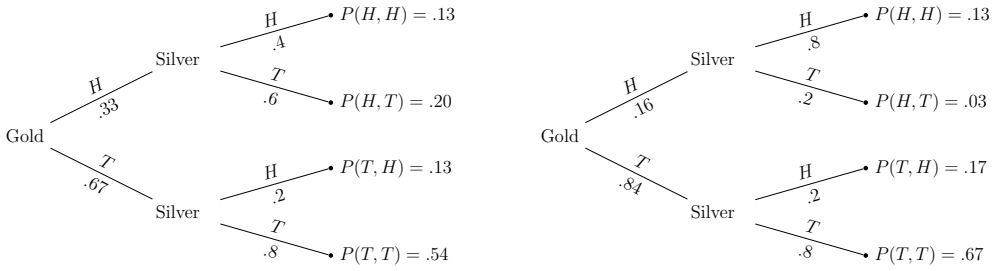
ited the poor properties of the score function S. In Section 3.5, we elaborate on the connection between the ex post score function and the expected score in the context of multiple forecasts.

The above makes clear that applying certain scoring rules to the joint forecast can be problematic. This suggests that the joint distribution should be factored into marginal distributions (or, in the case of dependence, marginal and conditional distributions) and then sequentially scored. However, as we demonstrate in the following section, scoring the decomposed joint distribution can also lead to results that could be viewed as problematic.

3.3.2 Sequential Scoring for Dependent Assessments

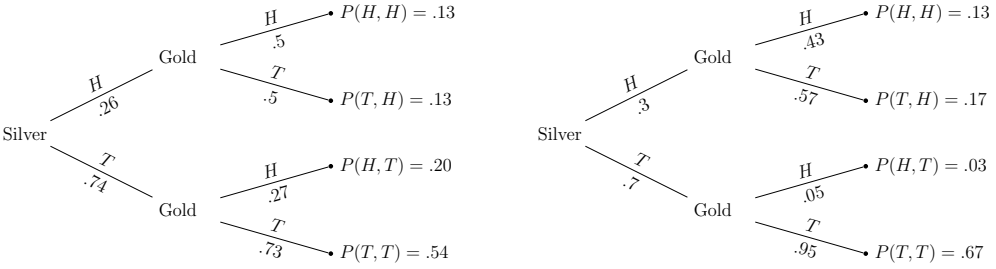
When probabilistic dependence is involved, the sequential scoring procedure in Equation (3.1) corresponds with adding the scores for a marginal distribution and a sequence of conditional distributions. The specific conditional distributions scored depends on the realized outcome ω. We do not evaluate predictions conditioned on an unobserved event. We return to our coin flip example to show that forecast rankings under this sequential scoring procedure depend on the order in which the assessments are conditioned (or evaluated).

Once again, suppose that two coins are flipped. However, in contrast to the previous example, we now assume that the probability of the silver coin landing H is dependent on the result of the gold flip. As before, two probability assessments are provided for each coin. These forecasts are presented in Figures 3.4a and 3.4b, where the probabilities for the silver coin are conditioned on the outcome of the gold coin flip. Alternatively, the same information can be expressed in terms of the marginal probabilities for the silver coin and the



(a) Forecast 1 (b) Forecast 2

Figure 3.4: Probabilities for Dependent Coins



(a) Forecast 1 (b) Forecast 2

Figure 3.5: Flipped Tree Conditional Probabilities

conditional probabilities for the gold coin, as given in Figures 3.5a and 3.5b.

We observe the joint outcome (H, H), and we adopt the sequential scoring procedure, this time using the Spherical scoring rule. For each forecast, we sum the scores for the marginal forecast for the gold coin, and the conditional forecast for the silver coin, given a realization of H for the gold coin. We then score each forecast after reversing the order of conditioning, that is, we score the marginal forecast for the silver coin and the conditional forecast for the gold coin. For example, conditioning on the gold coin, Forecast 1 receives the score

$$S_S\left(H, [.33, .67]\right) + S_S\left(H, [.4, .6]\right) = 1.00.$$

Alternatively, conditioning on the silver coin, Forecast 1 receives a score of

$$S_S\left(H, [.26, .74]\right) + S_S\left(H, [.5, .5]\right) = 1.04.$$

Table 3.2 summarizes the resulting Spherical scores under each procedure.

Table 3.2: Marginal/Conditional vs. Joint Spherical Scores for Dependent Coin Flip Forecasts

Forecast	$G,\ S\|G$	$S,\ G\|S$	Joint
Forecast 1	1.00	1.04	.22
Forecast 2	1.16	.99	.18
Best	Forecast 2	Forecast 1	Forecast 1

The forecast rankings are shown to be dependent upon the order of conditioning. On the one hand, if we first score the marginal assessment for the gold coin (G) and then score the assessment for the silver coin (S) conditional on G ($S|G$), Forecast 2 obtains the highest cumulative score. On the other hand, if the order of conditioning and scoring is reversed, then Forecast 1 obtains the highest cumulative score.

This discrepancy can be understood as follows. When scoring a sequence of conditional forecasts under the Spherical rule, the order of conditioning is essentially a choice of what information from the joint distribution, corresponding with unobserved outcomes, will be used in computing the score. For example, when the forecast for the silver coin is conditioned on the outcome of the gold coin, the sequential score depends on the probability $P(H, H)$ (the observed outcome) and the probability $P(H, T)$ (unobserved). When the order of conditioning is reversed, the sequential score depends on $P(H, H)$ and

$P(T, H)$. The joint Spherical score, meanwhile, depends on all four components of the joint forecast. Thus, the order of conditioning arbitrarily *localizes* the measurement of forecast accuracy. This causes the joint Spherical or Quadratic score to fundamentally differ from the marginal/conditional score. It also means that the marginal/conditional rankings may contradict one another depending on the order in which the forecasts are scored.

In real-world forecasting situations, where the cumulative or average score is used as a measurement of forecast accuracy, sensitivity to the way in which information is decomposed could be seen as problematic. Practitioners should be aware that scores under some rules, such as the Spherical and Quadratic, are dependent on not just the chosen scoring rule (i.e., the functional form of S) and the forecaster's information, but also upon the order of scoring – giving rankings a degree of arbitrariness that may be difficult to explain.

In decision analysis applications, where obtaining a joint forecast is often performed sequentially, it is most practical to first elicit a marginal assessment and subsequently a set of conditional assessments in an order designed to facilitate the elicitation process. If a sequential scoring procedure is adopted, it would then seem natural to sum scores in the same order. However, the assessor and the forecasters need to understand that the ranking of competing forecasts depends upon the scoring procedure itself, which is independent of the underlying information.

One possible reservation about this finding is that the example above might be unusual in producing rankings that vary according to the order of conditioning. The following section tests this hypothesis.

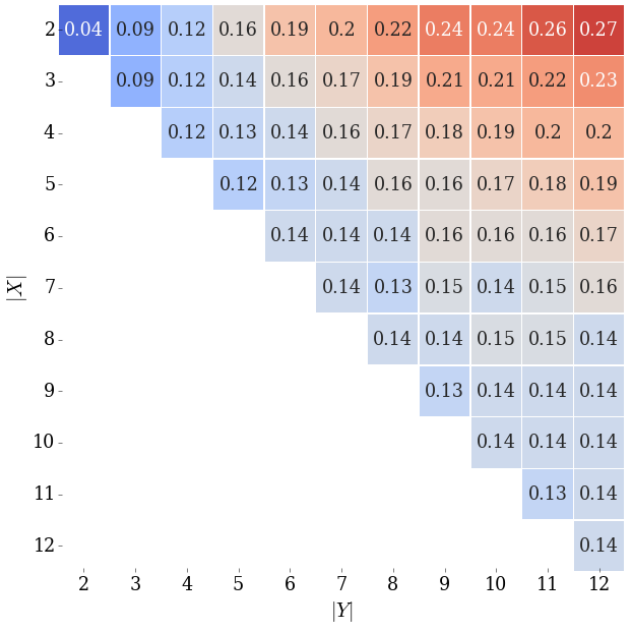
3.3.3 Sensitivity of Rankings for Sequential Scoring of Conditional Assessments

Let X, Y be two discrete uncertainties taking on $|X|$ and $|Y|$ possible outcomes. Suppose that two competing experts provide joint distribution forecasts $\mathbf{r}(1), \mathbf{r}(2)$ for (X, Y). From this information, we form the conditional and marginal distributions $\mathbf{r}^X(j),\ \mathbf{r}^{Y|X}(j),\ \mathbf{r}^Y(j),\ \mathbf{r}^{X|Y}(j)$ for $j = 1, 2$. Assume without loss of generality that $(X, Y) = (1, 1)$ is observed. For a given scoring rule S, if we find that

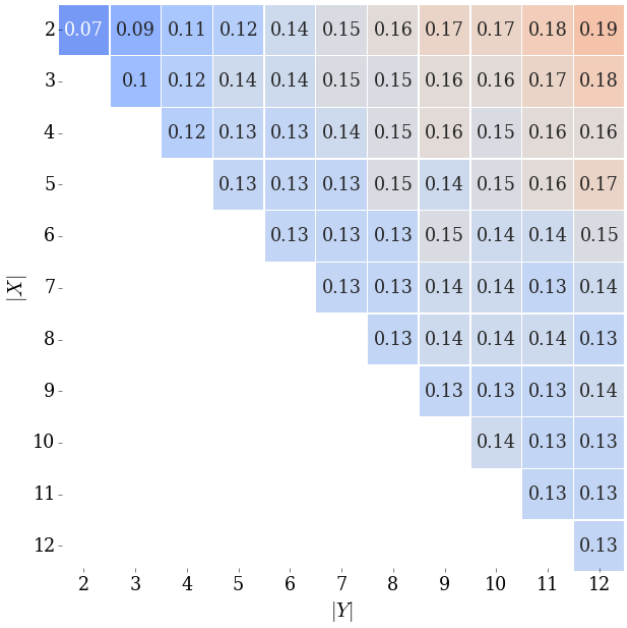
$$\begin{aligned} &\left[S\left(1, \mathbf{r}^X(1)\right) + S\left(1, \mathbf{r}^{Y|X=1}(1)\right) - S\left(1, \mathbf{r}^X(2)\right) + S\left(1, \mathbf{r}^{Y|X=1}(2)\right)\right] \\ \times &\left[S\left(1, \mathbf{r}^Y(1)\right) + S\left(1, \mathbf{r}^{X|Y=1}(1)\right) - S\left(1, \mathbf{r}^Y(2)\right) + S\left(1, \mathbf{r}^{X|Y=1}(2)\right)\right] \\ &< 0, \end{aligned} \tag{3.2}$$

then the ranking of the experts (under a sequential additive scoring procedure) is contingent on the order of conditioning for X and Y.

We estimate the frequency of these rank-order reversals through simulation. To this end, we sample pairs of joint distributions, $\mathbf{r}(1)$ and $\mathbf{r}(2)$, independently and uniformly from the probability simplex. For fixed $|X|$ and $|Y|$, we draw $10,000$ pairs $(\mathbf{r}(1), \mathbf{r}(2))$ and calculate the proportion of samples satisfying Equation (3.2), which are those whose rankings are reversed with the order of conditioning. These proportions, for varying dimensions of X and Y, are reported in Figure 3.6 for the Quadratic and Spherical scoring rules. As the dimensions of X and Y become more disparate, the proportion of samples that exhibit ranking inconsistency is seen to increase. In the case of $|X| = 2, |Y| = 2$, ranking changes are observed in only 4% of samples when using the Quadratic scoring rule. However, with $|X| = 2, |Y| = 12$, that proportion rises to approximately 27%. A similar pattern emerges when Spherical scoring is used. These results suggest that ranking inconsistencies



(a)



(b)

Figure 3.6: Proportion of samples with rank-reversal under the (a) Quadratic scoring rule, and (b) Spherical scoring rule, for varying cardinalities of X, Y

may arise in many real-world applications when conditional distributions are scored sequentially.

We now turn to the task of identifying and constructing scoring rules that are immune to the inconsistencies found in our motivating examples.

3.4 Additive and Strongly Additive Scoring Rules

This section describes *additive* and *strongly additive* strictly proper scoring rules, which have desirable properties for scoring multiple assessments. Recall that for J uncertainties, we write Ω to denote the joint sample space, with outcomes $\omega = (\omega^1, \omega^2, \ldots, \omega^j)$, and $\mathcal{P}$ to represent the set of all joint distributions over Ω. Before proceeding, we define $\mathcal{Q} \subset \mathcal{P}$ to be the set of joint distributions corresponding with mutual independence between the j uncertainties:

$$\mathcal{Q} := \left\{ \mathbf{q} \in \mathcal{P} : q_\omega = \prod_{j \in [J]} q^j_{\omega^j} \text{ for all } \omega \in \Omega \right\}.$$

For notational brevity, we write $\mathbf{q} = \prod_{j \in [J]} \mathbf{q}^j$, for $\mathbf{q} \in \mathcal{Q}$. For additive scoring rules, we adapt the definition given in [64].

Definition 3.1. *Additive Scoring Rule*

We say a scoring rule S is additive if S is strictly proper on $\mathcal{P}$ and, for strictly proper scoring rules S^j on $\mathcal{P}^j$, $j \in [J]$, we have

$$S(\omega, \mathbf{q}) = \sum_{j=1}^{J} S^j(\omega^j, \mathbf{q}^j) \tag{3.3}$$

for all $\omega \in \Omega$ and $\mathbf{q} \in \mathcal{Q}$.

An additive scoring rule incentivizes truth-telling by a forecaster with respect to his or her knowledge of the joint distribution and admits a sequential,

additive scoring procedure when the forecast implies independence. Parry [64] originally introduced such rules for use in parameter estimation when the scoring rules S and S^j, $j \in [J]$ are applied to density functions. Additive scoring rules, as measurements of the quality of the forecast $\mathbf{q} \in \mathcal{P}$, have an attractive ex post consistency property: for independent uncertainties, an additive scoring rule S will always assign a higher score for the joint forecast when a more accurate forecast was made for the marginal prediction $\mathbf{q}^j$. In the coin flip example in the previous section, we observed that the Quadratic (and Spherical) scoring rules, when applied directly to a joint distribution, do not satisfy this requirement.

Strongly additive rules, defined below, have the property that the joint score is equal to a sum of conditional scores, *independent* of the order in which the uncertainties are assessed.

Definition 3.2. *Strongly Additive Scoring Rule*

A scoring rule S is strongly additive if S is strictly proper on $\mathcal{P}$ and, for strictly proper scoring rules $S^{\pi(j)}$ on $\mathcal{P}^{\pi(j)}$, $j \in [J]$, we have

$$S(\omega, \mathbf{r}) = \sum_{j=2}^{J} S^{\pi(j)}(\omega^{\pi(j)}, \mathbf{r}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}}) + S^{\pi(1)}(\omega^{\pi(1)}, \mathbf{r}^{\pi(1)}) \tag{3.4}$$

for all $\omega \in \Omega$ and $\mathbf{r} \in \mathcal{P}^+$, and for every arbitrary permutation $\pi(\cdot)$ of the indices $1, \ldots, J$.

Put simply, strongly additive scoring rules reward equivalent information, encoded in a joint distribution, equivalently, over any decomposition into conditional distributions. In applications where competing experts submit a sequence of dependent forecasts, using a scoring rule that is strongly additive

guarantees that scores are independent of the order in which the experiments are performed. In contrast, our previous examples showed that scores under the Quadratic and Spherical rules depend upon the ordering of the assessments. Requiring Definition 3.2 to hold only for $\mathbf{r} \in \mathcal{P}^+$ ensures that the associated conditional distributions are well defined. A strongly additive rule also necessarily satisfies the weaker additivity requirement given in Definition 3.1 on $\mathcal{P}^+$.

Example 3.1. *An example of a strongly additive (and additive) scoring rule is the Logarithmic scoring rule. For any* $\mathbf{r} \in \mathcal{P}^+$ *and permutation* $\pi(\cdot)$*, we have:*

$$\begin{aligned}\ln(r_\omega) &= \ln\left(\prod_{j=2}^{J} r_{\omega^j}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}} \times r_{\omega^1}^{\pi(1)}\right) \\ &= \sum_{j=2}^{J} \ln\left(r_{\omega^j}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}}\right) + \ln\left(r_{\omega^1}^{\pi(1)}\right).\end{aligned}$$

In this example, the functional form of the scoring rules on the right-hand and left-hand sides of Equation (3.1) are equivalent: the joint Logarithmic score is equal to the sum of the Logarithmic scores for the conditional distributions. However, we note that the scoring rules S and $S^j, j \in [J]$ in Definitions 3.1 and 3.2 are allowed to take different functional forms. In Section 3.4.2, we will demonstrate that under mild assumptions, the Logarithmic scoring rule $S(\omega, \mathbf{p}) = \ln(p_\omega)$ is the only strictly proper rule, up to equivalence, that satisfies the strong additivity property.

3.4.1 Constructing Additive Scoring Rules

This section provides a method for constructing additive scoring rules. Suppose we are given a set of strictly proper scoring rules:

$$S^j : \Omega^j \times \mathcal{A}^j \to \mathbb{R},\ j \in [J].$$

It has been shown that a sequential rule of the form

$$S(\omega, \mathbf{r}) := \sum_{j \in [J]} S^j \left(\omega^j, \mathbf{r}^{j|\omega^{j-1},\ldots,\omega^1} \right)$$

will be strictly proper and additive [64]. In applications where forecasts are provided over time and scored in sequence, this implies that the sequential score will assign a higher total score to better marginal predictions.

However, when a joint forecast will be directly provided, applying the sequential score requires decomposition into marginal and conditional distributions. In the process, one must choose an order of conditioning, which can alter the score. Thus, in this case, we would like additive scoring rules that can be applied directly to a joint forecast.

We construct a scoring rule S on $\mathcal{A}^+$ (the interior of $\mathcal{A}$) as follows. (We restrict attention to component-wise positive forecasts to avoid dividing by 0.) Let $\mathbf{r} \in \mathcal{A}^+$. For each joint outcome $\omega = (\omega^1, \omega^2, \ldots, \omega^j)$, there is an associated entry in $\mathbf{r}$, $r_{(\omega^1,\ldots,\omega^j)}$. Define the vector $\mathbf{r}(j, \omega) \in \mathbb{R}^{n^j}$, whose kth component is

$$\mathbf{r}(j, \omega)_k := r_{(\omega^1,\ldots,\omega^{j-1},\omega_k^j,\omega^{j+1},\ldots,\omega^j)}$$

for $k = 1, \ldots, n^j$. In other words, $\mathbf{r}(j, \omega)$ is the vector containing the elements of $\mathbf{r}$ formed by varying the jth component of ω within Ω^j, while holding ω^i fixed for all $i \neq j$. We then define S as

$$S(\omega, \mathbf{r}) := \sum_{j \in [J]} S^j(\omega^j, \mathbf{r}(j, \omega)). \tag{3.5}$$

For these rules, the score $S(\omega, \mathbf{r})$ depends on only some components of the joint distribution vector $\mathbf{r}$. In particular, the rule defined in Equation (3.5) is equivalent to the example of a discrete local scoring rule described in Eq. 13 of [31].

When $\mathbf{r} \in \mathcal{P}^+$, we have

$$\mathbf{r}(j, \omega) = \\ P(\omega^1 = \omega^1, \ldots, \omega^{j-1} = \omega^{j-1}, \omega^{j+1} = \omega^{j+1}, \ldots, \omega^n = \omega^n)\mathbf{r}^{j|\omega^1,\ldots,\omega^{j-1},\omega^{j+1},\ldots,\omega^j}, \tag{3.6}$$

where the probability on the right-hand side is computed with respect to $\mathbf{r}$. In other words, the vector $\mathbf{r}(j, \omega)$ is proportional to the conditional distribution for the jth uncertainty given observations $\omega^i = \omega^i$ for $i \neq j$. We now demonstrate that S, as defined in Equation (3.5), is additive and strictly proper.

Theorem 3.1. *The scoring rule given in Equation* (3.5) *is strictly proper on* $\mathcal{P}^+$ *and, for* $\mathbf{q} \in \mathcal{Q}^+$*, satisfies*

$$S(\omega, \mathbf{q}) = \sum_{j \in [J]} S^j(\omega^j, \mathbf{q}^j).$$

Proof. From Equation (3.6) and the fact that the rules S^j, $j \in [J]$ are 0-homogeneous, it follows that for a general $\mathbf{r} \in \mathcal{P}^+$

$$S(\omega, \mathbf{r}) = \sum_{j \in [J]} S^j(\omega^j, \mathbf{r}^{j|\omega^1,\ldots,\omega^{j-1},\omega^{j+1},\ldots,\omega^j}),$$

which immediately implies that S is additive. We now demonstrate strict propriety on $\mathcal{P}^+$. To this end, let $\mathbf{p}, \mathbf{r} \in \mathcal{P}^+$. We have

$$E_{\mathbf{p}}[S(\omega, \mathbf{r})] = \sum_{\omega \in \Omega} p_\omega \left\{ \sum_{j \in [J]} S^j \left[\omega^j, \mathbf{r}(j, \omega)\right] \right\}. \tag{3.7}$$

Let $\Omega^{/j} := \Omega^1 \times \ldots \times \Omega^{j-1} \times \Omega^{j+1} \times \ldots \times \Omega^J$. Rearranging terms on the right-hand side of Equation (3.7) yields:

$$\begin{aligned} E_{\mathbf{p}}[S(\omega, \mathbf{r})] &= \sum_{j \in [J]} \left\{ \sum_{(\omega^1, \ldots, \omega^{j-1}, \omega^{j+1}, \ldots, \omega^j) \in \Omega^{/j}} \mathbf{p}(j, \omega)^\top S^j \left[\omega^j, \mathbf{r}(j, \omega)\right] \right\} \\ &\leq \sum_{j \in [J]} \left\{ \sum_{(\omega^1, \ldots, \omega^{j-1}, \omega^{j+1}, \ldots, \omega^j) \in \Omega^{/j}} \mathbf{p}(j, \omega)^\top S^j \left[\omega^j, \mathbf{p}(j, \omega)\right] \right\} \\ &= E_{\mathbf{p}}[S(\omega, \mathbf{p})], \end{aligned} \tag{3.8}$$

where the inequality follows from strict propriety of the S^j.

We need to show that this inequality holds strictly when $\mathbf{r} \neq \mathbf{p}$. To obtain a contradiction, suppose $E_{\mathbf{p}}[S(\omega, \mathbf{r})] = E_{\mathbf{p}}[S(\omega, \mathbf{p})]$ for some $\mathbf{r} \neq \mathbf{p}$. By strict propriety of S^j, $j \in [J]$, it follows from Equation (3.8) that $\mathbf{p}$ and $\mathbf{r}$ must satisfy the system of equations:

$$\begin{aligned} \mathbf{p}(j, \omega)/\mathbf{1}^\top \mathbf{p}(j, \omega) = \mathbf{r}(j, \omega)/\mathbf{1}^\top \mathbf{r}(j, \omega), \\ \forall j, (\omega^1, \ldots, \omega^{j-1}, \omega^{j+1}, \ldots, \omega^j) \in \Omega^{/j}. \end{aligned} \tag{3.9}$$

From Equation (3.9), we see that if $\omega_1, \omega_2 \in \Omega$ differ in one dimension, that is, $\omega_1^j \neq \omega_2^j$, and $\omega_1^i = \omega_2^i$ for all $i \neq j$, then we have:

$$\frac{p_{\omega^1}}{r_{\omega^1}} = \frac{p_{\omega^2}}{r_{\omega^2}}.$$

But then, by repeating this argument one component (or dimension) at a time, we obtain

$$\frac{p_\omega}{r_\omega} = \frac{p_\eta}{r_\eta}$$

for any $\omega, \eta \in \Omega$. This implies that $\lambda \mathbf{r} = \mathbf{p}$ for some $\lambda > 0$. Imposing the constraint $\mathbf{1}^\top \mathbf{r} = 1$, we must have $\mathbf{r} = \mathbf{p}$, contradicting our assertion that $\mathbf{r} \neq \mathbf{p}$. □

We apply the construction method given above to derive the additive scoring rules given in the following examples.

Example 3.2. *Additive Quadratic Scoring Rule*

For each j, choose S^j to be the Quadratic scoring rule:

$$S^j(\omega, \mathbf{r}) := 2\left(\frac{r_\omega}{\mathbf{1}^\top \mathbf{r}}\right) - \frac{||\mathbf{r}||_2^2}{[\mathbf{1}^\top \mathbf{r}]^2}$$

for $j \in [J]$. Then

$$S(\omega, \mathbf{r}) = \sum_{j=1}^{J} \left\{ 2\left(\frac{r_\omega}{\mathbf{1}^\top \mathbf{r}(j,\omega)}\right) - \frac{||\mathbf{r}(j,\omega)||_2^2}{[\mathbf{1}^\top \mathbf{r}(j,\omega)]^2} \right\}$$

is strictly proper on $\mathcal{P}^+$ and for $\mathbf{q} \in \mathcal{Q}^+$:

$$S(\omega, \mathbf{q}) = \sum_{j \in [J]} 2q^j_{\omega^j} - ||\mathbf{q}^j||_2^2.$$

Example 3.3. *Additive Logarithmic Scoring Rule*

Let $S^j(\omega, \mathbf{r}) = \ln(r_\omega / \mathbf{1}^\top \mathbf{r})$ for $j \in [J]$. Then

$$S(\omega, \mathbf{r}) = \sum_{j=1}^{J} \ln\left(\frac{r_\omega}{\mathbf{1}^\top \mathbf{r}(j,\omega)}\right)$$

is strictly proper on $\mathcal{P}^+$ and for $\mathbf{q} \in \mathcal{Q}^+$:

$$S(\omega, \mathbf{q}) = \sum_{j \in [J]} \ln\left(q^j_{\omega^j}\right). \tag{3.10}$$

In particular, Example 3.3 demonstrates that S is not unique given our choice of $S^j, j \in [J]$, since the simple Logarithmic rule $S(\omega, \mathbf{q}) = \ln(q_\omega)$ for $\mathbf{q} \in \mathcal{Q}$ will also satisfy Equation (3.10). However, for a general $\mathbf{r} \in \mathcal{P}$, we have

$$\sum_{j=1}^{J} \ln\left(\frac{r_\omega}{\mathbf{1}^\top \mathbf{r}(j,\omega)}\right) \neq \ln(r_\omega).$$

3.4.1.1 Numerical Example

This numerical example applies the Additive Quadratic scoring rule. We return to the independent coin flip example of Section 3.3. Suppose that we obtain the forecast $\mathbf{r} = [.18, .42, .12, .28]$ for the events

$$[(H,H),(H,T),(T,H),(T,T)].$$

The forecast $\mathbf{r}$ implies that the coin flips are independent. As usual, we assume (H,H) is the observed outcome.

To compute the score, we form $\mathbf{r}(j,(H,H))$ for $j = 1,2$, where $j = 1$ corresponds with the gold coin and $j = 2$ the silver coin. For $j = 1$,

$$\mathbf{r}(1,(H,H)) = [P(H,H) = .18, P(T,H) = .12],$$

where $\mathbf{r}(1,(H,H))$ includes the components of the joint distribution found by varying the outcome of the gold coin from the observed value. We similarly find

$$\mathbf{r}(2,(H,H)) = [P(H,H) = .18, P(H,T) = .42]$$

by varying the outcome of the silver coin. Neither vector includes the probability $P(T,T) = .28$ corresponding with the unobserved outcome for *both* the gold and silver coins. The Additive Quadratic score for the joint forecast is expressed as

$$\begin{aligned} S((H,H),\mathbf{r}) &= \left\{2\left(\frac{.18}{\mathbf{1}^\top[.18,.12]}\right) - \frac{||[.18,.12]||_2^2}{[\mathbf{1}^\top[.18,.12]]^2}\right\} \\ &\quad + \left\{2\left(\frac{.18}{\mathbf{1}^\top[.18,.42]}\right) - \frac{||[.18,.42]||_2^2}{[\mathbf{1}^\top[.18,.42]]^2}\right\} \\ &= 2\,(.6) - ||[.6,.4]||_2^2 + 2\,(.3) - ||[.3,.7]||_2^2 \\ &= .7. \end{aligned}$$

The resulting score corresponds exactly with the sum of Quadratic scores for the implied marginal predictions $\mathbf{r}^1 = [.6, .4]$ for the gold coin and $\mathbf{r}^2 = [.3, .7]$ for the silver coin, as calculated in Table 3.1. Unlike the traditional Quadratic scoring rule, the Additive Quadratic scoring rule rewards a higher joint score for more accurate marginal predictions when the coins are independent.

We now compare scores under the Additive Quadratic scoring rule with those obtained in the sequential scoring procedure in Equation (3.1). Suppose that we will score the forecast $\mathbf{r} = [.13, .20, x, .67 - x]$ for the events $[(H,H),(H,T),(T,H),(T,T)]$, which, for $x = .13$ is exactly Forecast 1 given in Figure 3.4a. We will vary x and compare the resulting scores. Suppose that (H,H) is the observed outcome.

The sequential score of Equation (3.1) can be computed differently depending on whether we condition upon the outcome of the gold coin or that of the silver. We can also produce a joint score using the Additive Quadratic rule given in Example 3.2. These different scores, as functions of the joint probability distribution, are calculated as follows:

Sequential S|G :

$$2(.33) - \|[.33, .67]\|_2^2 + 2\left(\frac{.13}{\mathbf{1}^\top[.13, .20]}\right) - \left\|\frac{[.13, .20]}{\mathbf{1}^\top[.13, .20]}\right\|_2^2$$

Sequential G|S :

$$2(.13 + x) - \|[.13 + x, .67 - x + .20]\|_2^2 + 2\left(\frac{.13}{\mathbf{1}^\top[.13, x]}\right) - \left\|\frac{[.13, x]}{\mathbf{1}^\top[.13, x]}\right\|_2^2$$

Additive Quadratic :

$$2\left(\frac{.13}{\mathbf{1}^\top[.13, .20]}\right) - \left\|\frac{.13, .20}{\mathbf{1}^\top[.13, .20]}\right\|_2^2 + 2\left(\frac{.13}{\mathbf{1}^\top[.13, x]}\right) - \left\|\frac{[.13, x]}{\mathbf{1}^\top[.13, x]}\right\|_2^2.$$

Each scoring rule satisfies the additivity requirement. Figure 3.7 plots the scores calculated under each method as a function of the probability $x = P(T, H)$. The Sequential S|G score is .37 irrespective of x. In contrast, the Sequential G|S score does depend on x and exceeds .37 except on the (approximate) interval $(.2, .27)$. The Additive Quadratic score decreases monotonically in x, because shifting weight to the probability $x = P(\text{Gold} = T, \text{Silver} = H)$ corresponds to a less accurate conditional forecast for $P(\text{Gold} = H|\text{Silver} = T)$. When $x \approx .264$, the forecast $\mathbf{r} = [.13, .20, x, .67 - x]$ implies independence between the gold and silver coin, and all three scoring methods produce the same result.

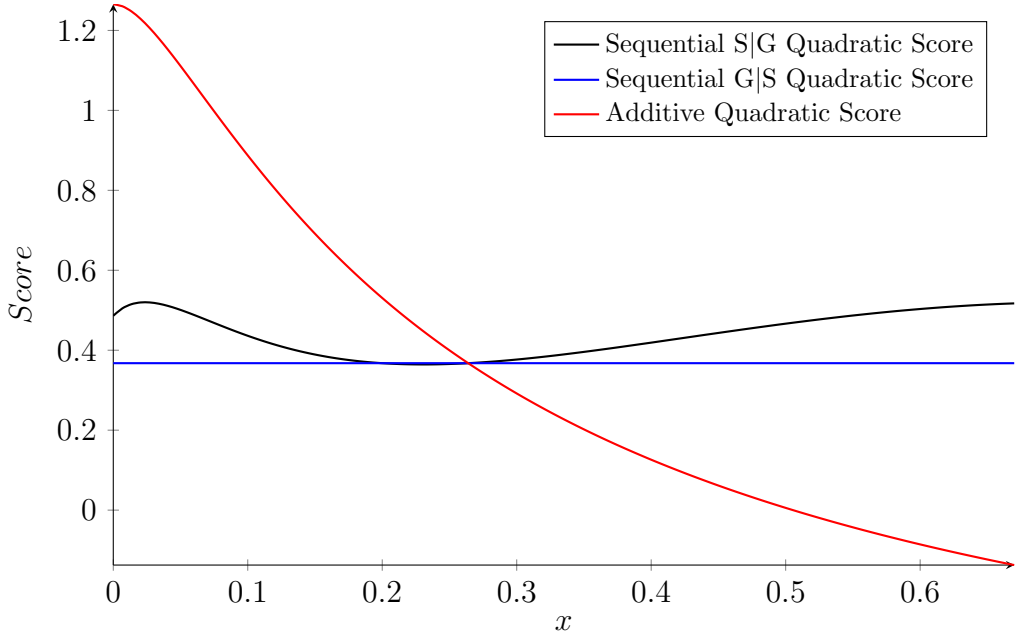


Figure 3.7: Sequential Quadratic Scores vs. Additive Quadratic Score for $\mathbf{r} = [.13, .20, x, .67 - x]$

The locality of the Additive Quadratic score ensures that a forecast is

scored based only on the relative magnitudes of the joint probabilities that correspond to at least one observed marginal outcome.

However, one issue with the additive scoring rules described above is that increasing the weight of the forecast corresponding to the realized outcome may not increase the score. For example, for the coin flip scenario, consider the forecasts

$$\mathbf{r}^1 = [.1, .1, .1, .7]$$

and

$$\mathbf{r}^2 = [.2, .2, .2, .4].$$

Clearly, the second forecast appears to be more accurate given the observation (H, H). Not only is the probability higher for the observed joint outcome, the probabilities placed on joint outcomes corresponding to one of the coins landing heads are also larger. However, it can be seen that the two forecasts receive equivalent scores under the Additive Quadratic scoring rule.

A simple corrective is to add a Logarithmic term to the score function:

$$S(\omega, \mathbf{r}) := \theta \ln(r_\omega) + (1-\theta) \sum_{j=1}^{J} \left\{ 2 \left(\frac{r_\omega}{\mathbf{1}^\top \mathbf{r}(j,\omega)} \right) - \frac{||\mathbf{r}(j,\omega)||_2^2}{[\mathbf{1}^\top \mathbf{r}(j,\omega)]^2} \right\},$$

where $0 < \theta < 1$ is a weighting parameter. This rule is additive and strictly proper, and guarantees a higher score when a greater probability is placed on the observed outcome. Of course, the same adjustment can be made to any of the additive scoring rules constructed in this section.

We now turn our attention to the identification of strongly additive scoring rules.

3.4.2 Strongly Additive Scoring Rules and the Logarithmic Scoring Rule

Having presented a method for constructing a wide variety of additive scoring rules, we now demonstrate that, under mild assumptions, there is a single scoring rule that satisfies the strong additivity property. The remainder of this section will consider only scoring rules $S(\omega, \mathbf{r})$ that are continuous on $\mathcal{P}$ and differentiable in their second argument. The main result of the section is given below.

Theorem 3.2. *Let S and $S^j, j \in [J]$ be strictly proper scoring rules on $\mathcal{P}$ and $\mathcal{P}^j, j \in [J]$, respectively, that satisfy the strong additivity criteria (Definition 3.2). Then $S(\omega, \mathbf{r}) = a \ln(r_\omega) + b$, where $a > 0$ and $b \in \mathbb{R}$ are arbitrary constants.*

We will use the following result [4, 13, 30].

Theorem 3.3. *Suppose $|\Omega| > 2$, and let $S : \Omega \times \mathcal{P}^+ \to \mathbb{R} \cup -\infty$ be a strictly local strictly proper scoring rule, in the sense that we have*

$$S(\omega, \mathbf{r}) = f(r_\omega)$$

for some function $f : (0, 1) \to \mathbb{R} \cup -\infty$. Then $S(\omega, \mathbf{r}) = \ln(r_\omega)$ up to additive and multiplicative constants.

We will also need the following lemma.

Lemma 3.1. *Let S and $S^j, j \in [J]$ be scoring rules on $\mathcal{P}, \mathcal{P}^j, j \in [J]$ that satisfy for all $\mathbf{r} \in \mathcal{P}^+$:*

$$S(\omega, \mathbf{r}) = \sum_{j=2}^{J} S^{\pi(j)}(\omega^{\pi(j)}, \mathbf{r}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}}) + S^{\pi(1)}(\omega^{\pi(1)}, \mathbf{r}^{\pi(1)}) \tag{3.11}$$

for some ω and permutation π. Let r_y, r_z be components of the vector $\mathbf{r}$ that satisfy $y^{\pi(1)} = z^{\pi(1)}, y^{\pi(1)} \neq \omega^{\pi(1)}, z^{\pi(1)} \neq \omega^{\pi(1)}$ where $y^{\pi(1)} \in \Omega^{\pi(1)}, z^{\pi(1)} \in \Omega^{\pi(1)}$. Then for every $\mathbf{r} \in \mathcal{P}^+$:

$$\frac{\partial}{\partial r_y} S(\omega, \mathbf{r}) = \frac{\partial}{\partial r_z} S(\omega, \mathbf{r}).$$

Proof. Consider the vector $\mathbf{p} = \mathbf{r} + t(\boldsymbol{\delta}_y - \boldsymbol{\delta}_z)$ with $t > 0$. For $t < \min\{r_y, r_z\}$, we have $\mathbf{p} \in \mathcal{P}^+$. Furthermore, for any such t we must have that

$$S(\omega, \mathbf{p}) = S(\omega, \mathbf{r})$$

as our choice of y, z ensures that the right side of Equation (3.11) remains constant for shifts of mass between outcomes y and z.

Hence:

$$0 = \lim_{t \to 0} \frac{S(\omega, \mathbf{r} + t(\boldsymbol{\delta}_y - \boldsymbol{\delta}_z)) - S(\omega, \mathbf{r})}{t} = (\boldsymbol{\delta}_y - \boldsymbol{\delta}_z)^\top \nabla_{\mathbf{r}} S(\omega, \mathbf{r})|_{\mathbf{r}}$$

and the result follows. □

We are now ready to prove Theorem 3.2.

Proof. (Theorem 3.2)

Let $\omega = (\omega^1, \omega^2, \ldots, \omega^j) \in \Omega$ be chosen arbitrarily. Let $y \in \Omega$ satisfy $y^j \neq \omega^j, \forall j \in [J]$, i.e., the joint outcomes ω and y differ in every dimension. Suppose first that $z \in \Omega, z \neq \omega$ satisfies $z^k = y^k$ for some $k \in [J]$. By strong additivity, we have for any $\mathbf{r} \in \mathcal{P}^+$,

$$S(\omega, \mathbf{r}) = \sum_{j \in [J] \setminus k} S^{\pi(j)}(\omega^{\pi(j)}, \mathbf{r}^{\pi(j)|\omega^{\pi(j-1)}, \ldots, \omega^k}) + S^k(\omega^k, \mathbf{r}^k), \tag{3.12}$$

where π permutes the remaining $J-1$ dimension of Ω after k is removed. As Equation (3.12) holds for all $\mathbf{r}$ in $\mathcal{P}^+$, we apply Lemma 3.1 and conclude that

$$\frac{\partial}{\partial r_y} S(\omega, \mathbf{r}) = \frac{\partial}{\partial r_z} S(\omega, \mathbf{r}).$$

Now we deal with the case of $z \in \Omega, z \neq \omega$ for which $z^j \neq y^j$ for all $j \in [J]$. By assumption, there exists l such that $\omega^l \neq z^l$. Again, by strong additivity, we have

$$S(\omega, \mathbf{r}) = \sum_{j \in [J] \setminus l} S^{\pi(j)}(\omega^{\pi(j)}, \mathbf{r}^{\pi(j)|\omega^{\pi(j-1)}, \ldots, \omega^l}) + S^l(\omega^l, \mathbf{r}^l).$$

Pick $x \in \Omega$ that satisfies $x^l = z^l$ and $x^k = y^k$ for some $k \in [J]$. Applying Lemma 3.1 yields

$$\frac{\partial}{\partial r_x} S(\omega, \mathbf{r}) = \frac{\partial}{\partial r_z} S(\omega, \mathbf{r}).$$

We have previously observed that

$$\frac{\partial}{\partial r_x} S(\omega, \mathbf{r}) = \frac{\partial}{\partial r_y} S(\omega, \mathbf{r})$$

and thus have shown that

$$\frac{\partial}{\partial r_z} S(\omega, \mathbf{r}) = \frac{\partial}{\partial r_y} S(\omega, \mathbf{r}) := C$$

for all $z \neq \omega$.

Suppose that distributions $\mathbf{p}, \mathbf{q}$ satisfy $p_\omega = q_\omega$. By the Mean Value Theorem,

$$S(\omega, \mathbf{p}) - S(\omega, \mathbf{q}) = \nabla_{\mathbf{r}} S(\omega, \mathbf{r})|_{\mathbf{c}}^{\top} (\mathbf{p} - \mathbf{q}) \tag{3.13}$$

for some $\mathbf{c} \in \mathcal{P}^+$. (Here we rely on the convexity of $\mathcal{P}^+$.) Note that

$$0 = \mathbf{1}^\top (\mathbf{p} - \mathbf{q}) = (p_\omega - q_\omega) + \sum_{y \neq \omega} (p_y - q_y),$$

which in turn means that we must have $\sum_{y \neq \omega}(p_y - q_y) = 0$. We can then conclude that

$$\nabla_{\mathbf{p}} S(\omega, \mathbf{p})|_{\mathbf{c}}^{\top}(\mathbf{p} - \mathbf{q}) = C \sum_{y \neq \omega}(p_y - q_y) = 0. \tag{3.14}$$

From Equations (3.13) and (3.14), we conclude that $S(\omega, \mathbf{p}) = S(\omega, \mathbf{q})$ whenever $p_\omega = q_\omega$. Thus, we can write $S(\omega, \mathbf{p}) = f(p_\omega)$ on $\mathcal{P}^+$. As x was arbitrary, it follows from Theorem 3.3 that

$$S(\omega, \mathbf{p}) = \ln(p_\omega) \tag{3.15}$$

up to additive and multiplicative constants on $\mathcal{P}^+$. Finally, we extend Equation (3.15) to $\mathcal{P}$ using the continuity of S. □

We also have the following corollary.

Corollary 3.1. *Let S and $S^j, j \in [J]$ satisfy the postulates of Theorem 3.2. Then $S^j(\omega^j, \mathbf{r}^j) = \ln(r^j_{\omega^j})$ up to additive and multiplicative constants.*

Proof. We have

$$S(\omega, \mathbf{r}) = \ln\left(\prod_{j=1}^{J} r_{\omega^j}^{j|\omega^{j-1},\ldots,\omega^1}\right) = \sum_{j=2}^{J} S^j(\omega^j, \mathbf{r}^{j|\omega^{j-1},\ldots,\omega^1}) + S^1(\omega^1, \mathbf{r}^1).$$

As the left-hand side depends only on the probability $r_{\omega^j}^{j|\omega^{j-1},\ldots,\omega^1}$ for each j, so too does the right-hand side. Since the conditional distributions $\mathbf{r}^{j|\omega^{j-1},\ldots,\omega^j}$ are arbitrary, it must be that $S^j(\omega^j, \mathbf{r}^j)$ equals $f(r^j_{\omega^j})$ for all $\mathbf{r}^j \in \mathcal{P}^{j+}$, for some function f. Differentiation on both sides of the above equality then yields $f'(p) = 1/p$, and the result follows. □

Theorem 3.2 and Corollary 3.1 demonstrate that the only scoring procedure for which the joint score is equal to a sum of conditional scores, regardless of the order of conditioning, is the method using the Logarithmic scoring rule presented in Example 3.1. Thus, when the Logarithmic scoring rule is used to score dependent assessments, both forecasters and end users can be confident that expert rankings are not dependent upon arbitrary choices regarding the order of conditioning.

The argument given to prove Theorem 3.2 involves showing that a strongly additive scoring rule must be *strictly* local in the sense that $S(\omega, \mathbf{r}) = f(r_\omega)$ for some function f. This can be shown as follows. Suppose that the joint score $S(\omega, \mathbf{r})$ depends on the assessment for p_y for $y \neq \omega$, i.e., on the probability placed on an unobserved outcome. Then we can find, under some order of conditioning, a sequential score that remains unaltered when p_y is changed.

It has been argued by others that strict locality might be a desirable property for a scoring rule to possess [14, 12], because strict locality guarantees that competing forecasts will be ranked exclusively according to the probability placed on the observed outcome. As [76] note, this property implies that the Logarithmic rule is the only strictly proper rule consistent with the likelihood principle.

The Logarithmic scoring rule, in fact, satisfies an even more restrictive property than strong additivity: logarithmic scores are additive over any decomposition of a distribution into component parts. To avoid introducing additional notation, we illustrate this characteristic with a simple example and note that the same principles hold in general.

Suppose we wish to obtain a forecast for the outcome of a given plate appearance in a Major League baseball game. A plate appearance is defined

here as resulting in one of these outcomes:

Single (S), Double (D), Triple (T), Home Run (HR), Walk (W), Out (O).

A distributional forecast $\mathbf{r}$ for the plate appearance would include probabilities for each outcome. However, we could also obtain information on the plate appearance in two steps by decomposition of the distribution $\mathbf{r}$. For instance, we could first assess probabilities for the events

Hit, Walk, Out,

which, in terms of the distribution $\mathbf{r}$ can be represented as

$$\mathbf{r}^1 = [r_S + r_D + r_T + r_{HR}, r_W, r_O].$$

We could then assess the individual conditional probabilities for the events

Single, Double, Triple, Home Run

given a hit:

$$\mathbf{r}^2 = [r_S, r_D, r_T, r_{HR}]/(r_S + r_D + r_T + r_{HR}).$$

Using Logarithmic scoring, after observing the outcome of the plate appearance, we obtain the same score whether we sum the scores for $\mathbf{r}^1$ and, if a hit occurs, $\mathbf{r}^2$, or score $\mathbf{r}$ directly. For example, if the batter singles we have

$$\ln(r_S) = \ln(r_S + r_D + r_T + r_{HR}) + \ln\left(\frac{r_S}{r_S + r_D + r_T + r_{HR}}\right).$$

A generalized version of the decomposition applied to $\mathbf{r}$ above allows for the factoring of a joint distribution into its associated marginal and conditional distributions. Thus, additivity of scores over arbitrary decompositions implies

that the scoring rule is strongly additive. It then follows that the logarithmic scoring rule is also the only rule that is additive over arbitrary decompositions of the forecast distribution.

The Shannon entropy has similar additivity properties. The Shannon entropy of a joint distribution is equal to the expected sum of entropies for any decomposition into conditional distributions. Furthermore, when uncertainties are independent, the Shannon entropy of the joint distribution is equal to the sum of entropies for the marginal distributions.

We now demonstrate that, in a more general sense, the additive properties of scoring rules and entropy measures are closely related.

3.5 Additive Scoring Rules and Entropy Measures

Connections between proper scoring rules and information theory are well established. In particular, the negative entropy $-H(\mathbf{p}) = -\mathbf{p}^\top S(\mathbf{p})$ associated with a strictly proper scoring rule can be viewed as a measurement of the amount of uncertainty in the distribution $\mathbf{p}$ [44]. Intuitively, a lower expected score implies higher uncertainty. Here, we extend literature by fully characterizing the relationship between the ex post additivity properties of scoring rules and the additive properties of convex entropy measures satisfying Definition 2.1.

We further assume that an entropy measure H has at least one subgradient at every point in $\mathcal{A}$, so that H can be used to generate a strictly proper scoring rule. The convexity of H already guarantees the existence of a subgradient at every $\mathbf{p} \in \mathcal{A}^+$. We define the following notions of additivity and strong additivity for entropy measures.

Definition 3.3. *Additive Entropy*

Let H and $H^j, j \in [J]$ be entropy measures (Definition 2.1) defined on $\mathcal{A}$ and $\mathcal{A}^j, j \in [J]$, respectively. We say that H is additive if H satisfies

$$H(\mathbf{q}) = \sum_{j \in [J]} H^j(\mathbf{q}^j) \tag{3.16}$$

for all $\mathbf{q} \in \mathcal{Q}$.

Similarly, an entropy function is *strongly additive* if it satisfies the following criteria.

Definition 3.4. *Strongly Additive Entropy*

Let H and $H^{\pi(j)}, j \in [J]$ be entropy measures (Definition 2.1) defined on $\mathcal{A}$ and $\mathcal{A}^{\pi(j)}, j \in [J]$, respectively. We say that H is strongly additive if

$$H(\mathbf{p}) = \sum_{j=2}^{J} E_{\mathbf{p}} \left[H^{\pi(j)}(\mathbf{p}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}}) \right] + H^{\pi(1)}(\mathbf{p}^{\pi(1)}) \tag{3.17}$$

for all $\mathbf{p} \in \mathcal{P}^+$, and for every arbitrary permutation $\pi(\cdot)$ of the indices in $[J]$.

Here, the expectation is taken over the realization for the conditional distributions. Both the additivity and strong additivity conditions have been identified as potentially desirable properties of an information measure [3, 26]. (In this context, the functions H and H^j, $j \in [J]$ usually take the same functional forms.) Entropy measures satisfying Definition 3.4 measure the uncertainty contained in the joint distribution $\mathbf{p}$ as equal to the total expected information contained in the associated marginal and conditional distributions. Definition 3.3 relaxes this condition to apply only in the case of independence. Taking one or both properties as axioms that an information function must

satisfy (in tandem with various additional axioms) leads to characterization theorems for some entropy measures. See [28] for an overview of these results.

An example of a strongly additive entropy measure is the (negative) Shannon entropy. An entropy measure that is additive, but not strongly additive, is the (negative) Renyi entropy [66]:

$$R_\alpha(\mathbf{p}) = -\frac{1}{1-\alpha}\left[(\mathbf{1}^\top \mathbf{p}) \ln\left(\sum_{x\in\Omega}\left(\frac{p_\omega}{\mathbf{1}^\top \mathbf{p}}\right)^\alpha\right)\right]. \tag{3.18}$$

In the context of strictly proper scoring rules, the additive properties of the associated generalized entropy functions have received some attention. In particular, [54] noted that the entropy measures of a class of weighted scoring rules satisfy pseudo-additive properties, although the connection between these properties and the properties of the score function itself is not examined. As we now show, the additive properties of scoring rules and entropy measures are, in fact, closely related.

The following lemma will be needed in the proofs in this section; see [67] and [24, Proposition 2.3] for details.

Lemma 3.2.

1. *Suppose* $f : \mathcal{A}^+ \to \mathbb{R}$ *and* $g : \mathcal{A}^+ \to \mathbb{R}$ *are convex functions. Then*

$$\partial(f+g)(\mathbf{p}) = \partial f(\mathbf{p}) + \partial g(\mathbf{p}).$$

2. *Suppose that* $f : \mathcal{A}^+ \to \mathbb{R}$ *is a convex function and* $\mathbf{A}$ *an* $n \times m$ *matrix, and let* $g(\mathbf{p}) = f(\mathbf{Ap})$*. Then*

$$\partial g(\mathbf{p}) = \mathbf{A}^\top \partial f(\mathbf{Ap}).$$

3. *Let $f : \mathcal{A}^+ \to \mathbb{R}$ be a convex function and let $g(\mathbf{p}, t) = tf(\frac{\mathbf{p}}{t})$ be the perspective function of f. Then*

$$\partial g(\mathbf{p}, t) = \left\{ \left[\mathbf{v}, f\left(\frac{\mathbf{p}}{t}\right) - \frac{\mathbf{p} \cdot \mathbf{v}}{t} \right] : \mathbf{v} \in \partial f\left(\frac{\mathbf{p}}{t}\right) \right\}.$$

The following result relates additive entropy functions to the additive scoring rules described in Definition 3.1.

Theorem 3.4. *Let $H : \mathcal{A} \to \mathbb{R}$ be an entropy function. Then there is an additive strictly proper scoring rule with $\mathbf{p}^\top S(\mathbf{p}) = H(\mathbf{p})$ if, and only if,*

1. *H is additive, and*
2. *$H(\mathbf{p}) \geq H(\prod_{j \in [J]} \mathbf{p}^j)$ for all $\mathbf{p} \in \mathcal{P}$, with the inequality strict when $\mathbf{p} \notin \mathcal{Q}$.*

Proof. First, suppose that S is additive and strictly proper on $\mathcal{P}$. By taking expectations, $H(\mathbf{p}) := E_{\mathbf{p}}[S(\omega, \mathbf{p})]$ will be additive as well. Furthermore, for $\mathbf{q} = \prod_{j \in [J]} \mathbf{p}^j$, we have

$$H(\mathbf{p}) \geq E_{\mathbf{p}}[S(\omega, \mathbf{q})] = \sum_{j=1}^{J} E_{\mathbf{p}}[S(\omega^j, \mathbf{p}^j)] = \sum_{j=1}^{J} E_{\mathbf{p}^j}[S(\omega^j, \mathbf{p}^j)] = H(\mathbf{q})$$

where the inequality follows from propriety and the first equality from the additivity of S. By strict propriety of S, the inequality will be strict for $\mathbf{q} \neq \mathbf{p}$.

For the other direction, define the function $\hat{H} : \mathcal{A} \to \mathbb{R}$ as follows:

$$\hat{H}(\mathbf{p}) := H\left(\prod_{j \in [J]} \mathbf{p}^j\right) = \sum_{j \in [J]} H^j(\mathbf{M}^j \mathbf{p}) = \sum_{j \in [J]} H^j(\mathbf{p}^j),$$

where $\mathbf{M}^j$ is a suitably chosen matrix that marginalizes $\mathbf{p}$ with respect to the jth uncertainty (i.e., with $\mathbf{M}^j\mathbf{p} = \mathbf{p}^j$). $\hat{H}$ as defined is convex and 1-homogeneous on $\mathcal{A}$. Let $\mathbf{q} \in \mathcal{Q}^+$ (the same construction given below also applies at the boundary). We will show that if $\mathbf{v}(\mathbf{q}) \in \partial\hat{H}(\mathbf{q})$, then $\mathbf{v}(\mathbf{q})$ is a strict subgradient of H at $\mathbf{q}$. By the additivity of H, $\mathbf{v}(\mathbf{q})^\top\mathbf{q} = \hat{H}(\mathbf{q}) = H(\mathbf{q})$ and thus we need establish only that $\mathbf{v}(\mathbf{q})^\top\mathbf{p} < H(\mathbf{p})$ for all $\mathbf{p} \in \mathcal{P},\ \mathbf{p} \neq \mathbf{q}$.

From parts 1 and 2 of Lemma 3.2 and the definition of $\hat{H}$, it follows that $\mathbf{v}(\mathbf{q}) \in \partial\hat{H}(\mathbf{q})$ if, and only if,

$$\mathbf{v}(\mathbf{q}) = \sum_{j\in[J]} \left(\mathbf{M}^j\right)^\top \mathbf{v}^j(\mathbf{q}^j) \tag{3.19}$$

for some choice of $\mathbf{v}^j(\mathbf{q}^j) \in \partial H^j(\mathbf{q}^j)$. Therefore, for $\mathbf{p} \in \mathcal{P},\ \mathbf{p} \neq \mathbf{q}$, we have

$$\mathbf{v}(\mathbf{q})^\top\mathbf{p} = \sum_{j\in[J]} \mathbf{v}^j(\mathbf{q}^j)^\top\mathbf{p}^j.$$

There are two cases to consider. On the one hand, if $\mathbf{p}^j \neq \mathbf{q}^j$ for some j, then, using the strict convexity of H^j for $j \in [J]$, we find

$$\mathbf{v}(\mathbf{q})^\top\mathbf{p} < \sum_{j\in[J]} H^j(\mathbf{p}^j) = \hat{H}(\mathbf{p}) \leq H(\mathbf{p}).$$

On the other hand, if $\mathbf{p}^j = \mathbf{q}^j$ for all j, then we find

$$\mathbf{v}(\mathbf{q})^\top\mathbf{p} = \hat{H}(\mathbf{p}) < H(\mathbf{p}),$$

where the strict inequality follows by assumption. Thus, we have shown that $\mathbf{v}(\mathbf{q})$ is indeed a strict subgradient of H at $\mathbf{q}$.

To complete the proof, it follows from Equation (3.19) that for $x \in \mathcal{X}$:

$$\boldsymbol{\delta}_\omega^\top\mathbf{v}(\mathbf{q}) = \boldsymbol{\delta}_\omega^\top \sum_{j\in[J]} \left(\mathbf{M}^j\right)^\top \mathbf{v}^j(\mathbf{q}^j) = \sum_{j\in[J]} \boldsymbol{\delta}_{\omega^j}^\top\mathbf{v}^j(\mathbf{q}^j). \tag{3.20}$$

By Theorem 2.1, the right-hand side of (3.20) is a sum of strictly proper scores applied to the marginal distributions $\mathbf{q}^j, j \in [J]$, whereas the left-hand side is the score under a strictly proper rule for the joint forecast $\mathbf{q}$. □

Theorem 3.4 says the following: an additive scoring rule will have an additive entropy function; but not every additive entropy function will generate an additive scoring rule. Theorem 3.4 is a discrete state-space analogue to [64, Theorem 1]; however, our result requires the second postulate, which has the following interpretation. For a given $\mathbf{p} \in \mathcal{P}$, the distribution $\prod_{j\in[J]} \mathbf{p}^j \in \mathcal{Q}$ has the same marginal distributions as $\mathbf{p}$, but implies independence between the J uncertainties. A scoring rule with entropy H satisfying the second postulate has the property that the expected score is higher when the joint forecast implies dependency. In the context of information theory, taking $-H$ to be a generalized measure of entropy, the postulate reads $-H(\mathbf{p}) \leq -H\left(\prod_{j\in[J]} \mathbf{p}^j\right)$. In this case, the amount of uncertainty in the distribution $\mathbf{p}$, measured as $-H(\mathbf{p})$, is always lower when relationships between the uncertainties are known. The Shannon entropy is an uncertainty measure that satisfies this intuitive property [3]. In contrast, the Renyi entropy does *not* satisfy $R_\alpha(\mathbf{p}) \geq R_\alpha(\prod_{j\in[J]} \mathbf{p}^j)$ for all $\mathbf{p} \in \mathcal{P}$, and, as illustrated in the following example, does not have an associated additive proper scoring rule.

Example 3.4. *Renyi Scoring Rule*

The Renyi entropy function $R_\alpha(\mathbf{p})$, given in Equation 3.18, is strictly convex on $\mathcal{P}$ for $0 < \alpha < 1$ [45]. Furthermore, the additivity requirement given in Definition 3.3 is satisfied with $H^j(\mathbf{p}^j) = R_\alpha(\mathbf{p}^j)$ for all j. Therefore, for α between 0 and 1, it follows that $R_\alpha(\mathbf{p})$ as defined in Equation (3.18) *is an additive entropy measure.*

Using Theorem 2.1, we derive the proper scoring rule on $\mathcal{P}$ whose entropy is R_α. Specifically, we form the rule $S(\omega, \mathbf{r}) = \frac{\partial}{\partial r_\omega} R_\alpha(\mathbf{r})$. Directly computing the partial derivative,

$$\frac{\partial}{\partial r_\omega} R_\alpha(\mathbf{r}) =$$

$$-\frac{1}{1-\alpha}\left[\ln\left(\sum_{\eta\in\Omega}(r_\eta/\mathbf{1}^\top\mathbf{r})^\alpha\right)\right]$$

$$-\frac{1}{1-\alpha}\left[\frac{1}{\sum_{\eta\in\Omega}(r_\eta/\mathbf{1}^\top\mathbf{r})^\alpha}\left(\sum_{\eta\in\Omega}\alpha(r_\eta/\mathbf{1}^\top\mathbf{r})^{\alpha-1}\frac{-r_\eta}{(\mathbf{1}^\top\mathbf{r})^2}+\frac{\alpha r_\omega^{\alpha-1}}{(\mathbf{1}^\top\mathbf{r})^\alpha}\right)\right].$$

Setting $\mathbf{1}^\top\mathbf{r} = 1$ and simplifying yields

$$S(\omega, \mathbf{r}) = -\frac{1}{1-\alpha}\left[\ln\left(\sum_{\eta\in\Omega} p_\eta^\alpha\right) + \frac{\alpha r_\omega^{\alpha-1}}{\sum_{\eta\in\Omega} r_\eta^\alpha} - \alpha\right]. \tag{3.21}$$

In an analagous manner, we can derive proper scoring rules S^j on $\mathcal{P}^j$ with the same functional form. It is trivial to construct an example where the rule given in Equation (3.21) *does not satisfy Definition 3.1.*

When we assume differentiability of the entropy measure, the proof for the "only if" portion of the result can be framed in terms of an optimization problem. The stipulation that $H(\mathbf{p}) \geq H(\prod_{j\in[J]} \mathbf{p}^j)$ on $\mathcal{P}^+$ is equivalent to

$$H(\mathbf{p}) \geq \sum_{j\in[J]} H^j(\mathbf{M}^j\mathbf{p})$$

by the additivity of H, where we recall that $\mathbf{M}^j$ is the matrix that marginalizes $\mathbf{p}$. That H and each H^j are 1-homogeneous allows extending this inequality for $\mathbf{p} \in \mathcal{A}^+$. Hence,

$$\sup_{\mathbf{p}\in\mathcal{A}}\left\{\left[\sum_{j\in[J]} H^j(\mathbf{M}^j\mathbf{p})\right] - H(\mathbf{p})\right\} \leq 0. \tag{3.22}$$

The term on the left-hand side obtains its upper bound at any $\mathbf{q} \in \mathcal{Q}^+$; thus each $\mathbf{q}$ is a critical point of the bracketed function in Equation (3.22). When the functions H and $H^j,\ j \in [J]$ are differentiable, this implies

$$\nabla_{\mathbf{p}} \left\{ \left[\sum_{j \in [J]} H^j(\mathbf{M}^j \mathbf{p}) \right] - H(\mathbf{p}) \right\} \Big|_{\mathbf{q}} = \mathbf{0}$$

for $\mathbf{q} \in \mathcal{Q}^+$, or equivalently, by the chain rule,

$$\left[\sum_{j \in [J]} \left(\mathbf{M}^j \right)^\top \nabla_{\mathbf{q}^j} H^j(\mathbf{q}^j) \right] - \nabla_{\mathbf{q}} H(\mathbf{q}) = \mathbf{0}.$$

For $\omega \in \Omega$, a direct computation demonstrates that

$$\begin{aligned} &\boldsymbol{\delta}_\omega^\top \left[\sum_{j \in [J]} \left(\mathbf{M}^j \right)^\top \nabla_{\mathbf{q}^j} H^j(\mathbf{q}^j) \right] - \boldsymbol{\delta}_\omega^\top \nabla_{\mathbf{q}} H(\mathbf{q}) \\ &= \sum_{j \in [J]} \boldsymbol{\delta}_{\omega^j}^\top \nabla_{\mathbf{q}^j} H^j(\mathbf{q}^j) - \boldsymbol{\delta}_\omega^\top \nabla_{\mathbf{q}} H(\mathbf{q}) = 0, \end{aligned}$$

or

$$\sum_{j \in [J]} \boldsymbol{\delta}_{\omega^j}^\top \nabla_{\mathbf{q}^j} H^j(\mathbf{q}^j) = \boldsymbol{\delta}_\omega^\top \nabla_{\mathbf{q}} H(\mathbf{q}). \tag{3.23}$$

It then follows from Theorem 2.1 that the first term in Equation (3.23) is a sum of strictly proper scores generated from the H^j, whereas the second term is the joint score derived from diffentiation of H.

We now turn to strongly additive entropy measures and their associated proper scores. A preliminary weak result is as follows. If H is *strongly* additive, then any scoring rule associated with H will be additive on $\mathcal{Q}^+$. To prove this, first note that strongly additive H will satisfy the weaker additivity requirement for $\mathbf{q} \in \mathcal{Q}^+$. Furthermore, for each $j \in [J]$, Jensen's inequality gives

$$E[H^j(\mathbf{p}^{j|\omega^{j-1},\ldots,\omega^1})] \geq H^j(E[\mathbf{p}^{j|\omega^{j-1},\ldots,\omega^1}]) = H^j(\mathbf{p}^j)$$

for $\mathbf{p} \in \mathcal{P}^+$. Therefore,

$$\begin{aligned} H(\mathbf{p}) &= \sum_{j=2}^{J} E_{\mathbf{p}}[H^{\pi(j)}(\mathbf{p}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}})] + H^{\pi(1)}(\mathbf{p}^{\pi(1)}) \\ &\geq \sum_{j=1}^{J} H^j(\mathbf{p}^j) \\ &= H(\prod_{j\in J} \mathbf{p}^j), \end{aligned}$$

with the inequality being strict provided that $\mathbf{p} \neq \prod_{j\in J} \mathbf{p}^j$ (i.e., $\mathbf{p} \notin \mathcal{Q}$). Then, using Theorem 3.4, we conclude that there is an additive $\mathcal{P}^+$-proper scoring rule S with entropy function H.

A stronger result holds as well. As the following demonstrates, a strictly proper scoring rule will in fact be *strongly* additive if, and only if, its corresponding entropy measure is also strongly additive.

Theorem 3.5. *Let S be a strongly additive scoring rule. Then for $\mathbf{p}, \mathbf{r} \in \mathcal{P}^+$:*

$$\begin{aligned} &E_{\mathbf{p}}[S(\omega, \mathbf{r})] \\ &= \sum_{j=2}^{J} E_{\mathbf{p}}\left[E_{\mathbf{p}^{j|\omega^{j-1},\ldots,\omega^1}}[S^{\pi(j)}(\omega^{\pi(j)}, \mathbf{r}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}})]\right] + E_{\mathbf{p}^{\pi(1)}}[S(\omega^{\pi(1)}, \mathbf{r}^{\pi(1)})] \end{aligned}$$

for any permutation π. In particular, the entropy measure $H(\mathbf{p}) = \mathbf{p}^\top S(\mathbf{p})$ will be strongly additive.

Proof. For $\mathbf{p}, \mathbf{r} \in \mathcal{P}^+$, it follows from the strong additivity of S that for any choice of permutation π,

$$E_{\mathbf{p}}[S(\omega, \mathbf{r})] = \sum_{j=2}^{J} E[S^{\pi(j)}(\omega^{\pi(j)}, \mathbf{r}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}})] + E[S^{\pi(1)}(\omega^{\pi(x)}, \mathbf{r}^{\pi(1)})]. \tag{3.24}$$

For each j, we have

$$\begin{aligned} & E[S^{\pi(j)}(\omega^{\pi(j)}, \mathbf{r}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}})] \\ &= E\left[E[S^{\pi(j)}(\omega^{\pi(j)}, \mathbf{r}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}})|\omega^{\pi(j-1)}, \ldots, \omega^{\pi(1)}]\right] \\ &= E_{\mathbf{p}}\left[E_{\mathbf{p}^{j|\omega^{j-1},\ldots,\omega^1}}[S^{\pi(j)}(\omega^{\pi(j)}, \mathbf{r}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}})]\right]. \end{aligned} \tag{3.25}$$

Substituting Equation (3.25) back into Equation (3.24) for each j proves the first statement. Choosing $\mathbf{r} = \mathbf{p}$ shows that $H(\mathbf{p})$ is strongly additive. □

Theorem 3.6. *Let H and $H^j, j \in [J]$ be entropy functions, defined on $\mathcal{P}$ and $\mathcal{P}^j, j \in [J]$, respectively, that satisfy the strong additivity requirement. Then any strictly proper scoring rule S with entropy function H will be strongly additive.*

Proof. The proof is straightforward but notationally cumbersome. For compactness, we define the vector-valued map $\mathbf{c}^j : \mathcal{A}^+ \times \Omega \to \mathcal{A}^{j+}$ as

$$\mathbf{c}^j(\mathbf{p}, \omega) := \frac{\mathbf{C}^j_\omega \mathbf{p}}{\chi(\{\eta \in \Omega : \eta^{j-1} = \omega^{j-1}, \ldots, \eta^1 = \omega^1\})^\top \mathbf{p}}, \tag{3.26}$$

where $\mathbf{C}^j_\omega$ is a $|\Omega^j| \times |\Omega|$ matrix, with each row corresponding to a $\omega^j \in \Omega^j$, and where each row satisfies

$$\mathbf{C}^j_\omega(\omega^j) = \chi(\{\eta \in \Omega : \eta^j = \omega^j, \eta^{j-1} = \omega^{j-1}, \ldots, \eta^1 = \omega^1\}).$$

When $\mathbf{p} \in \mathcal{P}$, the denominator in Equation (3.26) is simply

$$P(\omega^{j-1} = \omega^{j-1}, \ldots, \omega^1 = \omega^1)$$

so that Equation (3.26) is just a rewriting of the standard formula for conditional probability in terms of matrix multiplication. In other words, $c^j(\mathbf{p}, \omega) =$

$\mathbf{p}^{j|\omega^{j-1},\ldots,\omega^1}$ for $\mathbf{p} \in \mathcal{P}$. For an entropy measure H, we define $H(c^j(\mathbf{p}))$ to be the n-dimensional vector with components $H(c^j(\mathbf{p},\omega))$ for $\omega \in \Omega$ and note that $E[H(c^j(\mathbf{p},\omega))] = \mathbf{p}^\top H(c^j(\mathbf{p}))$ for $\mathbf{p} \in \mathcal{P}$.

In what follows, all statements will hold for any permutation of the indices $j \in [J]$, so we no longer reference the permutation operator π. As a preliminary, we note that the strong additivity condition for entropy measures given in Definition 3.4 can be extended from $\mathcal{P}^+$ to $\mathcal{A}^+$. When H is strongly additive on $\mathcal{P}^+$, for $\lambda\mathbf{p} \in \mathcal{A}^+$, we have

$$\begin{aligned} H(\lambda\mathbf{p}) = \lambda H(\mathbf{p}) &= \lambda \left[\sum_{j=2}^{J} E_{\mathbf{p}}[H^j(\mathbf{p}^{j|\omega^{j-1},\ldots,\omega^1})] + H^1(\mathbf{p}^1) \right] \\ &= \lambda \sum_{j=2}^{J} \mathbf{p}^\top H^j(\mathbf{c}^j(\mathbf{p})) + H^1(\lambda\mathbf{p}^1) \\ &= \sum_{j=2}^{J} \lambda\mathbf{p}^\top H^j(\mathbf{c}^j(\lambda\mathbf{p})) + H^1(\mathbf{M}^1(\lambda\mathbf{p})). \end{aligned} \tag{3.27}$$

The final equality follows from the extension of the domain of $\mathbf{c}^j$ to $\mathcal{A}^+$ and the observation that $\mathbf{c}^j$, as defined above, is 0-homogeneous.

We start with the equality derived in Equation (3.27):

$$H(\mathbf{p}) = \sum_{j=2}^{J} \mathbf{p}^\top H^j(\mathbf{c}^j(\mathbf{p})) + H^1(\mathbf{M}^1(\mathbf{p})), \tag{3.28}$$

which holds for $\mathbf{p} \in \mathcal{A}^+$. It follows that for $\mathbf{p} \in \mathcal{P}^+$ we have

$$\partial H(\mathbf{p}) = \partial \left\{ \sum_{j=2}^{J} \mathbf{p}^\top H^j(\mathbf{c}^j(\mathbf{p})) + H^1(\mathbf{M}^1(\mathbf{p})) \right\}. \tag{3.29}$$

Using the first and second parts of Lemma 3.2, the right side is equivalent to

$$\sum_{j=2}^{J} \partial\mathbf{p}^\top H^j(\mathbf{c}^j(\mathbf{p})) + (\mathbf{M}^1)^\top \partial H^1(\mathbf{M}^1\mathbf{p}). \tag{3.30}$$

For each $y \in \Omega^{j-1} \times \ldots \times \Omega^1$, let

$$\chi_y = \chi(\{\omega \in \Omega : y^{j-1} = \omega^{j-1}, \ldots, y^1 = \omega^1\}).$$

We observe that for each j,

$$\mathbf{p}^\top H^j(\mathbf{c}^j(\mathbf{p})) = \sum_{y \in \Omega^{j-1} \times \ldots \times \Omega^1} (\chi_y^\top \mathbf{p}) H^j\left(\frac{\mathbf{C}_y^j \mathbf{p}}{\chi_y^\top \mathbf{p}}\right), \tag{3.31}$$

where a row $\mathbf{C}_y^j(\omega^j)$ of the matrix $\mathbf{C}_y^j$ satisfies

$$\mathbf{C}_y^j(\omega^j)^\top \mathbf{p} = P(\omega^j = \omega^j, \omega^{j-1} = y^{j-1}, \ldots, \omega^1 = y^1)$$

for $\omega^j \in \Omega^j$. Each term in the sum can be viewed as the composition of the perspective function of H^j with the affine map

$$\mathbf{p} \mapsto (\mathbf{C}_y^j \mathbf{p}, \chi_y^\top \mathbf{p}).$$

It then follows by applying all three parts of Lemma 3.2, that

$$\partial \mathbf{p}^\top H^j(\mathbf{c}^j(\mathbf{p}, X)) =$$
$$\sum_{y \in \Omega^{j-1} \times \ldots \times \Omega^1} \mathbf{B}^\top \left[\partial H^j\left(\frac{\mathbf{C}_y^j \mathbf{p}}{\chi_y^\top \mathbf{p}}\right), H^j\left(\frac{\mathbf{C}_y^j \mathbf{p}}{\chi_y^\top \mathbf{p}}\right) - \left(\frac{\mathbf{C}_y^j \mathbf{p}}{\chi_y^\top \mathbf{p}}\right)^\top \partial H^j\left(\frac{\mathbf{C}_y^j \mathbf{p}}{\chi_y^\top \mathbf{p}}\right)\right], \tag{3.32}$$

where

$$\mathbf{B} = \begin{bmatrix} \mathbf{C}_y^j \\ \chi_y \end{bmatrix}.$$

Because H^j is 1-homogeneous, the set

$$H^j\left(\frac{\mathbf{C}_y^j \mathbf{p}}{\chi_y^\top \mathbf{p}}\right) - \left(\frac{\mathbf{C}_y^j \mathbf{p}}{\chi_y^\top \mathbf{p}}\right)^\top \partial H^j\left(\frac{\mathbf{C}_y^j \mathbf{p}}{\chi_y^\top \mathbf{p}}\right) = \{0\}$$

for every y and j. Therefore, plugging Equation (3.32) back into Equation (3.30) gives:

$$\partial H(\mathbf{p}) = \sum_{j=2}^{J} \sum_{y \in \Omega^{j-1} \times \ldots \times \Omega^1} (\mathbf{C}_y^j)^\top \partial H^j \left(\mathbf{p}^{j|y}\right) + \left(\mathbf{M}^1\right)^\top \partial H^1(\mathbf{M}^1 \mathbf{p}).$$

In particular, for any $\mathbf{v}(\mathbf{p}) \in \partial H(\mathbf{p})$, there are

$$\mathbf{v}^j(\mathbf{p}^{j|\omega^{j-1},\ldots,\omega^1}) \in \partial H^j \left(\mathbf{p}^{j|\omega^{j-1},\ldots,\omega^1}\right)$$

for $j \in [J]$ such that, for $\omega \in \Omega$,

$$\boldsymbol{\delta}_\omega^\top \mathbf{v}(\mathbf{p}) = \boldsymbol{\delta}_\omega^\top \left[\sum_{j=2}^{J} \sum_{y \in \Omega^{j-1} \times \ldots \times \Omega^1} (\mathbf{C}_y^j)^\top \mathbf{v}^j(\mathbf{p}^{j|y}) + \left(\mathbf{M}^1\right)^\top \mathbf{v}^1(\mathbf{p}^1) \right],$$

which reduces by direct calculation to

$$\boldsymbol{\delta}_\omega^\top \mathbf{v}(\mathbf{p}) = \sum_{j \in [J]} \boldsymbol{\delta}_{\omega^j}^\top \mathbf{v}^j(\mathbf{p}^{j|\omega^{j-1},\ldots,\omega^1}). \tag{3.33}$$

By Theorem 2.1, the left-hand side of Equation (3.33) corresponds to the value $S(\omega, \mathbf{p})$ for a scoring rule S with entropy H. Similarly, the right-hand side is the sum of strictly proper scores $S^j(\omega^j, \mathbf{p}^{j|\omega^{j-1},\ldots,\omega^1})$ with S^j derived from H^j. For the same choice of $\mathbf{v}$, an identical argument shows that Equation (3.33) holds for any permutation of $j \in [J]$. This has demonstrated that any scoring rule whose entropy function is H satisfies the strong additivity condition given in Definition 3.4. □

Theorem 3.5 shows that when a strongly additive scoring rule is used, the expected score is also invariant to the order of conditioning. Theorem 3.6 demonstrates that when the entropy measure used to construct a scoring rule

is strongly additive – that is, the optimal expected score is invariant over all decompositions into marginal and conditional distributions – then the scoring rule will be strongly additive as well. Leveraging these results, together with the previously established fact that the Logarithmic rule is the only strongly additive strictly proper scoring rule, we now prove the following characterization of the Shannon entropy. The corollary presented below establishes that the Shannon entropy is the only measure of uncertainty (among the class of such functions that are twice differentiable) that is concave and strongly additive.

Corollary 3.2. *Suppose that the entropy functions H and $H^j, j \in [J]$ satisfy the strong additivity requirement given in Definition 3.4. Furthermore, suppose that H is twice continuously differentiable on $\mathcal{A}$. Then, on $\mathcal{P}$,*

$$H(\mathbf{p}) = \sum_{x \in \Omega} p_\omega \ln(p_\omega)$$

up to addition and positive scaling.

Proof. By Theorem 3.6, the functions H and $H^j, j \in [J]$ produce (using Theorem 2.1) proper scoring rules S and $S^j,\ j \in [J]$, respectively, that satisfy

$$S(\omega, \mathbf{p}) = \sum_{j=2}^{J} S^{\pi(j)}(\omega^{\pi(j)}, \mathbf{p}^{\pi(j)|\omega^{\pi(j-1)},\ldots,\omega^{\pi(1)}}) + S^{\pi(1)}(\omega^{\pi(x)}, \mathbf{p}^{\pi(1)})$$

for any permutation π of $[J]$. By Theorem 3.2, on $\mathcal{P}$, $S(\omega, \mathbf{p}) = \ln(p_\omega)$ up to a positive affine transformation. It follows that

$$H(\mathbf{p}) = E_{\mathbf{p}}\left[S(\omega, \mathbf{p})\right] = \sum_{x \in \Omega} p_\omega \ln(p_\omega)$$

up to multiplicative and additive constants. □

The connection between the additive properties of entropy measures and scoring rules is of theoretical interest. However, the results presented in this section also have practical implications. First, the relationship between additive entropy functions and additive scoring rules allows one to construct an additive rule from an additive entropy measure and vice versa. Second, using an additive (or strongly additive) scoring rule leads to additive (or strongly additive) ex ante incentives. For additive rules, this implies that the information-gathering incentives for the forecaster are the same whether the information is scored sequentially or scored in terms of the joint distribution. For strongly additive rules, information-gathering incentives remain unchanged over all possible conditioning permutations. Finally, if the underlying uncertainty has true sampling distribution $\mathbf{p}^*$, we have shown that the long-run average score under multiple observations, $E_{\mathbf{p}^*}[S(\omega, \mathbf{r})] = \lim_n \frac{1}{n} \sum_n S(\omega_n, \mathbf{r})$, inherits the additive properties of S. In contrast, in the independent coin example presented in Section 3.3, we observed that a non-additive rule (the Quadratic score applied to a joint distribution) produced illogical results irrespective of sample size.

3.6 Conclusion

Typically, when rewarding forecasts for multiple uncertainties, scores for each individual forecast are added or averaged. Alternatively, a joint forecast, when available, could be scored directly. When an strictly proper scoring rule is chosen as the reward function, both procedures establish ex ante incentives for honest reporting. However, depending on the scoring rule chosen, the ex post properties of one or both scoring methods may be undesirable.

Some rules, such as the Quadratic and Spherical scores, have illogical ex post properties when applied to joint forecasts. In particular, when forecasts

are made for independent uncertainties, a better marginal forecast may correspond with a worse score. In contrast, the additive rules we derived in Section 3.4 are ideal for scoring joint forecasts because, in cases of independence, they measure accuracy in terms of the marginal predictions.

However, sequentially adding scores can also produce arbitrary ex post results when the forecasts are made for dependent uncertainties. For example, under Quadratic and Spherical scoring, the average score is dependent on the order in which the variables are conditioned. As we demonstrated in Section 3.4.2, the Logarithmic rule is the only strictly proper score guaranteed to give the same ex post score independent of the order in which the forecasts are conditioned. This result is directly related to the strict locality of the Logarithmic scoring rule and the desirable additive properties of the Shannon entropy.

The additive properties of scoring rules and their associated entropy measures are closely related. Choosing an additive or strongly additive rule in sequential-scoring contexts guarantees that the associated entropy measure provides the same ex ante incentives independent of rearrangement of the underlying probabilistic information. Conversely, entropy measures with additive properties generate additive scoring rules. These results further demonstrate the fundamental link between choosing an entropy measure and choosing a scoring rule.

There are many ex post and/or ex ante properties that practitioners may consider when selecting a scoring rule (for example, simplicity of the functional form or risk preferences of the forecaster). Our work demonstrates the benefit of Logarithmic scoring when forecasts are provided for multiple uncertainties. The Logarithmic score treats equivalent probabilistic information

equivalently. Furthermore, practitioners need not worry about whether a forecast corresponds with a joint or marginal distribution when scoring. In this way, logical ex post scores are obtained irrespective of full understanding of the underlying probabilistic structure.

Chapter 4

Weighted Scoring Rules and Convex Risk Measures

4.1 Chapter Summary

This chapter examines two classes of proper weighted scoring rules with associated entropy measures based on ϕ-divergences. These rules are generalizations of the *Weighted Power* and *Weighted Pseudospherical* rules presented in [54]. We demonstrate that these scoring rules are tailored to optimization problems under uncertainty, where the aim is to maximize a concave certainty equivalent. Further, we demonstrate that every weighted scoring rule has the following economic interpretation: the entropy measure associated with a weighted scoring rule is equal to the maximum value of an optimization problem involving a convex risk measure. This result connects the theory of proper scoring rules with financial mathematics and risk evaluation.

4.2 Introduction

We develop a general connection between two types of proper scoring rules: *weighted* scoring rules and *tailored* scoring rules. Weighted proper scoring rules, introduced in [54], explicitly incorporate a baseline forecast, in the form of a distribution $\mathbf{q}$, into the score function. Tailored scoring rules are constructed directly from a decision problem, and align the information-gathering

incentives of the forecaster with those of the decision maker [52].

Jose et al. [54] demonstrated that two families of weighted rules, the Weighted Power and Weighted Pseudospherical scores, are each tailored to an optimization problem faced by a risk averse investor. We provide a novel and more general economic interpretation for weighted scoring rules: *every weighted scoring rule is tailored to an optimization problem under uncertainty involving the maximization of a concave certainty equivalent, or equivalently, the minimization of a convex risk measure.* Convex risk measures, introduced in [35], are a class of functionals that are used in the context of financial mathematics to compute the cash on hand required to back a risky position. Our result provides a new connection between proper scoring and financial mathematics, and makes explicit the relationship between the choice of scoring rule and an investor's quantification of risk.

After the presentation of this result, we introduce two classes of weighted proper scoring rules, ϕ-Divergence scoring rules and Scaled ϕ-Divergence scoring rules, which are tailored to optimization problems involving the *Optimized Certainty Equivalent* [11] and the *u-Mean Certainty Equivalent* [35], respectively (both members of the class of concave certainty equivalents). As we demonstrate, these families generalize the Weighted Power and Weighted Pseudospherical scores, respectively, and have similar economic interpretations. We derive the key properties of these scoring families and demonstrate how the main theoretical results presented in [54] follow directly from the connection between these scoring families and their associated convex risk measures.

The remainder of this section is organized as follows. Section 4.3 provides relevant background on weighted and tailored proper scoring rules. Section 4.4 details the general connection between weighted scoring rules and opti-

mization problems involving a convex risk measure. Section 4.5 introduces ϕ-Divergence and Scaled ϕ-Divergence weighted scoring rules, discusses their properties, and shows that they are tailored to optimization problems involving the maximization of a concave certainty equivalent (the negation of a convex risk measure).

4.2.1 Notation and Scoring Framework

This chapter focuses on scoring forecasts from $\mathcal{P}$, that is, it does not consider denormalized forecasts. Although there is no mathematical distinction between scoring normalized vs. denormalized forecasts, the literature on weighted scoring rules and tailored scoring rules generally portrays scoring rules and their entropy measures as functions restricted to $\mathcal{P}$. For consistency with our primary reference [54], we use the McCarthy/Savage characterization theorem for proper scoring rules, stated formally below.

Theorem 4.1. *[60, 46, 40]*

Let $S : \Omega \times \mathcal{P} \to [-\infty, \infty)$ be a (strictly) proper scoring rule. Then $H(\mathbf{p}) := E_{\mathbf{p}}[S(\omega, \mathbf{p})]$ is (strictly) convex on $\mathcal{P}$ and the vector $S(\mathbf{r})$ is a subgradient of H at $\mathbf{r}$ for each $\mathbf{r} \in \mathcal{P}$. Conversely, given a (strictly) convex function $H : \mathcal{P} \to \mathbb{R}$, the scoring rule given by

$$S(\omega_i, \mathbf{r}) = H(\mathbf{r}) + \boldsymbol{\delta}_i^\top \mathbf{v}(\mathbf{r}) - \mathbf{r}^\top \mathbf{v}(\mathbf{r})$$

for $\omega_i \in \Omega$ is (strictly) proper for any choice of subgradient $\mathbf{v}(\mathbf{r}) \in \partial H(\mathbf{r})$.

Consistent with Theorem 4.1, this chapter will treat entropy measures as strictly convex functions on $\mathcal{P}$. Note that given a strictly proper scoring rule S, the entropy function $H(\mathbf{p}) = \sup_{\mathbf{r} \in \mathcal{P}} \mathbf{p}^\top S(\mathbf{r})$ is strictly convex and closed (as it is the supremum of affine functions).

In addition, because this section does not (explicitly) consider joint forecast and marginal/conditional forecasts, we simplify the notation as follows: for a probability distribution $\mathbf{p} \in \mathcal{P}$, we define $p_i := p_{\omega_i}$ for an outcome $\omega_i \in \Omega$.

We now formally define weighted and tailored proper scoring rules.

4.3 Weighted and Tailored Proper Scoring Rules

4.3.1 Weighted Scoring Rules

Under a weighted scoring rule, the reward for a forecast $\mathbf{r}$ depends directly on the improvement over a baseline distribution $\mathbf{q}$. Given a realization $\omega_i \in \Omega$, a forecaster receives the score $S(\omega_i, \mathbf{r}, \mathbf{q})$, where $\mathbf{r}$ is the reported distribution and $\mathbf{q} \in \mathcal{P}$ is the baseline. The optimal expected score function for a weighted scoring rule, $H(\mathbf{p}, \mathbf{q})$, is also a function of the baseline and is minimized as a function of $\mathbf{p}$, when $\mathbf{p} = \mathbf{q}$. Without loss of generality, we can assume that $H(\mathbf{q}, \mathbf{q}) = 0$; if this is not the case for a scoring rule S, we can instead consider the equivalent rule $S(\omega_i, \mathbf{r}, \mathbf{q}) - H(\mathbf{q}, \mathbf{q})$. The function $H(\mathbf{p}, \mathbf{q})$ is convex in its first argument and can be viewed as a measure of divergence of $\mathbf{p}$ from $\mathbf{q}$. In light of Theorem 4.1, every proper weighted scoring rule is a subgradient of the convex divergence function $H(\mathbf{p}, \mathbf{q}) = E_{\mathbf{p}}[S(\omega, \mathbf{p}, \mathbf{q})]$. Conversely, given a convex divergence measure, we can find a proper weighted scoring rule by taking subgradients.

By comparison, many common unweighted scoring rules, such as the Logarithmic, Quadratic, and Spherical scores, have entropy measures $H(\mathbf{p})$ that are minimized at the uniform distribution $\mathbf{q} = \left[\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right]$ [14]. Under these unweighted reward functions, a forecaster's expected score is lowest when he believes each possible outcome of the underlying uncertainty is equally likely.

The farther the expert deviates from this baseline, the more he increases his expected score.

While, in some applications, the assumption that the expert should be measured against a uniform baseline may be reasonable, many settings include a publicly available benchmark forecast. For example, in weather forecasting, there are readily available frequency data for various events, such as the proportion of days on which rain is observed in a given area. Or, in sports betting, there are often published odds that imply win probabilities for competing teams. In these contexts and many others, the value of a forecast is commensurate with its improvement over the publicly available baseline and should be rewarded accordingly.

Two generic families of weighted scoring rules, introduced in [54], are the *Weighted Power* and *Weighted Pseudospherical* scoring rules, which are defined in terms of their entropy measures as follows.

Example 4.1. *In terms of the entropy function, the Weighted Power scoring rules are defined by*

$$H_\beta^P(\mathbf{p}, \mathbf{q}) = \frac{1}{\beta(\beta-1)} \left[\sum_{i=1}^{n} q_i \left(\frac{p_i}{q_i} \right)^\beta - 1 \right] \tag{4.1}$$

for $\beta \in \mathbb{R}$. *The Weighted Pseudospherical scoring rules are likewise defined by the entropy function*

$$H_\beta^S(\mathbf{p}, \mathbf{q}) = \frac{1}{\beta-1} \left[\left[\sum_{i=1}^{n} q_i \left(\frac{p_i}{q_i} \right)^\beta \right]^{1/\beta} - 1 \right] \tag{4.2}$$

for $\beta \in \mathbb{R}$.

The properties of these rules and their generalizations have been further explored in [50] and [36]. Weighted scoring rules have also been developed for density functions [48] and cumulative distribution functions [55].

4.3.2 Tailored Scoring Rules

Tailored proper scoring rules were comprehensively defined in [52]. The goal of a tailored scoring rule is to align the incentives of a decision maker and a forecaster. Suppose that a decision maker faces a stochastic optimization problem of the form:

$$f(\mathbf{p}) = \sup_{\mathbf{x}\in\mathcal{X}} E_{\mathbf{p}}[u(\omega, \mathbf{x})], \tag{4.3}$$

where $\mathbf{x}$ is a vector of decision variables constrained to lie in the set $\mathcal{X}$, $\mathbf{p}$ is a discrete probability distribution, and u is a concave utility function. If a forecast is obtained for the distribution $\mathbf{p}$, a proper scoring rule

$$S(\omega_i, \mathbf{r}) = u(\omega_i, \mathbf{x}^*(\mathbf{r})),\ \mathbf{x}^*(\mathbf{r}) \in \arg\max_{\mathbf{x}\in\mathcal{X}} E_{\mathbf{r}}[u(\omega, \mathbf{x})] \tag{4.4}$$

is the tailored scoring rule associated with the decision problem. The entropy measure, that is, the expected score for an honest forecaster who believes $\mathbf{p}$, will be precisely $f(\mathbf{p})$ when the tailored scoring rule defined in Equation (4.4) is chosen. Thus, the information-gathering incentives the same for the forecaster and the decision maker in that an improved expected score for the forecaster corresponds to increased rewards in Equation (4.3). This suggests that when forecasts are used directly to solve a specific decision problem, a tailored rule should be chosen. However, one drawback is that the scoring function for a tailored rule will usually not have a closed form.

The functional form of the tailored scoring rule depends upon the risk preferences of the decision maker, which are captured by the utility function

u appearing in Equation (4.4). We do not directly consider forecaster risk aversion here. In the case of forecaster risk aversion, any scoring rule, including a tailored rule, might not be proper or strictly proper. On the other hand, if the forecaster has a known utility function v, then a risk-adjusted version of any proper score $v^{-1}(S)$ can be used instead, so that the forecaster's expected risk-adjusted score is

$$E_{\mathbf{p}}[v(v^{-1}(S(\omega, \mathbf{r}))] = E_{\mathbf{p}}[S(\omega, \mathbf{r})],$$

equivalent to the expected score under the original rule.

Scoring rules and their entropy measures have often been connected to specific decision problems. Prior to the formalization of the notion of a tailored scoring rule, it was observed that scoring rules arise naturally in the context of stochastic optimization problems [44]. The theoretical requirements for the existence of incentive-matching scoring rules have been laid out in [34]. Working in the other direction, researchers have identified the decision problems for which various scoring rules are tailored, providing an economic rationale for the score. As described in the next section, the authors in [54] identify the decision problems for which the Weighted Power and Weighted Pseudospherical scoring rules defined in Equations (4.1) and (4.2) are tailored. Tailored scores are applicable even when the decision problem is not in the form of problem (4.3). If a forecast $\mathbf{p}$ is used to solve a decision problem that itself is convex in $\mathbf{p}$, a tailored score can be derived by applying Theorem 4.1.

4.4 Tailored Weighted Scoring Rules and Convex Risk Measures

We now show that all weighted scoring rules are tailored to decision problems involving a risk averse investor. We begin by summarizing the results presented in [54] that show that the Weighted Power Scoring Rules and Weighted Pseudospherical scoring rules are tailored to utility maximization problems under uncertainty.

4.4.1 Tailored Scoring Rules and the Weighted Power and Weighted Pseudospherical Scores

As shown in [54, Theorem 1], the Weighted Power and Weighted Pseudospherical scoring rules have entropy functions equal to the maximum obtainable expected reward for a risk averse investor (who we will also refer to as the "bettor" or "decision maker") with a HARA (hyperbolic absolute risk aversion) utility function. Specifically, the following relationships hold:

$$H_\beta^P(\mathbf{p},\mathbf{q}) = \sup_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] - E_{\mathbf{q}}[\mathbf{x}] \tag{4.5}$$

and

$$H_\beta^S(\mathbf{p},\mathbf{q}) = \sup_{\mathbf{x}\in\mathbb{R}^n} \left\{ E_{\mathbf{p}}[u_\beta(\mathbf{x})] \mid E_{\mathbf{q}}[\mathbf{x}] \leq 0 \right\}, \tag{4.6}$$

where the notation $u_\beta(\mathbf{x})$ is shorthand for the vector

$$[u_\beta(x_1), u_\beta(x_2), \ldots, u_\beta(x_n)].$$

In the optimization problems above, u_β takes the form:

$$u_\beta(x) = \left\{ \begin{array}{ll} \frac{1}{\beta-1}(1+\beta x)^{(\beta-1)/\beta} - \frac{1}{\beta-1}, & \text{for } \beta x \geq -1 \\ -\infty, & \text{o.w.} \end{array} \right\} \tag{4.7}$$

for $\beta \in \mathbb{R}$. The duality relationship exploited in [54] to establish Equations (4.5) and (4.6) extended the known connections between utility maximization problems and information-theoretic divergence measures (see the list of references given in [54, p. 1153]). The results also establish that the Weighted Power and Weighted Pseudospherical scoring rules are tailored to the decision problems (4.5) and (4.6), respectively.

The decision problems (4.5) and (4.6) have real-world interpretations described in detail in [54, 61]. Both can be interpreted in terms of an investor buying and selling Arrow-Debreu securities in a complete market with the goal of maximizing risk-adjusted rewards. For each possible future states in Ω, such a security pays \$1 if ω_i occurs, and \$0 otherwise. The market prices for these securities are $\$q_1, \$q_2, \ldots, \$q_n$.

In problem (4.5), the investor seeks to maximize expected utility over two time periods by buying and selling these securities. In period 1, the investor trades securities (represented by the decision vector $\mathbf{x}$) at market prices determined by the underlying probability distribution $\mathbf{q}$, getting an immediate reward of $E_{\mathbf{q}}[\mathbf{x}]$. In period 2, after observing ω_i, the expected utility for the investments made in period 1 is given by $E_{\mathbf{p}}[u_\beta(\mathbf{x})]$. Problem (4.6) has a similar interpretation. In this case, the investor seeks to maximize the expected utility of his payouts subject to the self-financing constraint $E_{\mathbf{q}}[\mathbf{x}] \leq 0$, which enforces that purchases of securities must be offset by corresponding short-selling at the market prices.

Problem (4.6) also represents the optimal buying/selling strategy for a contingent claim in a complete market [61]. A contingent claim is represented as a vector $\mathbf{y} \in \mathbb{R}^n$ where each component y_i represents the future reward if ω_i is observed. For example, the vector $\mathbf{y}$ could represent the possible prices

at some future time of a stock. In a complete market, the price for buying or selling the asset $\mathbf{y}$ is equal to $E_{\mathbf{q}}[\mathbf{y}]$ under the risk neutral probability distribution $\mathbf{q}$. The future payoff vector available to an investor buying or selling the asset can be represented as a vector $\mathbf{x} = s(\mathbf{y} - \mathbf{1}E_{\mathbf{q}}[\mathbf{y}])$, where s in $\mathbb{R}$ is the amount bought or sold. Thus, all feasible payout vectors satisfy $E_{\mathbf{q}}[\mathbf{x}] = 0$, which is equivalent to the constraint $E_{\mathbf{q}}[\mathbf{x}] \leq 0$ in problem (4.6) because u_β is non-decreasing.

Finally, problem (4.6) models a risk averse investor with utility function u and distribution $\mathbf{p}$ betting against a non-strategic, risk neutral opponent with distribution $\mathbf{q}$. The constraint $E_{\mathbf{q}}[\mathbf{x}] \leq 0$, equivalently $E_{\mathbf{q}}[-\mathbf{x}] \geq 0$, stipulates that the opponent's expected rewards are positive. If the opponent is risk averse with a known utility function, the problem can be reformulated as in (4.6) [54].

The remainder of the chapter establishes a more general relationship between weighted scoring rules and optimization problems involving a risk averse investor. We demonstrate that every weighted scoring rule can be viewed as a tailored scoring rule in which the underlying decision problem involves the minimization of a convex risk measure.

4.4.2 Convex Risk Measures, Concave Certainty Equivalents, and Weighted Proper Scoring Rules

This section develops an exact relationship between weighted scoring rules and convex risk measures. In financial regulation, the latter prescribe the cash on hand needed to back a risky financial position (described in terms of a random variable X) and satisfy a number of intuitive properties [35, 37]. The class of convex risk measures generalizes the class of so-called *coherent risk*

measures described originally in [6]. Convex risk measures arise naturally in the context of optimization problems under uncertainty, including applications in robust and distributionally robust optimization [69, 68, 5]. The formal definition is given below.

Definition 4.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{V}$ be a linear space of bounded random variables that map Ω to $\mathbb{R}$. A functional $\rho : \mathcal{V} \to \mathbb{R}$ is called a convex risk measure if ρ satisfies the following properties:*

1. *Monotonicity: If $X \geq Y$ almost surely, then $\rho(X) \leq \rho(Y)$*
2. *Translation invariance: $\rho(X + c) = \rho(X) - c$, $c \in \mathbb{R}$*
3. *Convexity: $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y)$, $\forall \lambda \in [0, 1]$*

The first axiom in Definition 4.1 specifies that if every outcome of lottery X dominates the corresponding outcome in lottery Y, then X carries less risk compared to Y. The second axiom says that if a risk-free amount c is added to every possible outcome of X, then the risk of the new position decreases by exactly c. The third axiom captures the value of diversification. Generally, convex risk measures are normalized to satisfy $\rho(0) = 0$, and a position X is deemed "acceptable" if $\rho(X)$ is less than or equal to 0. In this case, by axiom two, the amount $\rho(X)$ is the minimum cash injection into the position X that makes X acceptable.

For a convex risk measure ρ, the negation $-\rho$ is known as a concave certainty equivalent. The value $-\rho(X)$ can be viewed as an indifferent buying price for a lottery X [39]. (In particular, if the lottery X is acquired at price

$-\rho(X)$, the resulting position $Y := X - (-\rho(X))$ satisfies $\rho(Y) = 0$.) The traditional von-Neumann-Morgenstern (vNM) certainty equivalent given by

$$\rho(X) := u^{-1}\left(E_P[u(X)]\right)$$

is a concave certainty equivalent only when the utility function u is exponential or piece-wise linear [11]. In general, the vNM certainty equivalent, which in vNM utility theory is the indifferent *selling* price for a lottery X, will not satisfy the second axiom. However, interpreting a concave certainty equivalent as the buying price for X, axiom two encodes the logical property that adding a risk-free reward of c to a lottery is worth c to the purchaser. In the spirit of utility maximization, a number of concave certainty equivalents are constructed directly from a prespecified utility function and can similarly be viewed as valuation measures for uncertain positions [35, 11, 39]. These include the Optimized Certainty Equivalent and u-Mean Certainty Equivalent described in Sections 4.5.2 and 4.5.3.

As demonstrated in [35], each convex risk measure satisfying certain regularity requirements admits a so-called robust representation.

Theorem 4.2. *[35]*

Let $|\Omega| = n$. Then any convex risk measure defined for random variables $\mathbf{x} : \Omega \to \mathbb{R}$ can be expressed as:

$$\rho(\mathbf{x}) = \sup_{\mathbf{p} \in \mathcal{P}} \; E_{\mathbf{p}}[-\mathbf{x}] - \alpha(\mathbf{p}), \tag{4.8}$$

where $\alpha(\cdot)$ is a closed, convex function.

The converse is also true: any functional in the form of the right-hand side of Equation (4.8) defines a convex risk measure. For discrete random variable

$\mathbf{x}$, we can choose

$$\rho(\mathbf{x}) := \sup_{\mathbf{p}\in\mathcal{P}} \{E_{\mathbf{p}}[-\mathbf{x}] - H(\mathbf{p},\mathbf{q})\}, \tag{4.9}$$

where $H(\mathbf{p},\mathbf{q})$ is the closed entropy function for a weighted scoring family with baseline $\mathbf{q}$. By definition, $\rho(\mathbf{x}) \geq E_{\mathbf{q}}[-\mathbf{x}] - H(\mathbf{q},\mathbf{q})$, which reduces to $-\rho(\mathbf{x}) \leq E_{\mathbf{q}}[\mathbf{x}]$ under the assumption that $H(\mathbf{q},\mathbf{q}) = 0$. Thus, a risk measure constructed using a weighted score divergence corresponds to risk aversion in that the expected risk-adjusted rewards $-\rho(\mathbf{x})$ under the baseline measure $\mathbf{q}$ are dominated by the expected value.

We now demonstrate that when we choose a risk measure ρ according to Equation (4.9), the entropy measure of the weighted scoring rule is equal to the maximal value of an optimization problem.

Proposition 4.1. *Define a convex risk measure over discrete random variables* $\mathbf{x} : \Omega \to \mathbb{R}$ *as*

$$\rho(\mathbf{x}) := \sup_{\mathbf{p}\in\mathcal{P}} \; E_{\mathbf{p}}[-\mathbf{x}] - H(\mathbf{p},\mathbf{q}),$$

where $H(\mathbf{p},\mathbf{q})$ *is the entropy for a weighted scoring family with baseline* $\mathbf{q}$*. Then we have that*

$$H(\mathbf{p},\mathbf{q}) = \sup_{\mathbf{x}\in\mathbb{R}^n} \{-\rho(\mathbf{x}) \mid E_{\mathbf{p}}[\mathbf{x}] \leq 0\}. \tag{4.10}$$

Proof. Using strong Lagrangian duality, which holds because the original optimization problem on the left-hand-side is feasible (for example, take $\mathbf{x} = \mathbf{0}$), we find

$$\sup_{\mathbf{x}\in\mathbb{R}^n} \{-\rho(\mathbf{x}) \mid E_{\mathbf{p}}[\mathbf{x}] \leq 0\} = \inf_{\lambda\geq 0} \sup_{\mathbf{x}\in\mathbb{R}^n} -\rho(\mathbf{x}) - \lambda E_{\mathbf{p}}[\mathbf{x}].$$

Substituting the definition of ρ and simplifying:

$$\inf_{\lambda \geq 0} \sup_{\mathbf{x} \in \mathbb{R}^n} \left\{ - \sup_{\mathbf{r} \in \mathcal{P}} \{ E_{\mathbf{r}}[-\mathbf{x}] - H(\mathbf{r}, \mathbf{q}) \} - \lambda E_{\mathbf{p}}[\mathbf{x}] \right\}$$

$$= \inf_{\lambda \geq 0} \sup_{\mathbf{x} \in \mathbb{R}^n} \inf_{\mathbf{r} \in \mathcal{P}} E_{\mathbf{r}}[\mathbf{x}] - \lambda E_{\mathbf{p}}[\mathbf{x}] + H(\mathbf{r}, \mathbf{q}).$$

Because the set $\mathcal{P}$ is convex and compact and because the function $H(\cdot, \mathbf{q})$ is closed, a minimax theorem attributable to [74] for convex-concave functions can be applied here to interchange the order of the inner supremum and infimum. We are then left with the equivalent problem:

$$\inf_{\lambda, \mathbf{r} \in \mathcal{P}} \sup_{\mathbf{x} \in \mathbb{R}^n} \{ E_{\mathbf{r}}[\mathbf{x}] - \lambda E_{\mathbf{p}}[\mathbf{x}] + H(\mathbf{r}, \mathbf{q}) \}. \tag{4.11}$$

The inner maximization problem is unbounded unless $\lambda \mathbf{p} = \mathbf{r}$. Because $\mathbf{r}, \mathbf{p} \in \mathcal{P}$, we have that $\lambda \mathbf{p}^\top \mathbf{1} = \mathbf{r}^\top \mathbf{1}$, which implies that $\lambda = 1$ and $\mathbf{r} = \mathbf{p}$. With these additional constraints, it is clear that the value of the optimization problem (4.11) is $H(\mathbf{p}, \mathbf{q})$. □

The optimization problem in Equation (4.10) generalizes the decision problem (4.6) described previously, although, as will be demonstrated in Section 4.5.2, problem (4.5) is also a special case. Proposition 4.1 establishes that the entropy measure underpinning every weighted scoring rule can be viewed as the optimal risk-adjusted reward available to an investor who buys or sells a contingent claim in a complete market. Alternatively, the problem can be viewed in terms of a risk averse bettor wagering against a risk neutral, non-strategic opponent with distribution $\mathbf{p}$.

In this optimization problem, the form of the entropy measure underpinning the weighted rule determines the risk attitude of the investor. The argument $\mathbf{p}$ in the entropy measure $H(\mathbf{p}, \mathbf{q})$ appears in the *constraint* of the

right side of Equation (4.10); thus, the associated scoring rule is tailored to the decision problem when forecasts are given for the underlying risk neutral distribution. When $H(\mathbf{p}, \mathbf{q})$ can be interpreted as a convex divergence in either $\mathbf{p}$ or $\mathbf{q}$, then the parameter $\mathbf{q}$ that appears in the definition of $\rho(\mathbf{x})$ can also be elicited using a tailored rule.

The remainder of this chapter is devoted to exploiting the abstract relationship established in Proposition 4.1 in useful ways. The next section introduces ϕ-Divergence and Scaled ϕ-Divergence weighted scoring rules. These rules correspond with well-studied risk measures and generalize the Weighted Power and Weighted Pseudospherical rules, respectively. We subsequently demonstrate how Proposition 4.1 can be used to generalize the theoretical results presented in [54].

4.5 ϕ-Divergence and Scaled ϕ-Divergence Weighted Scoring Rules

This section describes two general classes of weighted scoring rules based on the ϕ-Divergence measure, as well as their relationship with specific convex risk measures.

4.5.1 Definition and Properties

A ϕ-Divergence provides a measure of the difference between the distributions $\mathbf{p}$ and $\mathbf{q}$ [27]. Let

$$\frac{\mathbf{p}}{\mathbf{q}} := \left[\frac{p_1}{q_1}, \frac{p_2}{q_2}, \ldots, \frac{p_n}{q_n}\right],$$

where $n = |\Omega|$. Further, let $\phi : \mathbb{R} \to (-\infty, \infty]$ be a proper, closed, convex function such that 1 is in the interior of the domain of ϕ, and $\phi(1) = 0$. For

compactness, we introduce the notation

$$\phi\left(\frac{\mathbf{p}}{\mathbf{q}}\right) := \left[\phi\left(\frac{p_1}{q_1}\right), \phi\left(\frac{p_2}{q_2}\right), \ldots, \phi\left(\frac{p_n}{q_n}\right)\right].$$

The ϕ-divergence of $\mathbf{p}$ from $\mathbf{q}$ for $\mathbf{p}, \mathbf{q} \in \mathcal{P}$ is given by

$$H_\phi(\mathbf{p}, \mathbf{q}) := E_{\mathbf{q}}\left[\phi\left(\frac{\mathbf{p}}{\mathbf{q}}\right)\right] = \sum_{i=1}^{n} q_i \phi\left(\frac{p_i}{q_i}\right), \tag{4.12}$$

where $0\phi(\frac{0}{0}) := 0$, $0\phi(\frac{p}{0}) := \lim_{q\to 0}\ q\phi(\frac{p}{q})$, and $q\phi(\frac{0}{q}) := \lim_{p\to 0}\ q\phi(\frac{p}{q})$. The ϕ-divergence is jointly convex in $\mathbf{p}$ and $\mathbf{q}$, and an application of Jensen's inequality demonstrates that $H_\phi(\mathbf{p}, \mathbf{q})$ attains a minimum of 0 by setting $\mathbf{p} = \mathbf{q}$. In general, ϕ-divergences are not symmetric in $\mathbf{p}$ and $\mathbf{q}$, and thus cannot be regarded as a distance metric. The Kullback-Leibler divergence is a special case when $\phi(t) = t\ln(t) - t + 1$. Given a convex function ϕ, we define the *adjoint* function $\hat{\phi}(t) = t\phi\left(\frac{1}{t}\right)$ and note that $\hat{\phi}$ is also convex. The divergence associated with $\hat{\phi}$ satisfies the relationship $H_{\hat{\phi}}(\mathbf{p}, \mathbf{q}) = H_\phi(\mathbf{q}, \mathbf{p})$.

Viewing $\mathbf{q}$ as the baseline, it is clear that $H_\phi(\mathbf{p}, \mathbf{q})$ is the entropy function for a weighted scoring rule. Indeed, this rule is a special case of a *quasi-Bregman* weighted scoring rule, introduced in [36]. We now derive the form of the weighted scoring rule associated with the divergence $H_\phi(\mathbf{p}, \mathbf{q})$.

Proposition 4.2. *Let $v_i \in \partial\phi\left(\frac{r_i}{q_i}\right)$ for $i = 1, \ldots, n$. Then the scoring rule $S : \Omega \times \mathcal{P} \to [-\infty, \infty)$ given by*

$$S_\phi(\omega_i, \mathbf{r}, \mathbf{q}) = H_\phi(\mathbf{r}, \mathbf{q}) + v_i - \sum_{j=1}^{n} r_j v_j \tag{4.13}$$

is a strictly proper scoring rule satisfying $E_{\mathbf{p}}[S_\phi(x, \mathbf{p}, \mathbf{q})] = H_\phi(\mathbf{p}, \mathbf{q})$.

Proof. This follows immediately from Theorem 4.1 and standard results concerning subgradient calculus [67]. □

In Equation (4.13), the scoring function is fully determined by the subgradients of the univariate convex function ϕ.

Example 4.2. *Weighted Power Scoring Rule*

We illustrate Proposition 4.2 by deriving the scoring function $S_\beta^P(x, \mathbf{r}, \mathbf{q})$ *associated with the Weighted Power scoring rule in Equation* (4.1)*. The entropy measure for the Weighted Power scoring rule,* $H_\beta^P(\mathbf{p}, \mathbf{q})$*, is a* ϕ*-divergence with* $\phi(t) = \frac{t^\beta - 1}{\beta(\beta-1)}$*,* $\beta \in \mathbb{R}$ *[54]. Differentiating* $\phi(t)$ *yields*

$$\phi'(t) = \frac{\beta t^{(\beta-1)}}{\beta(\beta-1)}.$$

From (4.13)*, it follows that the weighted scoring rule associated with* $H_\phi(\mathbf{p}, \mathbf{q})$ *is given by*

$$\begin{aligned} S_\beta^P(x, \mathbf{r}, \mathbf{q}) =& \frac{E_\mathbf{q}\left[(\frac{\mathbf{r}}{\mathbf{q}})^\beta\right] - 1}{\beta(\beta-1)} + \frac{\beta\left(\frac{r_x}{q_x}\right)^{(\beta-1)}}{\beta(\beta-1)} - \frac{\beta E_\mathbf{r}\left[\left(\frac{\mathbf{r}}{\mathbf{q}}\right)^{(\beta-1)}\right]}{\beta(\beta-1)} \\ =& \frac{\left(\frac{r_x}{q_x}\right)^{(\beta-1)} - 1}{(\beta-1)} - \frac{E_\mathbf{r}\left[\left(\frac{\mathbf{r}}{\mathbf{q}}\right)^{(\beta-1)}\right] - 1}{\beta}. \end{aligned} \tag{4.14}$$

S_β^P *matches exactly the Weighted Power score function described in [54].*

This choice of ϕ *yields the adjoint* $\hat{\phi}(t) = \frac{t^{1-\beta} - t}{\beta(\beta-1)}$ *and*

$$H_{\hat{\phi}}(\mathbf{p}, \mathbf{q}) = \frac{E_\mathbf{q}\left[\left(\frac{\mathbf{p}}{\mathbf{q}}\right)^{1-\beta} - \frac{\mathbf{p}}{\mathbf{q}}\right]}{\beta(\beta-1)} = \frac{E_\mathbf{q}\left[\left(\frac{\mathbf{p}}{\mathbf{q}}\right)^{1-\beta}\right] - 1}{\beta(\beta-1)} = H_{1-\beta}^P(\mathbf{p}, \mathbf{q}).$$

By the definition of the adjoint, the right side is equal to $H_\beta^P(\mathbf{q}, \mathbf{p})$*.*

If any component of the distribution $\mathbf{r}$ or $\mathbf{q}$ takes the value 0, ϕ-Divergence scoring rules may not be well-defined. First, we examine the scenario where

the baseline $\mathbf{q}$ is in $\mathcal{P}^+$ and r_i is 0 for some $\omega_i \in \Omega$. In this case, the divergence $H_\phi(\mathbf{r}, \mathbf{q})$ is finite when $\lim_{t\to 0} \phi(t) < \infty$. If 0 is in the interior of the domain of ϕ, then this limit will be finite and the subgradients of ϕ at 0 will be non-vertical [67]. Thus, this criterion provides a sufficient condition for the scoring rule in Equation (4.13) to be well-defined when $r_i = 0$.

We now turn to the case where $\mathbf{r} > 0$ and $q_i = 0$ for at least one $\omega_i \in \Omega$. This scenario is likely in applications where the baseline $\mathbf{q}$ is derived from observed frequencies. For example, when forecasting rare events, or in a setting where data are sparse, an outcome ω_i may have no empirical realizations (meaning $q_i = 0$ in the empirical distribution), but nevertheless could potentially have positive probability. In this case, there is still interest in obtaining a forecast for the positive probability of observing ω_i.

When $r_i > 0, q_i = 0$,

$$\lim_{q_i \to 0} q_i \phi\left(\frac{r_i}{q_i}\right) = r_i \lim_{t\to\infty} \frac{\phi(t)}{t}. \tag{4.15}$$

When the limit on the right-hand side is finite, $H_\phi(\mathbf{r}, \mathbf{q})$ will be as well. Alternatively, if $\lim_{t\to\infty} \phi(t)/t = \infty$, then $H_\phi(\mathbf{r}, \mathbf{q}) = \infty$. In this latter case, it follows from Equation (4.13) that the score $S(\omega_i, \mathbf{r}, \mathbf{q})$ will be infinite or undefined as well.

On the other hand, suppose that $\lim_{t\to\infty} \phi(t)/t = C < \infty$. Then for any $v(t_0) \in \partial\phi(t_0)$ the subgradient inequality implies

$$\phi(t) \geq \phi(t_0) + v(t_0)(t - t_0) \quad \Longrightarrow \quad \frac{\phi(t)}{t} \geq \frac{\phi(t_0)}{t} + v(t_0)\frac{(t - t_0)}{t}$$

for all $t > 0$. Taking limits on both sides yields $C \geq v(t_0)$. As this holds for all $t_0 > 0$, we must have $\lim_{t\to\infty} v(t) < \infty$ so that each term in (4.13) is finite. Thus, the scoring rule (4.13) is well-defined in this case. These observations are summarized in the following proposition.

Proposition 4.3. *Let S_ϕ be a proper scoring rule with entropy function H_ϕ of the form of Equation (4.12). Suppose that $q_i = 0, r_i > 0$ for some $\omega_i \in \Omega$. If $H_\phi(\mathbf{r}, \mathbf{q}) < \infty$, then $S_\phi(\omega_i, \mathbf{r}, \mathbf{q}) \in \mathbb{R}$ for all $\mathbf{r} \in \mathcal{P}$. Otherwise, if $H_\phi(\mathbf{r}, \mathbf{q}) = \infty$, then S_ϕ is infinite or undefined for all $\mathbf{r}$ with $r_i > 0$.*

In the context of distributionally robust optimization, ϕ-divergences can be used to define an ambiguity set of the form $\{\mathbf{r} \in \mathcal{P} : H_\phi(\mathbf{r}, \mathbf{q}) \leq z\}$, where $\mathbf{q}$ is often taken to be an empirical distribution derived from observations. The ϕ-divergences that are finite when $q_i = 0, r_i > 0$ are said to be able to *pop* scenarios [7]. For these divergence measures, the consideration of distributions $\mathbf{r}$ is not limited to those that are absolutely continuous with respect to the empirical distribution. By Proposition 4.3, ϕ-divergences that pop scenarios will also have well-defined score functions when $q_i = 0$ for some ω_i.

Next, we introduce the class of Scaled ϕ-Divergence scoring rules, which are defined in terms of the entropy measure

$$\bar{H}_\phi(\mathbf{p}, \mathbf{q}) := \inf_{\lambda \geq 0} E_{\mathbf{p}} \left[\phi \left(\lambda \frac{\mathbf{q}}{\mathbf{p}} \right) \right], \tag{4.16}$$

where ϕ is a convex function satisfying the conditions described in the definition of a ϕ-divergence, with the added stipulation that $\phi(1) = 0 = \min_{t \geq 0} \phi(t)$. In the derivation of the scoring rule, we also require $\mathbf{q} << \mathbf{p}$, meaning that the measure $\mathbf{q}$ is absolutely continuous with respect to $\mathbf{p}$. This allows the baseline forecast to contain 0 values, but it restricts the scored forecast to be positive wherever $\mathbf{q}$ is. The following lemma summarizes the important properties of the entropy measure (4.16).

Lemma 4.1. *Let $\mathbf{q} << \mathbf{p}$. Then*

1. *The function $f(\mathbf{p}, \lambda) := E_{\mathbf{p}} \left[\phi \left(\lambda \frac{\mathbf{q}}{\mathbf{p}} \right) \right]$ is jointly convex in $\mathbf{p}, \lambda$.*

2. $\bar{H}_\phi(\mathbf{p}, \mathbf{q}) := \inf_{\lambda > 0} E_{\mathbf{p}} \left[\phi \left(\lambda \frac{\mathbf{q}}{\mathbf{p}} \right) \right]$.

Proof. Indirect discussion of the convexity properties and optimizer of the expression

$$\bar{H}_\phi(\mathbf{p}, \mathbf{q}) := \inf_{\lambda \geq 0} E_{\mathbf{p}} \left[\phi \left(\lambda \frac{\mathbf{q}}{\mathbf{p}} \right) \right] \tag{4.17}$$

is provided in [41, 35, 11]. For completeness, a direct proof follows.

1. Joint convexity of the function $f(\mathbf{p}, \lambda) = E_{\mathbf{p}} \left[\phi \left(\lambda \frac{\mathbf{q}}{\mathbf{p}} \right) \right]$, for $\mathbf{p} > \mathbf{0}$ and $\lambda \geq 0$, follows from the joint convexity of the function $p_i \phi(q_i \lambda / p_i)$ in λ and p_i. The latter is established by recognizing $p_i \phi(q_i \lambda / p_i)$ as the composition of the perspective function of ϕ with the affine map $(\lambda, p_i) \to (q_i \lambda, p_i)$.

   We now extend joint convexity to $\mathbf{p}$ with $p_i = 0$. Consider two points (λ^1, p_i^1) and (λ^2, p_i^2) such that either $p_i^1 = 0$ or $p_i^2 = 0$, which implies that $q_i = 0$. In this case, for any $0 \leq \gamma \leq 1$, we have

$$(\gamma p_i^1 + (1-\gamma) p_i^2) \phi \left(\frac{(\gamma \lambda^1 + (1-\gamma)\lambda^2) q_i}{\gamma p_i^1 + (1-\gamma) p_i^2} \right) = \gamma p_i^1 \phi(0) + (1-\gamma) p_i^2 \phi(0)$$

$$= \gamma p_i^1 \phi \left(\frac{\lambda^1 q_i}{p_i^1} \right) + (1-\gamma) p_i^2 \phi \left(\frac{\lambda^2 q_i}{p_i^2} \right).$$

   This extends joint convexity to $\{(\lambda, \mathbf{p}) : \lambda \geq 0, \mathbf{q} << \mathbf{p}\}$.

2. If either $q_i = 0$ or $p_i = q_i = 0$, then the equation $p_i \phi \left(\frac{\lambda q_i}{p_i} \right) = p_i \phi(0)$ does not depend on λ. Thus, the optimal λ solves

$$\inf_{\lambda \geq 0} \sum_{i : p_i > 0, q_i > 0} p_i \phi \left(\frac{\lambda q_i}{p_i} \right).$$

Choose $\lambda' = \min_i \ p_i/q_i > 0$. We have

$$\sum_{i:p_i>0,q_i>0} p_i \phi\left(\frac{\lambda' q_i}{p_i}\right) = \sum_{i:p_i>0,q_i>0} p_i \phi(t_i),$$

where $0 < t_i \leq 1$ for each i. By the convexity of ϕ and by $\phi(0) \geq \phi(1)$, we obtain

$$\sum_{i:p_i>0,q_i>0} p_i \phi\left(\frac{\lambda' q_i}{p_i}\right) \leq \sum_{i:p_i>0,q_i>0} p_i \phi(0),$$

which implies that $\lambda = 0$ is suboptimal.

□

Under the conditions outlined above, the following proposition establishes that $\bar{H}_\phi(\mathbf{p}, \mathbf{q})$ is a convex divergence measure and thus generates a proper weighted scoring rule.

Proposition 4.4. *For* $\mathbf{q} << \mathbf{p}$*, the function*

$$\bar{H}_\phi(\mathbf{p}, \mathbf{q}) := \inf_{\lambda>0} E_{\mathbf{p}}\left[\phi\left(\lambda \frac{\mathbf{q}}{\mathbf{p}}\right)\right]$$

is a convex divergence measure in $\mathbf{p}$*.*

Proof. Joint convexity of the function $f(\mathbf{p}, \lambda) = E_{\mathbf{p}}\left[\phi\left(\lambda \frac{\mathbf{q}}{\mathbf{p}}\right)\right]$ is established for $\mathbf{q} << \mathbf{p}$ in the first part of Lemma 4.1. Convexity of $\bar{H}(\mathbf{p}, \mathbf{q})$ in $\mathbf{p}$ then follows from partial minimization [19].

Moreover, by Jensen's inequality and for any $\lambda > 0$, we have

$$\inf_{\mathbf{p}\in\mathcal{P}} E_{\mathbf{p}}\left[\phi\left(\lambda \frac{\mathbf{q}}{\mathbf{p}}\right)\right] \geq \phi(\lambda).$$

It follows that

$$\inf_{\mathbf{p}\in\mathcal{P}} \inf_{\lambda>0} E_{\mathbf{p}}\left[\phi\left(\lambda \frac{\mathbf{q}}{\mathbf{p}}\right)\right] = \inf_{\lambda>0} \inf_{\mathbf{p}\in\mathcal{P}} E_{\mathbf{p}}\left[\phi\left(\lambda \frac{\mathbf{q}}{\mathbf{p}}\right)\right] \geq \inf_{\lambda>0} \phi(\lambda) = 0,$$

where the last equality follows from the fact that $0 = \min_{t \geq 0} \phi(t)$. The lower bound is obtained by setting $\mathbf{p} = \mathbf{q}$, which proves the result. □

We have that $\bar{H}_\phi(\mathbf{p}, \mathbf{q}) \leq H_\phi(\mathbf{q}, \mathbf{p}) = H_{\hat{\phi}}(\mathbf{p}, \mathbf{q})$, which means that the optimal expected score under a weighted scoring rule generated from Equation (4.16) is always lower than the optimal expected score under the weighted scoring rule with entropy $H_{\hat{\phi}}(\mathbf{p}, \mathbf{q})$. The adjoint function, $\hat{\phi}(t) = t\phi(1/t)$, also attains its minimum at $t = 1$. Thus, the divergence measure

$$\bar{H}_{\hat{\phi}}(\mathbf{p}, \mathbf{q}) = \inf_{\lambda > 0} E_{\mathbf{p}} \left[\hat{\phi} \left(\lambda \frac{\mathbf{q}}{\mathbf{p}} \right) \right] = \inf_{\lambda > 0} E_{\mathbf{q}} \left[\lambda \phi \left(\frac{\mathbf{p}}{\lambda \mathbf{q}} \right) \right]$$

also underpins a weighted proper scoring rule.

The next proposition gives a method for finding the score function associated with $\bar{H}(\mathbf{p}, \mathbf{q})$.

Proposition 4.5. *Let* $\lambda^* \geq 0$, $v_i \in \partial\phi \left(\lambda^* \frac{q_i}{r_i} \right)$ *satisfy*

$$0 = \sum_{i=1}^{n} q_i v_i. \tag{4.18}$$

Then

$$\bar{S}(\omega_i, \mathbf{r}, \mathbf{q}) = \phi \left(\lambda^* \frac{q_i}{r_i} \right) - \lambda^* \frac{q_i}{r_i} v_i + \lambda^* \sum_{j=1}^{n} q_j v_j \tag{4.19}$$

is a proper scoring rule for $\mathbf{q} << \mathbf{r}$, *with* $E_{\mathbf{p}}[\bar{S}(\omega, \mathbf{p}, \mathbf{q})] = \bar{H}(\mathbf{p}, \mathbf{q})$.

Proof. Let $\mathbf{q} << \mathbf{p}$. First, note that

$$E_{\mathbf{p}}[\bar{S}(x, \mathbf{p}, \mathbf{q})] = E_{\mathbf{p}} \left[\phi \left(\lambda^* \frac{\mathbf{q}}{\mathbf{p}} \right) \right].$$

Furthermore, convexity of $E_{\mathbf{p}} \left[\phi \left(\lambda \frac{\mathbf{q}}{\mathbf{p}} \right) \right]$ in λ (Lemma 4.1) and condition (4.18) imply that λ^* is the optimal solution in

$$\inf_{\lambda \geq 0} E_{\mathbf{p}} \left[\phi \left(\lambda \frac{\mathbf{q}}{\mathbf{p}} \right) \right].$$

Thus, $E_{\mathbf{p}}[\bar{S}(x, \mathbf{p}, \mathbf{q})] = \bar{H}(\mathbf{p}, \mathbf{q})$.

Next, we show that $\bar{S}$ is strictly proper. We have

$$E_{\mathbf{p}}[\bar{S}(x, \mathbf{r}, \mathbf{q})] = \sum_i p_i \left[\phi\left(\lambda^* \frac{q_i}{r_i}\right) - \lambda^* \frac{q_i}{r_i} v_i + \lambda^* \sum_{j=1}^{n} q_j v_j \right].$$

We split the sum according to

$$E_{\mathbf{p}}[\bar{S}(x, \mathbf{r}, \mathbf{q})] =$$

$$\sum_{i:p_i>0, r_i>0} p_i \left[\phi\left(\lambda^* \frac{q_i}{r_i}\right) - \lambda^* \frac{q_i}{r_i} v_i + \lambda^* \sum_{j=1}^{n} q_j v_j \right]$$

$$+ \sum_{i:p_i=0 \text{ or } r_i=0} p_i \left[\phi\left(\lambda^* \frac{q_i}{r_i}\right) - \lambda^* \frac{q_i}{r_i} v_i + \lambda^* \sum_{j=1}^{n} q_j v_j \right].$$

That $\mathbf{q} << \mathbf{r}$ and $\mathbf{q} << \mathbf{p}$, along with Equation (4.18), imply that the right side sum is equal to

$$\sum_{i:p_i=0 \text{ or } r_i=0} p_i \phi(0) = \sum_{i:p_i=0 \text{ or } r_i=0} p_i S(\omega_i, \mathbf{p}, \mathbf{q}). \tag{4.20}$$

We now turn to the left side sum. As an optimal solution to

$$\inf_{\lambda \geq 0} E_{\mathbf{p}} \left[\phi\left(\lambda \frac{\mathbf{q}}{\mathbf{r}}\right) \right],$$

Lemma 4.1 implies $\lambda^* > 0$. Using Equation (4.18), the left side sum can then be written as

$$\sum_{i:p_i>0, r_i>0} p_i \left[\phi\left(\lambda^* \frac{q_i}{r_i}\right) - \lambda^* \frac{q_i}{r_i} v_i + \left(\frac{\lambda}{\lambda^*}\right) \lambda^* \sum_{j=1}^{n} q_j v_j \right],$$

where $\lambda > 0$ is arbitrary. Rearranging and factoring the term p_i yields the equivalent sum

$$\sum_{i:p_i>0,r_i>0} p_i \left[\phi\left(\lambda^* \frac{q_i}{r_i}\right) - v_i \left(\frac{r_i}{p_i}\frac{\lambda q_i}{r_i} - \frac{\lambda^* q_i}{r_i}\right)\right] \leq \sum_{i:p_i>0,r_i>0} p_i \phi\left(\lambda \frac{q_i}{p_i}\right),$$

where the inequality follows from $v_i \in \partial\phi\left(\lambda^* \frac{q_i}{r_i}\right)$. Finally, recombining the sums gives

$$\sum_{i:p_i>0} p_i \phi\left(\lambda \frac{q_i}{p_i}\right) + \sum_{i:p_i=0 \text{ or } r_i=0} p_i \phi(0) = E_{\mathbf{p}}\left[\phi\left(\lambda \frac{\mathbf{q}}{\mathbf{p}}\right)\right] \geq E_{\mathbf{p}}[\bar{S}(x, \mathbf{r}, \mathbf{q})].$$

The result then follows by taking the infimum with respect to $\lambda > 0$ on the left side of the inequality.

□

We close this section by demonstrating that Pseudospherical weighted scoring rules are special cases of Scaled ϕ-Divergence rules with the choice

$$\phi(t) = \frac{t^{1-\beta} - (1-\beta)t - \beta}{\beta(\beta-1)}. \tag{4.21}$$

In this case, the optimization problem (4.16) becomes

$$\inf_{\lambda \geq 0} \frac{1}{\beta(\beta-1)} \lambda E_{\mathbf{q}}\left[\left(\frac{\mathbf{p}}{\lambda \mathbf{q}}\right)^{\beta}\right] + \frac{\lambda}{\beta} + \frac{1}{(1-\beta)}. \tag{4.22}$$

Solving for the optimal λ yields

$$\lambda^* = \left[E_{\mathbf{q}}\left[\left(\frac{\mathbf{p}}{\mathbf{q}}\right)^{\beta}\right]\right]^{(1/\beta)}.$$

Substituting λ^* into (4.22) gives

$$\bar{H}_{\phi}(\mathbf{p}, \mathbf{q}) = H_{\beta}^{S}(\mathbf{p}, \mathbf{q}) = \frac{E_{\mathbf{q}}[(\frac{\mathbf{p}}{\mathbf{q}})^{\beta}]^{(1/\beta)} - 1}{\beta - 1}.$$

Up to this point, we have defined ϕ-Divergence and Scaled ϕ-Divergence scoring rules, derived the functional form of the score function, and identified conditions for which the score function is well-defined. Next, we demonstrate that these rules are tailored to a betting problem involving a convex risk measure.

4.5.2 ϕ-Divergence Scoring Rules and the Optimized Certainty Equivalent

This section will show that ϕ-Divergence scoring rules are tailored to a generic decision problem involving a risk averse investor placing wagers against a risk neutral opponent with the goal of maximizing his Optimized Certainty Equivalent (OCE). For the rest of the chapter, we assume that the decision maker's risk preferences are captured by a utility function having the following form.

Definition 4.2. *Normalized Utility Function [11]*

We will call a utility function $u : \mathbb{R} \to [-\infty, \infty)$ normalized if u is concave and non-decreasing with $-u$ a closed function. Furthermore, we assume u satisfies

$$u(0) = 0, 1 \in \hat{\partial} u(0),$$

where $\hat{\partial} u(x)$ is the superdifferential of u at x.

The domain of u, $\mathrm{dom}(u)$, is all points on the real line for which $u(x) > -\infty$, and we are assuming that 0 is in the interior of the domain. The HARA utility functions given in Equation (4.7) are special cases.

For a proper, closed, convex function $\phi : \mathbb{R} \to (-\infty, \infty)$, the *convex*

conjugate function [19] is given by:

$$\phi^*(t) = \sup_{x \in \text{dom}(\phi)} \{xt - \phi(x)\}.$$

The function ϕ^* is closed and convex, and since ϕ is closed, we have $(\phi^*)^* = \phi$ [67]. The domain of ϕ^* is those $t \in \mathbb{R}$ for which the supremum in (4.5.2) is finite. Similarly, for a concave function u, the concave conjugate function is defined as $u^*(t) = \inf_{x \in \text{dom}(u)} \{xt - u(x)\}$. Therefore, we have

$$-u^*(t) = \sup_{x \in \text{dom}(u)} \{-xt - (-u(x))\} = (-u)^*(-t),$$

making $-u^*(t)$ a closed, convex function. We will use the following lemma proved in the Appendix for this chapter.

Lemma 4.2. *Let $u : \mathbb{R} \to [-\infty, \infty)$ be a normalized utility function. Then $-u^*(t)$ is minimized at $t = 1$ and $-u^*(1) = 0$.*

As demonstrated in [11], the optimal expected score for ϕ-Divergence weighted scoring rules is related to an optimization problem involving the OCE, defined below.

Definition 4.3. *Optimized Certainty Equivalent [10]*

Let u be a normalized utility function and let X be a random variable. The optimized certainty equivalent (OCE) is defined as

$$OCE_u(X) := \sup_{\nu \in \mathbb{R}} \nu + E[u(X - \nu)].$$

The negative OCE is a convex risk measure.

The OCE interprets the value of a random position as a decision problem. X represents an uncertain future stream of income and ν captures the

level of current spending. The value of $OCE_u(X)$ is the expected reward to the decision maker under the optimal allocation of immediate and future consumption. The OCE and ϕ-divergences are linked, as demonstrated by the following result [11].

Theorem 4.3. *[11]*

Let ϕ be a proper, closed convex function and set $u(t) = -\phi^*(-t)$. Then for any $\mathbf{q} << \mathbf{p}$

$$H_\phi(\mathbf{q}, \mathbf{p}) = E_{\mathbf{p}}\left[\phi\left(\frac{\mathbf{q}}{\mathbf{p}}\right)\right] = \sup_{\mathbf{x}\in\mathbb{R}^n}\{OCE_u(\mathbf{x}) \mid E_{\mathbf{q}}[\mathbf{x}] = 0,\}$$

where $OCE_u(\mathbf{x}) = \sup_{\nu\in\mathbb{R}}\ \nu + E_{\mathbf{p}}[u(\mathbf{x} - \nu\mathbf{1})]$ (with $\mathbf{1}$ an n-dimensional vector of $1s$).

As we now show, the optimization problem involving the OCE described above is equivalent to the following generalization of the utility maximization problem considered in [54]:

$$z(\mathbf{p}, \mathbf{q}) = \sup_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u(\mathbf{x})] - E_{\mathbf{q}}[\mathbf{x}], \tag{4.23}$$

where $\mathbf{p}, \mathbf{q} \in \mathcal{P}$. We recover the optimization problem (4.5) considered in [54] by taking u to be a HARA utility function.

We will use Theorem 4.3 to prove the following.

Proposition 4.6. *Set $\phi(t) = -u^*(t)$ and denote by $\hat{\phi}$ the adjoint of ϕ. For $\mathbf{q} << \mathbf{p}$,*

$$H_{\hat{\phi}}(\mathbf{p}, \mathbf{q}) := E_{\mathbf{q}}\left[\hat{\phi}\left(\frac{\mathbf{p}}{\mathbf{q}}\right)\right]$$

is a ϕ-divergence and $H_{\hat{\phi}}(\mathbf{p}, \mathbf{q}) = z(\mathbf{p}, \mathbf{q})$.

Proof. Because $-u$ is convex and closed, we have that $\phi(t) = -u^*(t)$ if, and only if, $-\phi^*(-t) = u(t)$. Now observe that

$$z(\mathbf{p}, \mathbf{q}) = \sup_{x \in \mathbb{R}^n} E_{\mathbf{p}}[u(\mathbf{x})] - E_{\mathbf{q}}[\mathbf{x}] = \sup_{\mathbf{x} \in \mathbb{R}^n, \nu \in \mathbb{R}} \{E_{\mathbf{p}}[u(\mathbf{x})] + \nu \mid E_{\mathbf{q}}[\mathbf{x} + \nu \mathbf{1}] = 0, \}$$

where we have introduced the variable $\nu = -E_{\mathbf{q}}[\mathbf{x}]$. Making the substitution $\mathbf{y} = \mathbf{x} + \nu \mathbf{1}$, we obtain

$$\begin{aligned} z(\mathbf{p}, \mathbf{q}) &= \sup_{\mathbf{y} \in \mathbb{R}^n, \nu \in \mathbb{R}} \{E_{\mathbf{p}}[u(\mathbf{y} - \nu \mathbf{1})] + \nu \mid E_{\mathbf{q}}[\mathbf{y}] = 0\} \\ &= \sup_{\mathbf{y} \in \mathbb{R}^n} \{OCE_u(\mathbf{y}) \mid E_{\mathbf{q}}[\mathbf{y}] = 0\} \\ &= H_\phi(\mathbf{q}, \mathbf{p}) = H_{\hat{\phi}}(\mathbf{p}, \mathbf{q}), \end{aligned}$$

where the penultimate equality follows from Theorem 4.3. □

Proposition 4.6 demonstrates that ϕ-Divergence scoring rules are tailored to decision problem (4.23) and generalizes Theorem 1 (Part b) given in [54] to the class of normalized utility functions. It also provides an alternative interpretation for the optimal expected score function associated with the Weighted Power scoring rule. Taking $u_\beta(x)$ to be of the form in Equation (4.7), we find that the conjugate function $-u^*_\beta(t)$ is exactly as in Equation (4.21). Thus,

$$\hat{\phi}(t) = \frac{t^\beta - (1 - \beta) - \beta t}{\beta(\beta - 1)}.$$

It follows that

$$H_{\hat{\phi}}(\mathbf{p}, \mathbf{q}) = \frac{E_{\mathbf{q}}[(\frac{\mathbf{p}}{\mathbf{q}})^\beta] - 1}{\beta(\beta - 1)} = H^P_\beta(\mathbf{p}, \mathbf{q}).$$

By Proposition 4.6, we have

$$H^P_\beta(\mathbf{p}, \mathbf{q}) = \sup_{\mathbf{x} \in \mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] - E_{\mathbf{q}}[\mathbf{x}] = \sup_{\mathbf{y} \in \mathbb{R}^n} \left\{OCE_{u_\beta}(\mathbf{y}) \mid E_{\mathbf{q}}[\mathbf{y}] = 0\right\}.$$

Therefore, $H_\beta^P(\mathbf{p}, \mathbf{q})$ can be interpreted as the maximum OCE for a risk averse decision maker with HARA utility function betting against a risk neutral opponent.

Although, in real-world instances of problem (4.5), forecasts $\mathbf{p}$ that are not component-wise strictly positive are unlikely to be reasonable, the requirement $\mathbf{q} << \mathbf{p}$ is not necessary for this economic interpretation of ϕ-Divergence scoring rules to hold. Proposition 4.10 (presented in Section 4.7.2) asserts the more general version without this supposition.

Conversely, the distribution $\mathbf{q}$ containing 0s does not present technical problems. When $q_x = 0$ and $p_x > 0$, the optimization problem (4.5) is bounded above only when $\sup_{x \in \text{dom}(u)} u(x)$ is bounded. The latter criterion is equivalent to 0 being in the domain of $\phi(t) = -u^*(t)$, which in turn implies that $H_\phi(\mathbf{q}, \mathbf{p}) < \infty$.

Proposition 4.6 can be expanded to give a slightly different economic interpretation for ϕ-Divergence weighted scoring rules. For $\phi(t) = -u^*(t)$, we must also have

$$H_\phi(\mathbf{q}, \mathbf{p}) = \sup_{\mathbf{x} \in \mathbb{R}^n} \left\{ OCE_u(\mathbf{x}) \mid E_{\mathbf{q}}[\mathbf{x}] \leq 0 \right\}, \tag{4.24}$$

where we have relaxed the constraint $E_{\mathbf{q}}[\mathbf{x}] = 0$ in Theorem 4.3 to an inequality constraint. Indeed, any feasible solution $\mathbf{x}'$ to (4.24) satisfying $E_{\mathbf{q}}[\mathbf{x}'] < 0$ must have components satisfying $x_j < 0$ for j in some set $A \subseteq \{1, \ldots, n\}$. For the vector $\mathbf{y}$, with $y_j = 1$ for $j \in A$ and 0 otherwise, we have that $f(t) = OCE_u(\mathbf{x} + t\mathbf{y})$ is non-decreasing in t. Furthermore, there is some $t > 0$ for which $E_{\mathbf{q}}[\mathbf{x} + t\mathbf{y}] = 0$. Therefore, there exists an optimal solution for (4.24) that satisfies the stronger equality constraint.

The relation in Equation (4.24) is a special case of Proposition 4.1 with $\rho(\mathbf{x}) = -OCE_u(\mathbf{x})$. We can find an incentive-compatible ϕ-Divergence scoring rule for a decision maker seeking to maximize her OCE when either $\mathbf{p}$ or $\mathbf{q}$ is obtained from a forecast. When $\mathbf{q}$ is a known baseline distribution, and $\mathbf{p}$ is elicited from an expert, we can choose the scoring rule with entropy measure $H_{\hat{\phi}}(\mathbf{p}, \mathbf{q})$. Alternatively, when $\mathbf{p}$ is fixed and the distribution $\mathbf{q}$ will be obtained from a forecast, the scoring rule with entropy $H_{\phi}(\mathbf{q}, \mathbf{p})$ will align incentives.

We can also directly construct the proper scoring function S that is tailored to problem (4.5). For the derivation, assume that $\mathbf{q} > 0$ and that u is specified so that the optimal bets and expected score are finite. Further, assume for simplicity that the possible forecasts $\mathbf{r}$ are also component-wise positive. Under these assumptions, the score for a forecast $\mathbf{r}$ is exactly the ex post rewards for a decision maker who places bets assuming that $\mathbf{r}$ is the true distribution. Specifically, suppose the decision maker receives the forecast $\mathbf{r}$ and subsequently solves the optimization problem $\sup_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{r}}[u(\mathbf{x})] - E_{\mathbf{q}}[\mathbf{x}]$. In this case, after the optimal bets $\mathbf{x}^*$ are determined, and scenario $i \in \{1, \ldots, n\}$ is realized, the decision maker receives a reward of $u_i(\mathbf{x}^*) := u(x_i^*) - E_{\mathbf{q}}[\mathbf{x}^*]$. As we now show, the utility gain in scenario i can be interpreted as a ϕ-Divergence score $S_{\hat{\phi}}$ of the form in Equation (4.13).

Proposition 4.7. *Set $\phi(t) = -u^*(t), t \geq 0$. Define the score function*

$$S_{\hat{\phi}}(\omega_i, \mathbf{r}, \mathbf{q}) := u_i(\mathbf{x}^*),$$

where $\omega_i \in \Omega$ and $\mathbf{x}^* \in \mathbb{R}^n$ is the optimal solution to Equation (4.5) with $\mathbf{p}$ replaced by $\mathbf{r}$. Then $S = S_{\hat{\phi}}$, where the latter is of the form in Equation (4.13).

Proof. See the Chapter Appendix. □

The scoring function $S_{\hat{\phi}}(\omega_i, \mathbf{p}, \mathbf{q}) = u_i(\mathbf{x}^*)$ is the tailored scoring rule for betting problem (4.5) following the general construction method given in [52]. Propositions 4.6 and 4.7 justify incorporating the ratio r_i/q_i in the reward function S of a weighted scoring rule. The value of the report $\mathbf{r}$ relative to the baseline $\mathbf{q}$ can be measured in terms of the monetary rewards available to a decision maker betting according to $\mathbf{r}$ versus $\mathbf{q}$. As we have seen, under this paradigm, a generic risk averse decision maker should award a score that depends on the level of risk aversion, captured by ϕ, and the aforementioned ratio of the probabilities.

The above has shown that the ϕ-Divergence scoring rules (and thus the Weighted Power scoring rules) are related to an optimization problem involving a concave certainty equivalent. We next demonstrate that the entropy measure underlying the Pseudospherical scoring rule will have a similar interpretation.

4.5.3 Scaled ϕ-Divergence Scoring Rules and the u-Mean Certainty Equivalent

Like ϕ-Divergence scoring rules, Scaled ϕ-Divergence scoring rules are connected with a risk measure, in this case the u-Mean Certainty Equivalent. Before demonstrating this result, we show that Scaled ϕ-Divergence scoring rules are tailored to the simple betting problem

$$\bar{z}(\mathbf{p}, \mathbf{q}) = \sup_{\mathbf{x} \in \mathbb{R}^n} \left\{ E_{\mathbf{p}}[u(\mathbf{x})] \mid E_{\mathbf{q}}[\mathbf{x}] \leq 0 \right\}. \tag{4.25}$$

This concave maximization problem generalizes optimization problem (4.6) (Problem S in [54]), where u is assumed to be a HARA utility function, and has an identical interpretation.

Proposition 4.8. *Set $\phi(t) = -u^*(t)$, and suppose that $\mathbf{q} << \mathbf{p}$. Then*

$$\bar{H}_\phi(\mathbf{p}, \mathbf{q}) = \bar{z}(\mathbf{p}, \mathbf{q}).$$

Proof. The function $\phi(t) = -u^*(t)$ is convex and, by Lemma 4.2, satisfies $\phi(1) = 0 = \min_{t \geq 0} \phi(t)$. Further, by weak Lagrangian duality, we have:

$$\begin{aligned}
\bar{z}(\mathbf{p}, \mathbf{q}) &\leq \inf_{\lambda \geq 0} \sup_{\mathbf{x} \in \mathbb{R}^n} E_{\mathbf{p}}[u(\mathbf{x})] - \lambda E_{\mathbf{q}}[\mathbf{x}] \\
&= \inf_{\lambda \geq 0} \left\{ - \inf_{\mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^{n} p_i \left[\lambda \frac{q_i}{p_i} x_i - u(x_i) \right] \right\} \\
&= \inf_{\lambda \geq 0} \left\{ \sum_{i=1}^{n} p_i \left[- \inf_{x_i \in \mathbb{R}} \lambda \frac{q_i}{p_i} x_i - u(x_i) \right] \right\} \\
&= \inf_{\lambda \geq 0} E_{\mathbf{p}} \left[\phi \left(\lambda \frac{\mathbf{q}}{\mathbf{p}} \right) \right].
\end{aligned}$$

By assumption, $\mathbf{x} = \mathbf{0}$ is a feasible solution to problem (4.6). Because the constraint $E_{\mathbf{q}}[\mathbf{x}] \leq 0$ is affine in $\mathbf{x}$, the existence of a feasible solution implies that strong duality holds. □

It follows that the scoring rule defined in Proposition 4.5 is tailored to optimization problem (4.25). As we now demonstrate, Scaled ϕ-Divergence scores, like ϕ-Divergence scores, are also closely connected to a particular convex risk measure, the u-Mean Certainty Equivalent.

Definition 4.4. *u-Mean Certainty Equivalent [35]*

Let u be a normalized utility function, and let X be a random variable. The u-Mean Certainty Equivalent, $M_u(X)$, satisfies the following equation:

$$M_u(X) := \sup \{ \eta \in \mathbb{R} \mid E[u(X - \eta)] \geq 0 \}.$$

We interpret $M_u(X)$ as the maximum amount a risk averse decision maker with normalized utility function u would be willing to pay to acquire an uncertain lottery X. In particular, the standard indifferent buying price for a lottery X in vNM utility theory is given by

$$\sup\left\{\eta \in \mathbb{R} \mid E[u(X + w_0 - \eta)] \geq u(w_0)\right\},$$

where w_0 is the purchaser's current wealth. The expression above is equivalent to $M_{u'}(X)$, where

$$u'(x) = \left(\frac{d}{dx}u(x)|_{w_0}\right)^{-1}(u(x + w_0) - u(w_0))$$

is the wealth-normalized version of u.

The following lemma, which mirrors Proposition 4.6, demonstrates that the Scaled ϕ-Divergence entropy measure, $\bar{H}_\phi(\mathbf{p}, \mathbf{q})$ is equivalent to the optimal reward available to a decision maker maximizing his u-Mean Certainty Equivalent.

Lemma 4.3. *Suppose that u is a normalized utility function, and set $\phi(t) = -u^*(t)$. Let*

$$\kappa(\mathbf{p}, \mathbf{q}) := \sup_{\mathbf{x}\in\mathbb{R}^n}\left\{M_u(\mathbf{x}) \mid E_{\mathbf{q}}[\mathbf{x}] = 0\right\},$$

with $M_u(\mathbf{x}) := \sup\left\{\eta \mid E_{\mathbf{p}}[u(\mathbf{x} - \eta\mathbf{1})] \geq 0\right\}$. Then for $\mathbf{q} << \mathbf{p}$,

$$\kappa(\mathbf{p}, \mathbf{q}) = \inf_{\lambda\geq 0} E_{\mathbf{p}}\left[\lambda\phi\left(\frac{\mathbf{q}}{\lambda\mathbf{p}}\right)\right] = \bar{H}_{\hat{\phi}}(\mathbf{q}, \mathbf{p}).$$

Proof. The definition of $M_u(\mathbf{x})$ gives

$$\kappa(\mathbf{p}, \mathbf{q}) = \sup_{\eta\in\mathbb{R}}\sup_{\mathbf{x}\in\mathbb{R}^n}\left\{\eta \mid E_{\mathbf{p}}[-u(\mathbf{x} - \eta\mathbf{1})] \leq 0, E_{\mathbf{q}}[\mathbf{x}] = 0\right\}.$$

Substituting $\mathbf{y} = \mathbf{x} - \eta\mathbf{1}$, we form the equivalent concave maximization problem

$$\kappa(\mathbf{p}, \mathbf{q}) = \sup_{\mathbf{y} \in \mathbb{R}^n} \left\{ -E_{\mathbf{q}}[\mathbf{y}] \mid E_{\mathbf{p}}[-u(\mathbf{y})] \leq 0 \right\}. \tag{4.26}$$

Forming the Lagrangian,

$$\begin{aligned}
\kappa(\mathbf{p}, \mathbf{q}) &\leq \inf_{\lambda \geq 0} \sup_{\mathbf{y} \in \mathbb{R}^n} \left\{ -E_{\mathbf{q}}[\mathbf{y}] + \lambda E_{\mathbf{p}}[u(\mathbf{y})] \right\} \\
&= \inf_{\lambda \geq 0} \left\{ - \inf_{\mathbf{y} \in \mathbb{R}^n} \left\{ E_{\mathbf{q}}[\mathbf{y}] - \lambda E_{\mathbf{p}}[u(\mathbf{y})] \right\} \right\} \\
&= \inf_{\lambda \geq 0} - \sum_{i=1}^{n} p_i \left[\inf_{y_i \in \mathbb{R}} \left\{ y_i \frac{q_i}{p_i} - \lambda u(y_i) \right\} \right] \\
&= \inf_{\lambda \geq 0} E_{\mathbf{p}} \left[-(\lambda u)^* \left(\frac{\mathbf{q}}{\mathbf{p}} \right) \right] \\
&= \inf_{\lambda \geq 0} E_{\mathbf{p}} \left[\lambda \phi \left(\frac{\mathbf{q}}{\lambda \mathbf{p}} \right) \right],
\end{aligned}$$

where the final equality follows from the identity $(\lambda u)^*(t) = \lambda u^* \left(\frac{t}{\lambda} \right)$.

We now establish that strong duality holds. By assumption, 0 is in the interior of the domain of u. If u takes positive values to the right of 0, we can find a Slater point $\mathbf{y}' \in \mathbb{R}^n$ of all (possibly small) positive values such that $\mathbf{y}'$ satisfies $E_{\mathbf{p}}[-u(\mathbf{y}')] < 0$.

The other possible case is where $u(x) = 0$ for $x \geq 0$. Observe that

$$\phi(t) = -u^*(t) = \sup_{x \in \mathrm{dom}(u)} \{u(x) - xt\} = 0$$

for $0 \leq t \leq 1$ ($xt \geq u(x)$ for $x < 0$, because we assume $1 \in \hat{\partial} u(0)$). Therefore, picking $\lambda > \max_i \left\{ \frac{q_i}{p_i} \right\}$ gives

$$E_{\mathbf{p}} \left[\lambda \phi \left(\frac{\mathbf{q}}{\lambda \mathbf{p}} \right) \right] = 0.$$

However, by inspection, the optimal value of problem (4.26) is 0 as well (with $\mathbf{y} = \mathbf{0}$). $\square$

Lemma 4.3 allows entropy measure (4.2), which underpins the Weighted Pseudospherical scoring rule, to be expressed in terms of the concave certainty equivalent M_{u_β}, where u_β is a HARA utility function. Recall that

$$\phi(t) = -u_\beta^*(t) = \frac{t^{1-\beta} - (1-\beta)t - \beta}{\beta(\beta-1)}.$$

Let $\gamma = 1 - \beta$. We have

$$\hat{\phi}(t) = \frac{t\left[t^{\beta-1} - (1-\beta)t^{-1} - \beta\right]}{\beta(\beta-1)} = \frac{t^{1-\gamma} - \gamma - (1-\gamma)t}{\gamma(\gamma-1)} = -u_\gamma^*(t).$$

It follows from Lemma 4.3 that

$$\sup_{\mathbf{x}\in\mathbb{R}^n} \left\{ M_{u_\beta}(\mathbf{x}) \mid E_{\mathbf{q}}(\mathbf{x}) = 0 \right\} = \bar{H}_{(-u_\gamma^*)}(\mathbf{q}, \mathbf{p}) = H_{1-\beta}^S(\mathbf{q}, \mathbf{p}). \tag{4.27}$$

Thus, like the Weighted Power scoring rule, the Weighted Pseudospherical scoring rule is similarly tailored to an optimization problem involving a concave certainty equivalent.

Substituting $\mathbf{y} = \mathbf{x} - \eta\mathbf{1}$ into the optimization problem on the left side of Equation (4.27), we find that the entropy measure for the Weighted Pseudospherical scoring rule is equal to the optimal value of the concave maximization problem

$$\sup_{\mathbf{y}\in\mathbb{R}^n} \left\{ -E_{\mathbf{q}}[\mathbf{y}] \mid E_{\mathbf{p}}[u_\beta(\mathbf{y})] \geq 0 \right\}. \tag{4.28}$$

We conclude that the Weighted Pseudospherical score divergence $H_{1-\beta}^S(\mathbf{q}, \mathbf{p})$ is interpretable as the maximum expected loss under the distribution $\mathbf{q}$, given that the risk constraint $E_{\mathbf{p}}[u_\beta(\mathbf{y})] \geq 0$ is satisfied. In this setting, the Weighted Pseudospherical scoring rule is tailored to risk-assessment problem (4.28) when forecasts are provided for the distribution $\mathbf{q}$. Higher worst-case expected losses under the subjective distribution $\mathbf{q}$, taken over random random variables that

have positive expected HARA utility under the baseline distribution $\mathbf{p}$, correspond to higher optimal scores.

This section concludes with the following illustration of the generality of our weighted scoring families. The following example derives the ϕ-Divergence and Scaled ϕ-Divergence weighted scoring rules associated with the linear-exponential utility function, which is not included in the HARA family.

Example 4.3. *Linear-Exponential Utility Scoring Rules*

The linear-exponential utility function has the following normalized functional form:

$$u_{linex}(x) = L\left[bx - e^{-x/c}\right] - L, \tag{4.29}$$

where $c, b > 0$ are parameters and $L = 1/\left(b + \frac{1}{c}\right)$. Linear-exponential utility incorporates decreasing levels of risk aversion with greater wealth, along with other desirable properties [8, 56].

When the linear-exponential utility is used in optimization problems (4.5) *and* (4.6)*, the corresponding ϕ-Divergence and Scaled ϕ-Divergence scoring rules can be found through derivation of the conjugate u^*_{linex}. The resulting reward function is given by*

$$S(\omega_i, \mathbf{r}) = -cbL \ln\left(\frac{c}{L}\frac{q_i}{r_i} - cb\right) - c\frac{q_i}{r_i} + cE_{\mathbf{q}}\left[\ln\left(\frac{c}{L}\frac{\mathbf{q}}{\mathbf{r}} - cb\right)\right] + L(cb - 1).$$

The score is well-defined only when $\frac{q_i}{r_i} > Lb$. When this inequality fails to hold, the underlying optimization problem (4.5) *is unbounded.*

Similarly, the Scaled ϕ-Divergence score associated with u_{linex} is given by

$$S(\omega_i, \mathbf{r}) = -cbL \ln\left(\frac{\lambda^* c}{L}\frac{q_i}{r_i} - bc\right) - \lambda^* c\frac{q_i}{r_i} + Lbc,$$

where λ^* satisfies

$$cE_{\mathbf{q}}\left[\ln\left(\frac{\lambda^* c}{L}\frac{q_i}{r_i} - bc\right)\right] = 0.$$

In this case, the score will be well-defined and finite. However, computing the score itself involves numerically solving for λ^*, and thus the score function (and associated Scaled ϕ-Divergence) cannot be presented to the forecaster in closed form. Methods for eliciting forecasts when the tailored score function cannot be represented analytically are given in [52].

4.5.4 Robust Weighted Power and Weighted Pseudospherical Scoring Rules

This section uses Proposition 4.1 to prove robust characterizations for the Weighted Power and Pseudospherical scoring rule. This involves generalizing Theorem 2 in [54], which is reproduced below and labeled Theorem 4.4.

Theorem 4.4. *[54]*

The following equalities hold:

1.
$$\inf_{\mathbf{z}\in\mathcal{P}} H_\beta^S(\mathbf{p}, \mathbf{z}^\top\mathbf{Q}) = \sup_{\mathbf{x}\in\mathbb{R}^n}\{E_{\mathbf{p}}[u_\beta(\mathbf{x})] \mid \mathbf{Q}\mathbf{x} \leq 0.\} \tag{4.30}$$

2.
$$\inf_{\mathbf{z}\in\mathcal{P}} H_\beta^P(\mathbf{p}, \mathbf{z}^\top\mathbf{Q}) = \sup_{\mathbf{x}\in\mathbb{R}^n, z\in\mathbb{R}}\{E_{\mathbf{p}}[u_\beta(\mathbf{x})] - z \mid \mathbf{Q}\mathbf{x} \leq z\}. \tag{4.31}$$

The economic interpretations for the optimization problems considered in Theorem 4.4 are similar to those for betting problems (4.5) and (4.6), except that in problems (4.30) and (4.31), the betting market is assumed to be incomplete, with the rows of the matrix $\mathbf{Q}$ representing risk neutral probability distributions supporting market prices. The constraint $\mathbf{Q}\mathbf{x} \leq 0$ (and similarly

$\mathbf{Qx} \leq z$) imply that $E_{\mathbf{q}}[\mathbf{x}] \leq 0$ for every distribution $\mathbf{q}$ in the convex hull of these extreme distributions (i.e., the rows of $\mathbf{Q}$). As pointed out in [54], the equality established in Equation (4.30) follows from a more general duality relationship studied in [41].

Here, we generalize Theorem 4.4 by leveraging the connection between weighted scoring rules and convex risk measures. Let $\mathcal{Q} \subset \mathcal{P}$ be a convex set of probability distributions. We consider the following robust version of the *Bounded Shortfall Risk* [35]:

$$\rho(\mathbf{x}) := \inf_{\eta \in \mathbb{R}} \{\eta \mid \sup_{\mathbf{q} \in \mathcal{Q}} E_{\mathbf{q}}[\ell(-\eta \mathbf{1} - \mathbf{x})] \leq x_0\},$$

with $x_0 \in \mathbb{R}$ and ℓ a convex loss function. The optimal value for η represents the smallest constant function that must be added to the random variable $\mathbf{x}$ to ensure that the expected loss, with respect to the most adversarial distribution in the set $\mathcal{Q}$, is bounded by x_0. Choosing $\ell_\beta(t) := -u_{1-\beta}(-t)$ as the loss function, the associated concave certainty equivalent takes the form:

$$-\rho(\mathbf{x}) = \sup \left\{ \eta \in \mathbb{R} \mid \inf_{\mathbf{q} \in \mathcal{Q}} E_{\mathbf{q}}[u_{1-\beta}(\mathbf{x} - \eta \mathbf{1})] \geq x_0 \right\}, \tag{4.32}$$

which is a robust version of the u-Mean Certainty Equivalent given in Definition 4.4.

The penalty function $H(\mathbf{a})$ for $\mathbf{a} \in \mathcal{P}$ associated with the robust bounded shortfall risk is shown in [35, Proposition 14] to have the form:

$$H(\mathbf{a}) = \inf_{\lambda > 0} \ \lambda \left[x_0 + \inf_{\mathbf{q} \in \mathcal{Q}} E_{\mathbf{q}} \left[\ell^* \left(\frac{\mathbf{a}}{\lambda \mathbf{q}} \right) \right] \right].$$

The aforementioned choice of $\ell_\beta(t) := -u_{1-\beta}(-t)$ implies $\ell^*_\beta(t) = -u^*_{1-\beta}(t) = -\hat{u}^*_\beta(t)$. Thus, we can write

$$H(\mathbf{a}) = \inf_{\mathbf{q} \in \mathcal{Q}} \inf_{\lambda > 0} \ \lambda x_0 + E_{\mathbf{a}} \left[\phi \left(\lambda \frac{\mathbf{q}}{\mathbf{a}} \right) \right], \ \phi(t) := -u^*_\beta(t).$$

The inner optimization problem is the same as problem (4.22) except for the term λx_0. Solving for the optimal λ yields

$$H(\mathbf{a}) = \inf_{\mathbf{q}\in\mathcal{Q}} \left[\frac{E_{\mathbf{q}}\left[\left(\frac{\mathbf{a}}{\mathbf{q}}\right)^{\beta}\right]}{\beta x_0 + 1} \right]^{(1/\beta)} \left[\frac{\beta x_0 + 1}{\beta - 1}\right] - \frac{1}{\beta - 1}.$$

It now follows from Proposition 4.1, with the choice of $-\rho$ as in Equation (4.32), that

$$H(\mathbf{p}) = \sup_{\mathbf{x}\in\mathbb{R}^n} \sup_{\eta\in\mathbb{R}} \{\eta \mid \inf_{\mathbf{q}\in\mathcal{Q}} E_{\mathbf{q}}[u_{1-\beta}(\mathbf{x} - \eta\mathbf{1})] \geq x_0, E_{\mathbf{p}}[\mathbf{x}] \leq 0\}. \tag{4.33}$$

Substituting $\mathbf{y} = \mathbf{x} - \eta\mathbf{1}$ on the right side, the optimization problem given in (4.33) is equivalent to:

$$\sup_{\mathbf{y}\in\mathbb{R}^n} \{E_{\mathbf{p}}[-\mathbf{y}] \mid \inf_{\mathbf{q}\in\mathcal{Q}} E_{\mathbf{q}}[u_{1-\beta}(\mathbf{y})] \geq x_0\}. \tag{4.34}$$

A property of HARA utility functions is that $u_{1-\beta}(-u_\beta(-y)) = y$. This allows substituting $\mathbf{y} = -u_\beta(\mathbf{z})$ in problem (4.34) and reformulating as follows:

$$\sup_{\mathbf{z}\in\mathbb{R}^n} \{E_{\mathbf{p}}[u_\beta(\mathbf{z})] \mid \inf_{\mathbf{q}\in\mathcal{Q}} -E_{\mathbf{q}}[\mathbf{z}] \geq x_0\} = \sup_{\mathbf{z}\in\mathbb{R}^n} \{E_{\mathbf{p}}[u_\beta(\mathbf{z})] \mid \sup_{\mathbf{q}\in\mathcal{Q}} E_{\mathbf{q}}[\mathbf{z}] \leq x_0\}. \tag{4.35}$$

Equivalence between Equations (4.34) and (4.35) follows from there being a one-to-one correspondence between feasible solutions given by the relationship $y_i = -u_\beta(z_i)$, where $z_i = -u_{1-\beta}(y_i)$, for each y_i in the domain of $u_{1-\beta}$. Combining these observations yields

$$\begin{aligned} &\inf_{\mathbf{q}\in\mathcal{Q}} \left[\frac{E_{\mathbf{q}}\left[\left(\frac{\mathbf{p}}{\mathbf{q}}\right)^{\beta}\right]}{\beta x_0 + 1} \right]^{(1/\beta)} \left[\frac{\beta x_0 + 1}{\beta - 1}\right] - \frac{1}{\beta - 1} \\ &= \sup_{\mathbf{z}\in\mathbb{R}^n} \{E_{\mathbf{p}}[u_\beta(\mathbf{z})] \mid \sup_{\mathbf{q}\in\mathcal{Q}} E_{\mathbf{q}}[\mathbf{z}] \leq x_0\}. \end{aligned} \tag{4.36}$$

We are now ready to show the following generalization of Theorem 4.4.

Proposition 4.9. *Let $\mathcal{Q}$ be a convex subset of $\mathcal{P}$. Then,*

1.

$$\inf_{\mathbf{q}\in\mathcal{Q}} H_\beta^S(\mathbf{p},\mathbf{q}) = \sup_{\mathbf{x}\in\mathbb{R}^n}\left\{E_{\mathbf{p}}[u_\beta(\mathbf{x})] \mid \sup_{\mathbf{q}\in\mathcal{Q}} E_{\mathbf{q}}[\mathbf{x}] \le 0\right\}. \tag{4.37}$$

2.

$$\inf_{\mathbf{q}\in\mathcal{Q}} H_\beta^P(\mathbf{p},\mathbf{q}) = \sup_{\mathbf{x}\in\mathbb{R}^n, z\in\mathbb{R}}\left\{E_{\mathbf{p}}[u_\beta(\mathbf{x})] - z \mid \sup_{\mathbf{q}\in\mathcal{Q}} E_{\mathbf{q}}[\mathbf{x}] \le z\right\}. \tag{4.38}$$

Proof. 1. Follows immediately by setting $x_0 = 0$ in Equation (4.36).

2. First,

$$\begin{aligned}\sup_{\mathbf{x}\in\mathbb{R}^n, z\in\mathbb{R}}\left\{E_{\mathbf{p}}[u_\beta(\mathbf{x})] - z \mid \sup_{\mathbf{q}\in\mathcal{Q}} E_{\mathbf{q}}[\mathbf{x}] \le z\right\} &= \sup_{\mathbf{x}\in\mathbb{R}^n}\inf_{\mathbf{q}\in\mathcal{Q}}\{E_{\mathbf{p}}[u_\beta(\mathbf{x})] - E_{\mathbf{q}}[\mathbf{x}]\} \\ &\le \inf_{\mathbf{q}\in\mathcal{Q}}\sup_{\mathbf{x}\in\mathbb{R}^n}\{E_{\mathbf{p}}[u_\beta(\mathbf{x})] - E_{\mathbf{q}}[\mathbf{x}]\} \\ &= \inf_{\mathbf{q}\in\mathcal{Q}} H_\beta^P(\mathbf{p},\mathbf{q}),\end{aligned}$$

where the final equality is (4.5).

To obtain the reverse inequality, use Equation (4.36) to obtain

$$\begin{aligned}&\sup_{\mathbf{x}\in\mathbb{R}^n, z\in\mathbb{R}}\left\{E_{\mathbf{p}}[u_\beta(\mathbf{x})] - z \mid \sup_{\mathbf{q}\in\mathcal{Q}} E_{\mathbf{q}}[\mathbf{x}] \le z\right\} \\ &= \sup_{z\in\mathbb{R}}\left\{\inf_{\mathbf{q}\in\mathcal{Q}} \frac{\left[E_{\mathbf{q}}\left[\left(\frac{\mathbf{p}}{\mathbf{q}}\right)^\beta\right]\right]^{(1/\beta)}(\beta z+1)^{(\beta-1)/\beta}}{\beta-1} - \frac{1}{\beta-1} - z\right\} \\ &\ge \sup_{z:\ \beta z+1>0}\left\{\inf_{\mathbf{q}\in\mathcal{Q}} \frac{\left[E_{\mathbf{q}}\left[\left(\frac{\mathbf{p}}{\mathbf{q}}\right)^\beta\right]\right]^{(1/\beta)}(\beta z+1)^{(\beta-1)/\beta}}{\beta-1} - \frac{1}{\beta-1} - z\right\}.\end{aligned}$$

The inner optimization problem in $\mathbf{q}$ does not depend on z. In fact, an optimal solution $\mathbf{q}^*$ minimizes the Pseudospherical entropy measure:

$$H_\beta^S(\mathbf{p}, \mathbf{q}^*) = \inf_{\mathbf{q} \in \mathcal{Q}} H_\beta^S(\mathbf{p}, \mathbf{q})$$

for all values $z > -1/\beta$. Thus, we obtain

$$\sup_{\mathbf{x} \in \mathbb{R}^n, z \in \mathbb{R}} \left\{ E_{\mathbf{p}}[u_\beta(\mathbf{x})] - z \mid \sup_{\mathbf{q} \in \mathcal{Q}} E_{\mathbf{q}}[\mathbf{x}] \leq z \right\}$$

$$\geq \sup_{z:\ \beta z + 1 > 0} \left\{ \frac{c}{\beta - 1} (\beta z + 1)^{(\beta - 1)/\beta} - \frac{1}{\beta - 1} - z \right\},$$

where

$$c = \left[E_{\mathbf{q}^*} \left[\left(\frac{\mathbf{p}}{\mathbf{q}^*} \right)^\beta \right] \right]^{(1/\beta)}.$$

By direct computation, the optimal value of z is $z^* = \frac{c^\beta - 1}{\beta} > -1/\beta$. Moreover, we find that

$$\frac{c}{\beta - 1} (\beta z^* + 1)^{(\beta - 1)/\beta} - \frac{1}{\beta - 1} - z^* = H_\beta^P(\mathbf{p}, \mathbf{q}^*) \geq \inf_{\mathbf{q} \in \mathcal{Q}} H_\beta^P(\mathbf{p}, \mathbf{q}).$$

Thus, in total, we have shown that

$$\sup_{\mathbf{x} \in \mathbb{R}^n, z \in \mathbb{R}} \left\{ E_{\mathbf{p}}[u_\beta(\mathbf{x})] - z \mid \sup_{\mathbf{q} \in \mathcal{Q}} E_{\mathbf{q}}[\mathbf{x}] \leq z \right\} \geq \inf_{\mathbf{q} \in \mathcal{Q}} H_\beta^P(\mathbf{p}, \mathbf{q}).$$

□

This result extends [54, Theorem 2] (denoted Theorem 4.4 above) to arbitrary convex sets. As we have shown, these robust characterizations of the Weighted Power and Weighted Pseudospherical scoring rules follow directly from their relationships with convex risk measures.

4.5.5 Tailored scores for distributionally robust utility maximization

In the standard utility maximization problem (4.25), a forecast for the distribution $\mathbf{p}$ over future state values of the traded asset $\mathbf{y}$ is likely to be imprecise. In this case, the investor receiving the forecast might wish to hedge against distributions "close" to the forecasted distribution. This section derives the tailored score for a distributionally robust utility maximization problem of the form

$$\sup_{\mathbf{x}\in\mathbb{R}^n} \{f(\mathbf{p},\mathbf{x},t) \mid E_{\mathbf{p}}[\mathbf{x}] \leq 0\}, \tag{4.39}$$

where

$$f(\mathbf{p},\mathbf{x},t) = \inf_{\mathbf{r}\in\mathcal{P}:H_\phi(\mathbf{r},\mathbf{p})\leq t} E_{\mathbf{r}}[u(\mathbf{x})]. \tag{4.40}$$

Here, closeness to the forecast $\mathbf{p}$ is in terms of a ϕ-Divergence. When $u(\mathbf{x}) = e^{-x}$, the exponential utility function, $f(\mathbf{p},\mathbf{x},t)$ is a concave certainty equivalent for $\mathbf{x}$, so that the complete problem (4.39) is a special case of the optimization model presented in Proposition 4.1.

A tailored score for the forecast $\mathbf{p}$ in problem (4.39) can be found by considering the equivalent dual formulation, given in [9, corrolary 6.1]:

$$\sup_{\mathbf{x}:E_{\mathbf{q}}[\mathbf{x}]\leq 0,\lambda\geq 0,\eta\in\mathbb{R}} \left\{ E_{\mathbf{p}}\left[-\eta\mathbf{e} - t\lambda\mathbf{e} - \lambda\phi^*\left(\frac{u(\mathbf{x})-\eta}{\lambda}\right)\right]\right\}. \tag{4.41}$$

When u is concave, the optimization problem above is also. The tailored scoring rule for $\mathbf{p}$ can then be obtained in manner described in [52]:

$$S(\omega_i,\mathbf{r}) = -\eta' - t\lambda' - \lambda'\phi^*\left(\frac{u(x_i')-\eta'}{\lambda}\right),$$

where $\lambda',\eta',\mathbf{x}'$ are optimal solutions in the dual formulation (4.41) with $\mathbf{p}$ replaced by $\mathbf{r}$.

4.6 Conclusion

This chapter has described ϕ-Divergence and Scaled ϕ-Divergence weighted scoring rules that generalize the Weighted Power scoring rules and Weighted Pseudospherical scoring rules described in [54]. As we have demonstrated, the ϕ-Divergence and Scaled ϕ-Divergence rules have entropy measures equivalent to the optimal expected rewards in a betting problem involving a convex risk measure. We further showed a more general relationship: all weighted scoring rules are tailored to a corresponding decision problem involving a risk averse bettor seeking to maximize a concave certainty equivalent (the negative of a convex risk measure).

This result provides an economic interpretation for weighted scoring rules as providing the comparative value of probabilistic information in a simple betting market. Although all weighted scoring rules are tailored to a decision problem with this underlying structure, the specific choice of a scoring rule corresponds to the risk preference of the decision maker. For ϕ-Divergence and Scaled ϕ-Divergence rules, the score function depends upon both the form of the decision maker's utility function and the choice of risk measure.

The insights in this chapter provide a new connection between proper scoring rules and financial mathematics. Due to the large number of convex risk measures described in the literature, leveraging this relationship could lead to the identification of useful and interesting tailored scoring rules.

4.7 Appendix for Chapter 4

4.7.1 Additional Proofs

Proof. Proof of Lemma 2

By assumption $u(x) \leq x$ for all $x \in \text{dom}(u)$ and $u(0) = 0$. This gives

$$-u^*(1) = \sup_{x \in \text{dom}(u)} \{u(x) - x\} = 0.$$

Also,

$$\inf_t \sup_{x \in \text{dom}(u)} \{u(x) - tx\} \geq \inf_t \{u(0) - t \cdot 0\} = 0.$$

□

We now present a technical lemma that aids in proving subsequent results.

Lemma 4.4. *Let $\phi(t) = -u^*(t)$, where u is a normalized utility function. Then $x \in \partial\phi(t)$ if, and only if, $t \in \hat{\partial}u(-x)$.*

Proof. Recall that because $\phi(t)$ is convex and closed [67], we have

$$x \in \partial\phi(t) \iff t \in \partial\phi^*(x). \tag{4.42}$$

By definition $\phi^*(x) = (-u)^{**}(x) = -u(-x)$. Taking subgradients w.r.t to x on the right side yields $\partial\phi^*(x) = \hat{\partial}u(-x)$. Substituting this relationship into (4.42) completes the proof. □

Proof. Proof of Proposition 4.7

The bet $\mathbf{x}^*$ is optimal if, and only if, $\mathbf{0} \in \partial\{E_{\mathbf{p}}[u(\mathbf{x}^*)] - E_{\mathbf{q}}[\mathbf{x}]\}$, which reduces to the condition $\frac{q_i}{p_i} \in \hat{\partial}u(x_i^*)$ for all i. This optimality condition for x_i^* also implies that $u^*\left(\frac{q_i}{p_i}\right) = \inf_{x \in \text{dom}(u)} \frac{q_i}{p_i}x - u(x) = \frac{q_i}{p_i}x_i^* - u(x_i^*)$. Using this identity,

$$u_i(\mathbf{x}^*) = u\left(x_i^*\right) - E_{\mathbf{q}}[\mathbf{x}^*] = \frac{q_i}{p_i}x_i^* - u^*\left(\frac{q_i}{p_i}\right) - E_{\mathbf{q}}[\mathbf{x}^*].$$

Set $\phi(t) = -u^*(t)$ for $t \geq 0$ so that $\hat{\phi}(t) = -tu^*\left(\frac{1}{t}\right)$. We will show that $\frac{q_i}{p_i}x_i^* - u^*\left(\frac{q_i}{p_i}\right) \in \partial\hat{\phi}\left(\frac{p_i}{q_i}\right)$, which by Proposition 4.2 will demonstrate that $u_i(\mathbf{x}^*) = S_{\hat{\phi}}(\omega_i, \mathbf{p}, \mathbf{q})$. From Lemma 4.4, $\frac{q_i}{p_i} \in \hat{\partial}u(x_i^*)$ if, and only if, $-x_i^* \in \partial\{-u^*\left(\frac{q_i}{p_i}\right)\}$. Therefore, for all $z > 0$,

$$\begin{aligned} &\hat{\phi}\left(\frac{p_i}{q_i}\right) + \left(\frac{q_i}{p_i}x_i^* - u^*\left(\frac{q_i}{p_i}\right)\right)\left(z - \frac{p_i}{q_i}\right) \qquad (4.43) \\ &= z\left(-u^*\left(\frac{q_i}{p_i}\right) + \frac{q_i}{p_i}x_i^* - \frac{1}{z}x_i^*\right) \leq -zu^*\left(\frac{1}{z}\right) = \hat{\phi}(z). \end{aligned}$$

This establishes the subgradient inequality for $z > 0$.

For $z = 0$, using the fact that $\hat{\phi}\left(\frac{p_i}{q_i}\right) = -\frac{p_i}{q_i}u^*\left(\frac{q_i}{p_i}\right)$, the expression in Equation (4.43) reduces to $-x_i^*$. Observe that

$$\hat{\phi}(0) = \lim_{t\to 0} t\phi\left(\frac{1}{t}\right) = \lim_{s\to\infty}\frac{\phi(s)}{s}.$$

It was seen in Section 4.5 (in the lead-in to Proposition 4.3) that $-x^* \leq \lim_{s\to\infty}\frac{\phi(s)}{s}$. Having established

$$\hat{\phi}\left(\frac{p_i}{q_i}\right) + \left(\frac{q_i}{p_i}x_i^* - u^*\left(\frac{q_i}{p_i}\right)\right)\left(z - \frac{p_i}{q_i}\right) \leq \hat{\phi}(z)$$

for all $z \geq 0$, we conclude that $\frac{q_i}{p_i}x_i^* - u^*\left(\frac{q_i}{p_i}\right) \in \partial\hat{\phi}\left(\frac{p_i}{q_i}\right)$. □

4.7.2 Extension of Proposition 4.6

This appendix proves a more general version of Proposition 4.6 given in Section 4.5.2. We begin with the following Lemma on the limiting behavior of $\phi(t) = -u^*(t) = \sup_{x\in\text{dom}(u)}\{u(x) - tx\}$ for a normalized utility function u. Denote by $u'_-(x)$ and $u'_+(x)$ the right and left derivatives at a point $x \in \text{dom}(u)$. The superdifferential of u at x, $\hat{\partial}u(x)$, is exactly the set $\{v \in \mathbb{R} : u'_+(x) \leq v \leq u'_-(x)\}$. Standard results guarantee that $\hat{\partial}u(x)$ is non-empty for $x \in \text{int}(\text{dom}(u))$.

Lemma 4.5. *Let u be a normalized utility function, and set $\phi(t) = -u^*(t)$. Set $a := \inf\{x : x \in dom(u)\}$. Then $\lim_{t\to\infty} \phi(t)/t = -a$.*

Proof. We have that

$$\phi(t) = -u^*(t) = \sup_{x\in\mathrm{dom}(u)} \{u(x) - tx\} \geq u(y) - ty, \ \forall y \in \mathrm{dom}(u).$$

Dividing both sides by $t > 0$ and taking limits yields

$$\lim_{t\to\infty} \frac{\phi(t)}{t} \geq \lim_{t\to\infty} \frac{u(y)}{t} - y = -y, \ \forall y \in \mathrm{dom}(u).$$

Taking the supremum with respect to y on the right side gives the inequality $\lim_{t\to\infty} \phi(t)/t \geq -a$.

When $a = -\infty$, this inequality implies that $\lim_{t\to\infty} \phi(t)/t = \infty$. When $a > -\infty$ there are two cases. One is that u is defined on the left-closed domain $[a, b)$. The other is that u is defined on the left-open domain (a, b), where, because $-u$ is closed, we have $\lim_{x\searrow a} u(x) = -\infty$. In either case, for sufficiently large t, the optimization problem $\sup_{x\in\mathrm{dom}(u)} \{u(x) - tx\}$ has a finite optimal solution $x^* \in \mathrm{dom}(u)$.

Let $t_n \nearrow \infty$ be an increasing sequence and let

$$x_n^* \in \operatorname*{arg\,max}_{x\in\mathrm{dom}(u)} \{u(x) - t_n x\}.$$

The optimality condition for x_n^* is $t_n \in \hat{\partial} u(x_n^*)$, which holds if, and only if, $-x_n^* \in \partial\{-u^*(t_n)\} = \partial\phi(t_n)$ (see Lemma 4.4). By the convexity of ϕ, this implies that $\phi'_-(t_n) \leq -x_n^* \leq -a$. We conclude that $\lim_{t\to\infty} \phi'_-(t) \leq -a$.

Fixing $s \in \mathrm{dom}(\phi)$ the subgradient inequality yields

$$\phi(s) \geq \phi(t_n) + \phi'_-(t_n)(s - t_n) \Longleftrightarrow \phi(s) + \phi'_-(t_n)(t_n - s) \geq \phi(t_n).$$

Dividing both sides by t_n and letting $n \to \infty$ gives $\lim_{t\to\infty} \phi(t)/t \leq \lim_{t\to\infty} \phi'_-(t) \leq -a$. We have already shown that $\lim_{t\to\infty} \phi(t)/t \geq -a$, and thus the proof is complete.

□

We now prove a more general version of Proposition 4.6 that does not require $\mathbf{q} << \mathbf{p}$.

Proposition 4.10. *Let $\mathbf{p}, \mathbf{q} \in \mathcal{P}$, let u be a normalized utility function, and set $\phi(t) = -u^*(t)$. Then*

$$H_{\hat{\phi}}(\mathbf{p}, \mathbf{q}) := E_{\mathbf{q}}\left[\hat{\phi}\left(\frac{\mathbf{p}}{\mathbf{q}}\right)\right]$$

is a ϕ-divergence and $H_{\hat{\phi}}(\mathbf{p}, \mathbf{q}) = z(\mathbf{p}, \mathbf{q})$, where

$$z(\mathbf{p}, \mathbf{q}) = \sup_{\mathbf{x}\in\mathbb{R}^n:\ x_i \in dom(u),\ i=1,\ldots,n} E_{\mathbf{p}}[u(\mathbf{x})] - E_{\mathbf{q}}[\mathbf{x}]. \tag{4.44}$$

Proof. Set $\phi(t) = -u^*(t)$ and denote by $\hat{\phi}$ the adjoint of ϕ. We have observed that $\phi(\cdot)$, as defined above, is a proper, closed, convex function. As shown in Lemma 4.2, we have that $\phi(1) = 0$ indicating the quantity $H_\phi(\mathbf{q}, \mathbf{p})$ is indeed the ϕ-divergence of $\mathbf{q}$ from $\mathbf{p}$.

Now, consider optimization problem (4.44). Without loss of generality, assume that $p_i = 0, q_i > 0$ for $k \leq i \leq n$, with $1 \leq k \leq n$. Setting $a :=$

$\inf\{x|x \in \mathrm{dom}(u)\}$, we have

$$\begin{aligned}
z(\mathbf{p},\mathbf{q}) &= \sup_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u(\mathbf{x})] - E_{\mathbf{q}}[\mathbf{x}] \\
&= \sup_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{k-1} \{p_i u(x_i) - q_i x_i\} - \sum_{i=k}^{n} q_i a \\
&= -\inf_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{k-1} p_i \left(\frac{q_i}{p_i} x_i - u(x_i) \right) - \sum_{i=k}^{n} q_i a \\
&= -\sum_{i=1}^{k-1} p_i \left(\inf_{x_i\in\mathbb{R}} x_i \frac{q_i}{p_i} - u(x_i) \right) - \sum_{i=k}^{n} q_i a \\
&= \sum_{i=1}^{k-1} p_i \phi\left(\frac{q_i}{p_i}\right) - \sum_{i=k}^{n} q_i a.
\end{aligned}$$

However, by Lemma 3, we have

$$p_i \phi\left(\frac{q_i}{p_i}\right) = q_i \lim_{t\to\infty} \frac{\phi(t)}{t} = -q_i a$$

for $k \leq i \leq n$ (i.e. when $p_i = 0$) and the result follows. □

Chapter 5

Linear Scoring Rules for Decision Analysis Applications

5.1 Chapter Summary

This chapter describes a class of proper scoring rules that incentivize honest reporting of certain linear properties of the forecaster's underlying distribution. The linear properties of a distribution are those that can be expressed as an expectation, that is, a vector $\mathbf{f} = E_{\mathbf{p}}[g(\omega)]$. A strictly proper scoring rule for such linear properties is a function S that satisfies $E_{\mathbf{p}}[S(\omega, \mathbf{f})] \geq E_{\mathbf{p}}[S(\omega, \mathbf{g})]$ for arbitrary report $\mathbf{g}$ and $\mathbf{f} = E_{\mathbf{p}}[g(\omega)]$. Strictly proper scores for linear properties have been characterized in the measure-theoretic setting in [2] and extended in [38]. This previous work focused on scoring the output of a machine learning model for fitting purposes.

Here, we provide guidance on scoring linear forecasts in decision analysis applications, in which a human forecaster is asked to provide the answers to a series of queries concerning his subjective joint distribution. This forecast could then be used in a model by the decision analyst. We consider the special case in which the forecaster's underlying distribution is discrete, and the forecaster is asked to provide an arbitrary sequence of probability assessments pertaining to the underlying distribution. The provided forecast can be represented as $\mathbf{f}$ equal to $\mathbf{Xp}$, where $\mathbf{X}$ is a binary matrix and $\mathbf{p}$ is a discrete distribution. We begin by providing a construction method for strictly

proper scoring rules that incentivize honest reporting in this situation. We then demonstrate that the resulting rule is proper for distributional forecasts, and we characterize those matrices for which the rule is strictly proper. We further demonstrate that many existing distributional scoring rules can be represented in this form. Finally, we show how such rules arise naturally in the context of designing a tailored scoring rule for a generic distributionally robust optimization problem.

5.2 Introduction

As in previous chapters, we consider an uncertain experiment with sample space Ω consisting of $n > 1$ mutually exclusive possible outcomes. A forecaster's private subjective beliefs for the probability of each possible outcome is captured in a *distributional forecast* $\mathbf{p} = [p_{\omega_1}, p_{\omega_2}, \ldots, p_{\omega_n}]$, where p_ω is the probability of observing ω. Unlike in the previous chapters, we assume that the forecaster is asked to provide a forecast $\mathbf{f}$ that is representable as an affine function of his subjective probabilistic information. We assume that the forecasted quantity $\mathbf{f} = \mathbf{X}\mathbf{p}$, for some matrix $\mathbf{X}$ with binary elements. The forecast $\mathbf{f}$ could include

- Forecasts for cumulative probabilities.
- Forecasts for marginal or conditional distributions.
- Probabilistic estimates for events of interest.

We will assume that the goal of the forecasting exercise is to acquire expert information regarding an underlying joint distribution, as is typical in

decision analysis applications. In such a case, the elicitation procedure may be decomposed into simple forecasting tasks. Further, it might be that only certain information is well understood by the forecaster. For example, the marginal distribution for each uncertainty may be simple to forecast, but the full dependence structure between uncertainties may be more difficult to ascertain.

As an example, suppose Y^1, Y^2 are random variables with associated joint random variable $Y = (Y^1, Y^2)$ taking values in the set $\{(5,2),(5,0),(0,2),(0,0)\}$. We assume that forecasts are made according to the joint probabilities given in Figure 5.1. The following assessments can be formulated in terms of a linear function of the joint probabilities:

- The question "What is the marginal probability $P(Y^1 = 5)$?" corresponds to the forecast $[1,1,0,0]^\top[.08,.32.,48,.12] = .4$.
- The question "What is the probability that $Y^1 = 5$ or $Y^2 = 2$?" corresponds to the forecast $[1,1,1,0]^\top[.08,.32.,48,.12] = .88$.

These queries and the resulting forecast can be written in matrix form:

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix} (\mathbf{X}) \begin{bmatrix} .08 \\ .32 \\ .48 \\ .12 \end{bmatrix} (\mathbf{p}) = \begin{bmatrix} .4 \\ .88 \end{bmatrix} (\mathbf{f}).$$

This chapter gives a general method for constructing strictly proper scoring rules for linear forecasts like the one above. Commonly used scoring rules for distributions, such as the Logarithmic, Quadratic, and Spherical rules presented in Table 1.1, can be easily adapted to score the forecasts for linear

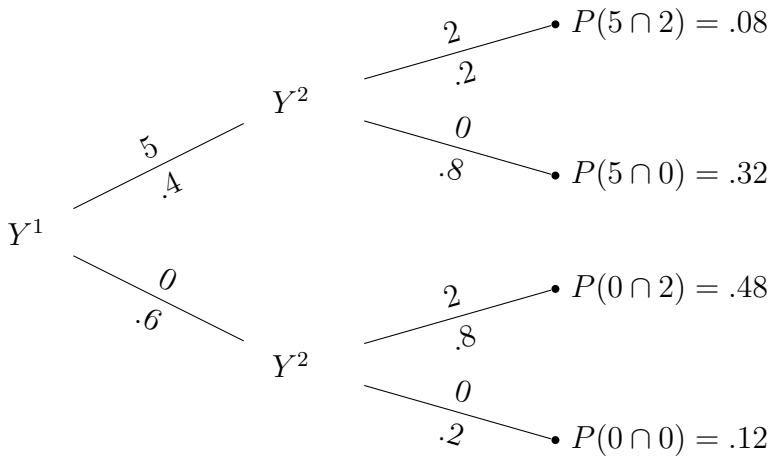


Figure 5.1: Probabilities for Y^1, Y^2

properties. Like distributional scoring rules, the scores for linear forecasts reward the accuracy of the prediction $\mathbf{f}$ while incentivizing honesty. In particular, if $\mathbf{f} = \mathbf{Xp}$ and $\mathbf{g} \neq \mathbf{f}$ are two linear forecasts, a proper scoring rule for linear properties $S^{lin}(\cdot)$ satisfies $E_{\mathbf{p}}[S^{lin}(\omega, \mathbf{f})] > E_{\mathbf{p}}[S^{lin}(\omega, \mathbf{g})]$. Under these rules, if a forecaster believes $\mathbf{p}$ to be the true distribution, he maximizes his expected score by reporting the linear forecast that corresponds with his subjective beliefs. As we demonstrate, there is no mathematical distinction between scoring rules for distributional forecasts and scoring rules for linear properties. As a result, we can construct scoring rules for distributions using the same principles.

The remainder of the chapter is organized as follows. Section 5.2.1 introduces the chapter-specific mathematical background and notation. Section 5.3 introduces a simple method for constructing scores for linear forecasts that directly follows from the primary characterization results for distributional scoring presented in Theorem 2.1. We also provide examples of these scor-

ing rules. Section 5.4 provides application areas for scoring rules described in Section 5.3.

5.2.1 Notation and Background

This chapter employs the following notation. We replace $\mathcal{P}$ (and similarly $\mathcal{A}$) with $\mathcal{P}^n$, where n reflects the fact that $\mathcal{P}^n \subset \mathbb{R}^n$. The forecaster is assumed to believe that the true underlying distribution over Ω is given by $\mathbf{p}$ and that his or her linear forecast comes from the convex *forecast set*:

$$\mathcal{F}(\mathbf{X}) := \{\mathbf{f} \in \mathbb{R}^\ell : \mathbf{f} = \mathbf{X}\mathbf{r}, \mathbf{r} \in \mathcal{A}\}, \tag{5.1}$$

where $\mathbf{X} \in \mathbb{R}^{\ell \times n}, \ell > 2$ is a *query matrix* made up of 0s and 1s, and is not the $\mathbf{0}$ matrix. The first row of the matrix $\mathbf{X}$ is assumed equal to $\mathbf{1}^\top$. Given $\mathbf{X}$ and $\mathbf{r}$, the forecast $\mathbf{f} = \mathbf{X}\mathbf{r}$ can be similarly interpreted as the expectation of a random vector under $\mathbf{r}$, where the ith outcome corresponds to the ith column of $\mathbf{X}$. Thus, forecast $\mathbf{f} = \mathbf{X}\mathbf{r}$ is a linear property of the distribution $\mathbf{r}$, as defined in [2].

Conversely, given a linear forecast $\mathbf{f}^* \in \mathcal{F}(\mathbf{X})$ and matrix $\mathbf{X}$, we can define the *truth set*

$$\mathcal{T}(\mathbf{f}^*) := \{\mathbf{r} \in \mathcal{A}^n : \mathbf{X}\mathbf{r} = \mathbf{f}^*\},$$

which is the closed and convex polytope containing all distributions consistent with $\mathbf{f}^*$. Because the truth set is convex for $\mathbf{f}$, the forecasts in $\mathcal{F}(\mathbf{X})$ are elicitable, meaning that there exists a scoring rule for these properties that incentivizes honest reporting [57]. In fact, many such rules will satisfy this criterion.

We view forecast $\mathbf{f} = \mathbf{X}\mathbf{p}$ as representing the answers provided in a series of forecasting tasks, where each row $\mathbf{X}(i)$ of $\mathbf{X}$ represents an individual query

relating to the underlying distribution. Typically, a row of $\mathbf{X}$ will take the form $\chi(A)$ for some set A. For example, in the numerical example given in the previous section, we have

$$\mathbf{X} = \begin{bmatrix} \chi(\{Y^1 = 5\}) \\ \chi(\{Y^1 = 5 \ \cup \ Y^2 = 2\}) \end{bmatrix}.$$

5.3 Scoring Linear Forecasts

We now directly extend the construction method for distributional scoring rules described in Theorem 2.1 to linear forecasts $\mathbf{f} \in \mathcal{F}(\mathbf{X})$, where $\mathbf{X} \in \mathbb{R}^{\ell \times n}$ is a query matrix. We define linear scoring rules as follows.

Definition 5.1. *Linear Scoring Rule*

Let $H : \mathcal{A}^\ell \to \mathbb{R}$ be an entropy function on $\mathcal{A}^\ell$. Define the linear scoring rule $S_H^{lin} : \Omega \times \mathcal{F}(\mathbf{X}) \to \mathbb{R}$ as $S_H^{lin}(\omega, \mathbf{f}) := (\mathbf{X}\boldsymbol{\delta}_\omega)^\top \mathbf{v}(\mathbf{f})$, where $\mathbf{v}(\mathbf{f}) \in \partial H(\mathbf{f})$ for each $\mathbf{f} \in \mathcal{F}(\mathbf{X})$.

Proposition 5.1 will demonstrate that S_H^{lin} can be viewed as a proper distributional scoring rule, and that S_H^{lin} is strictly proper over linear forecasts. This proposition will use the following lemma.

Lemma 5.1. *Let H be an entropy measure on $\mathcal{A}^n$, strictly convex on $\mathcal{P}^n$. If $\mathbf{v}(\mathbf{p}) \in \partial H(\mathbf{p})$, then for any $\mathbf{r} \neq \lambda \mathbf{p}$, for all $\lambda > 0$,*

$$\mathbf{p}^\top \mathbf{v}(\mathbf{p}) > \mathbf{p}^\top \mathbf{v}(\mathbf{r}).$$

Proof. Define $\hat{\mathbf{p}} = \mathbf{p}/\mathbf{1}^\top \mathbf{p}$ and $\hat{\mathbf{r}} = \mathbf{r}/\mathbf{1}^\top \mathbf{r}$ (i.e., consider the normalization of each forecast). Because the subgradient mapping for H is 0-homogeneous, Equation 5.2 holds if, and only if, $\mathbf{p}^\top \mathbf{v}(\hat{\mathbf{p}}) > \mathbf{p}^\top \mathbf{v}(\hat{\mathbf{r}})$ for subgradients $\mathbf{v}(\hat{\mathbf{p}})$

and $\mathbf{v}(\hat{\mathbf{r}})$ of H at $\hat{\mathbf{p}}$ and $\hat{\mathbf{r}}$, respectively. In turn, this inequality holds if, and only if, $\hat{\mathbf{p}}^\top \mathbf{v}(\hat{\mathbf{p}}) > \hat{\mathbf{p}}^\top \mathbf{v}(\hat{\mathbf{r}})$. By strict convexity of H on $\mathcal{P}^n$, the final strict inequality holds whenever $\hat{\mathbf{r}} \neq \hat{\mathbf{p}}$, which is indeed the case provided $\mathbf{r}$ is not equivalent to $\mathbf{p}$ up to positive scaling, as assumed in the proposition. □

Proposition 5.1. *The linear forecast scoring rule S_H^{lin} has the following properties:*

1. *The function $S(\omega, \mathbf{r}) := S_H^{lin}(\omega, \mathbf{X}\mathbf{r})$ is a proper distributional scoring rule on $\mathcal{A}^n$.*

2. *The rule S_H^{lin} is strictly proper over linear forecasts, by which we mean the following. Let $\mathbf{f} = \mathbf{X}\mathbf{p}$ for $\mathbf{p} \in \mathcal{P}^n$. Then $E_{\mathbf{p}}[S_H^{lin}(\omega, \mathbf{f})] > E_{\mathbf{p}}[S_H^{lin}(\omega, \mathbf{g})]$ for all $\mathbf{g} \neq \lambda \mathbf{f}, \lambda > 0, \mathbf{g} \in \mathcal{F}(\mathbf{X})$.*

Proof. Define $\hat{H}(\mathbf{p}) := H(\mathbf{X}\mathbf{p})$ for $\mathbf{p} \in \mathcal{A}^n$. To prove the first claim, note that the function $\hat{H}(\mathbf{p}) := H(\mathbf{X}\mathbf{p})$ is 1-homogeneous and convex in $\mathbf{p}$. Thus, the claim follows immediately from Theorem 2.1 by forming subgradients.

To prove the second claim, observe that the inequality $E_{\mathbf{p}}[S_H^{lin}(\omega, \mathbf{f})] > E_{\mathbf{p}}[S_H^{lin}(\omega, \mathbf{g})]$ for $\mathbf{f} = \mathbf{X}\mathbf{p}$ and $\mathbf{g}$ in $\mathcal{F}(\mathbf{X})$ is equivalent to $(\mathbf{X}\mathbf{p})^\top \mathbf{v}(\mathbf{f}) > (\mathbf{X}\mathbf{p})^\top \mathbf{v}(\mathbf{g})$ or

$$\mathbf{f}^\top \mathbf{v}(\mathbf{f}) > \mathbf{f}^\top \mathbf{v}(\mathbf{g}). \tag{5.2}$$

Because, by assumption, $\mathbf{g}$ is not equivalent to $\mathbf{f}$ up to positive scaling, the inequality in (5.2) holds by Lemma 5.1. □

As just demonstrated, linear scoring rules encourage a forecaster to report a forecast $\mathbf{f}$ that is consistent with his privately held probabilistic beliefs.

In particular, if $\mathbf{p}$ represents the forecaster's privately held subjective distribution, the forecaster maximizes his expected score by reporting $\mathbf{f} = \mathbf{X}\mathbf{p}$. Construction of the linear scoring rule requires existence of subgradients of H on $\mathcal{A}^\ell$ (the existence of subgradients on $\mathcal{A}^{\ell+}$ follows from the convexity of H). Conveniently, choosing strict convexity to hold on $\mathcal{P}^\ell$ allows us to use the same entropy functions used to generate distributional scoring rules to construct linear scoring rules.

Mathematically, as the preceding proposition illustrates, there is no distinction between proper distributional scores and strictly proper linear scores. For the distributional score $S(\omega, \mathbf{r}) = S_H^{lin}(\omega, \mathbf{X}\mathbf{r})$, we have $S_H^{lin}(\omega, \mathbf{f}) = S(\omega, \mathbf{r})$ for $\mathbf{r} \in \mathcal{T}(\mathbf{f})$. This observation is used to prove the following fact:

Proposition 5.2. *For $\mathbf{f} = \mathbf{X}\mathbf{p}$, we have that $E_\mathbf{p}[S_H^{lin}(\omega, \mathbf{f})] = E_\mathbf{p}[S_H^{lin}(\omega, \mathbf{g})]$ if, and only if, $\mathcal{T}(\mathbf{f}) = \lambda\mathcal{T}(\mathbf{g})$, for some $\lambda > 0$.*

Proof. As seen in the proof of Proposition 5.1, $E_\mathbf{p}[S_H^{lin}(\omega, \mathbf{f})] = E_\mathbf{p}[S_H^{lin}(\omega, \mathbf{g})]$ implies $\mathbf{g} = \lambda\mathbf{f}$, which means that $\mathcal{T}(\mathbf{g}) = \lambda\mathcal{T}(\mathbf{f})$. Conversely, if $\mathcal{T}(\mathbf{f}) = \lambda\mathcal{T}(\mathbf{g})$, then

$$E_\mathbf{p}[S_H^{lin}(\omega, \mathbf{f})] = E_\mathbf{p}[S(\omega, \mathbf{p})] = E_\mathbf{p}[S(\omega, \mathbf{p}/\lambda)] = E_\mathbf{p}[S_H^{lin}(\omega, \mathbf{g})].$$

□

Thus, we can interpret linear scoring rules from a distributional scoring perspective, in the sense that the reported linear forecast must admit a truth set that contains the forecaster's subjective distribution (up to scaling). The distributional scoring rule $S(\omega, \mathbf{r}) = S_H^{lin}(\omega, \mathbf{X}\mathbf{r})$ will seldom be strictly proper; given subjective distribution $\mathbf{p}$, we have $E_\mathbf{p}[S(\omega, \mathbf{p})] = E_\mathbf{p}[S(\omega, \mathbf{r})]$ for any $\mathbf{r} \in \mathcal{T}(\mathbf{X}\mathbf{p})$. Conversely, when $\mathbf{r} \notin \mathcal{T}(\mathbf{X}\mathbf{p})$, the underlying proper distributional

scoring rule distinguishes the two distributions, and thus the linear scoring rule distinguishes the linear forecasts $\mathbf{Xp}$ and $\mathbf{Xr}$. The following proposition defines when the linear forecast score $S_H^{lin}(\omega, \mathbf{Xr})$ is a strictly proper distributional scoring rule.

Proposition 5.3. *Let S be the distributional scoring rule defined by taking subgradients of the function $\hat{H}(\mathbf{p}) := H(\mathbf{Xp})$ (i.e., $S(\omega, \mathbf{r}) = S_H^{lin}(\omega, \mathbf{Xr})$), where H is an entropy measure. Then S is a strictly proper distributional scoring rule if, and only if, $\mathbf{X}$ has rank n.*

Proof. By definition, S_H^{lin} is a strictly proper distributional scoring rule if, and only if, $(\mathbf{Xp})^\top \mathbf{v}(\mathbf{Xp}) > (\mathbf{Xp})^\top \mathbf{v}(\mathbf{Xq})$ for all $\mathbf{p}, \mathbf{q} \in \mathcal{P}^n, \mathbf{p} \neq \mathbf{q}$. Applying Lemma 5.1, by the strict convexity of H on $\mathcal{P}^n$, this inequality is satisfied if, and only if, $\mathbf{Xp} \neq \lambda \mathbf{Xq}$ for all such choices of $\mathbf{p}, \mathbf{q} \in \mathcal{P}^n$ and $\lambda > 0$.

Thus, we conclude that the theorem follows if we demonstrate that the following statement holds if, and only if, $\mathbf{X}$ has rank n:

$$\mathbf{X}(\mathbf{p} - \lambda \mathbf{q}) \neq \mathbf{0}, \forall \mathbf{p}, \mathbf{q} \in \mathcal{P}, \mathbf{p} \neq \mathbf{q}, \lambda > 0. \tag{5.3}$$

If $\mathbf{X}$ is a rank n matrix, then Equation (5.3) holds because $(\mathbf{p} - \lambda \mathbf{q}) \neq \mathbf{0}$.

For the other direction, we show that if $\mathbf{X}$ is not rank n, then Equation (5.3) does not hold. If $\text{Rank}(\mathbf{X}) < n$, then there exists $\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \neq \mathbf{0}$, such that $\mathbf{Xy} = \mathbf{0}$. Clearly, we can write (in many ways) $\mathbf{y} = \mathbf{y}^+ - \mathbf{y}^-$ for $\mathbf{y}^+, \mathbf{y}^- \geq 0,\ \mathbf{y}^+ \neq 0, \mathbf{y}^- \neq 0$ and $\mathbf{y}^+ \neq \mathbf{y}^-$ up to positive scaling. Dividing both sides by $\mathbf{1}^\top \mathbf{y}^+$ forms the vector

$$\hat{\mathbf{y}} = \mathbf{y}^+ / \mathbf{1}^\top \mathbf{y}^+ - \mathbf{y}^- / \mathbf{1}^\top \mathbf{y}^+,$$

which will also satisfy $\mathbf{X}\hat{\mathbf{y}} = \mathbf{0}$. Thus, choosing $\mathbf{p} = \mathbf{y}^+ / \mathbf{1}^\top \mathbf{y}^-$ and $\lambda, \mathbf{q}$ such that $\lambda \mathbf{q} = \mathbf{y}^- / \mathbf{1}^\top \mathbf{y}^+$ means that Equation (5.3) will not hold. □

Proposition 5.3 demonstrates that the general construction method for linear score functions can be used to generate strictly proper scoring rules as well. The scoring rule is proper when the set $\mathcal{T}(\mathbf{f}^*)$ is a single point for every $\mathbf{f}^*$. Section 5.3.2 provides examples of strictly proper scoring rules created in this fashion.

Finally, we prove a partial converse to Proposition 5.1. This result can be seen as a special case of the characterization for scoring linear properties given in [2] (also see [71]), where it is shown that every linear scoring rule can be cast as a Bregman rule. However, in the case of the matrix-based construction (based on the Hendrickson and Buehler [46] view of scoring rules) presented in Definition 5.1, the characterization presented below has a simple form.

Proposition 5.4. *Let $\mathbf{X}$ be a query matrix, and let $\mathbf{G} \in \mathbb{R}^{n\times\ell}$ be a left generalized inverse for $\mathbf{X}$, that is, a matrix satisfying $\mathbf{XGX} = \mathbf{X}$. Suppose that the scoring rule $S : \Omega \times \mathcal{F}(\mathbf{X}) \to \mathbb{R}$ satisfies $E_{\mathbf{p}}[S(\omega, \mathbf{Xp})] > E_{\mathbf{p}}[S(\omega, \mathbf{g})]$ for all $\mathbf{g} \in \mathcal{F}(\mathbf{X})$. Then the following statements hold:*

1. *If $\mathbf{X}$ is rank(n), then $S(\mathbf{f}) = \mathbf{X}^\top \mathbf{v}(\mathbf{f})$, where $\mathbf{v}(\mathbf{f}) \in \partial H(\mathbf{f})$ for some entropy measure H defined on $\mathcal{F}(\mathbf{X})$.*

2. *If the linear scoring rule S is differentiable in $\mathbf{f}$, and if there exists a generalized inverse $\mathbf{G} \geq \mathbf{0}$ (element-wise), then $\mathbf{GX}S(\omega, \mathbf{f}) = \mathbf{X}^\top \nabla_{\mathbf{f}} H(\mathbf{f})$ for some entropy measure H.*

Proof. Let $H(\mathbf{p}) = E_{\mathbf{p}}[S(\omega, \mathbf{Xp})]$ be the convex and 1-homogeneous optimal score function associated with the underlying proper distributional scoring rule $S(\mathbf{p}) := S^{lin}(\mathbf{Xp})$.

1. Under the posited conditions, any generalized inverse $\mathbf{G}$ is a true inverse for $\mathbf{X}$, in that $\mathbf{GXp} = \mathbf{p}$ for all $\mathbf{p} \in \mathcal{A}$. Consider the function $\hat{H}(\mathbf{f}) = H(\mathbf{Gf})$ for $\mathbf{f} \in \mathcal{F}(\mathbf{X})$. $\hat{H}$ is convex in $\mathbf{f}$, and $\hat{H}(\mathbf{Xp}) = H(\mathbf{p})$. Taking subgradients with respect to $\mathbf{p}$ on both sides of this equality proves the first statement.

2. Define $\hat{H}(\mathbf{f}) = H(\mathbf{Gf})$ as in the proof of the first case. The condition $\mathbf{G} \geq \mathbf{0}$ guarantees that $\mathbf{Gf}$ is in the domain of H, and thus $\hat{H}$ is convex in $\mathbf{f}$. By rearranging, we have

$$H(\mathbf{p}) = \hat{H}(\mathbf{Xp}) + [(\mathbf{I} - \mathbf{GX})\mathbf{p}]^\top S(\mathbf{Xp}),$$

where $\mathbf{I}$ is the identity matrix.

Taking gradients on both sides w.r.t $\mathbf{p}$ (using the product rule) yields

$$\begin{aligned} S(\mathbf{p}) = \mathbf{X}^\top \nabla_{\mathbf{Xp}} H(\mathbf{Xp}) + (\mathbf{I} - \mathbf{GX})^\top S(\mathbf{p}) \\ + [\mathbf{X}(\mathbf{I} - \mathbf{GX})\mathbf{p}]^\top \mathbf{M}, \end{aligned}$$

where $\mathbf{M}$ is a matrix of partial derivatives associated with $S(\mathbf{p})$. The final term is $\mathbf{0}$ by definition of the left inverse. Thus, using the fact that $(\mathbf{I} - \mathbf{GX})$ is a projection matrix, and therefore equal to its transpose, re-arrangement yields

$$\mathbf{GX}S^{lin}(\mathbf{f}) = \mathbf{X}^\top \nabla_{\mathbf{f}} H(\mathbf{f}).$$

□

The second part of Proposition 5.4 shows if the matrix $\mathbf{X}$ has a positive left inverse, as in the examples in [65], the scoring rule $S(\mathbf{f})$ is of the form given in Definition 5.1 up to positive linear scaling. Where $\mathbf{X}$ lacks such an

inverse, the relationship in part two will still hold, although it is not clear that H will be convex. We now present a simple but useful extension of linear scoring rules.

5.3.1 An Extension of Linear Scoring Rules

As an extension to the linear scoring rules presented up to now, we demonstrate that forecasts outside of $\mathcal{F}(\mathbf{X})$ can be scored similarly provided there exists a one-to-one mapping from the forecast space to $\mathcal{F}(\mathbf{X})$. We assume $\mathbf{Xp} = z(\mathbf{f})$, where z is a function of the elicited forecast $\mathbf{f}$.

Proposition 5.5. *Assume* $\mathbf{X}$ *is a query matrix with* $\mathbf{X}(1) = \mathbf{1}^\top$*, and* H *an entropy measure on* $\mathcal{A}^\ell$*. Furthermore, let* $\mathbf{f} \in \mathbb{R}^c$ *and let* $z : \mathbb{R}^c \to \mathcal{F}(\mathbf{X})$ *be a one-to-one map. Then the scoring rule* $\hat{S}(\omega, \mathbf{f}) := S_H^{lin}(\omega, z(\mathbf{f}))$ *satisfies* $E_{\mathbf{p}}[\hat{S}(\omega, \mathbf{f})] > E_{\mathbf{p}}[\hat{S}(\omega, \mathbf{g})]$*, for* $\mathbf{f} = z^{-1}(\mathbf{Xp})$*, and for all* $\mathbf{g} \in \{\mathbf{f} : z(\mathbf{f}) = \mathbf{Xr}$ *for some* $\mathbf{r} \in \mathcal{A}^n\}$*.*

Proof. To obtain a contradiction, suppose that there exists $\mathbf{g} \neq \mathbf{f}$ satisfying the requirements in the proposition such that $E_{\mathbf{p}}[\hat{S}(\omega, \mathbf{f})] = E_{\mathbf{p}}[\hat{S}(\omega, \mathbf{g})]$. By the established properties of linear scoring rules, this can be the case only if $\lambda z(\mathbf{g}) = \mathbf{Xp}$. By assumption, there is an $\mathbf{r} \in \mathcal{A}^n$ such that $z(\mathbf{g}) = \mathbf{Xr}$, so the equality can be rewritten $\mathbf{Xp} = \lambda \mathbf{Xr}$. The first equality in this system of equations $\mathbf{1}^\top \mathbf{p} = \lambda \mathbf{1}^\top \mathbf{r}$ implies that $\lambda = 1$ and $\mathbf{Xp} = \mathbf{Xr}$. Because z is one-to-one, this implies $\mathbf{f} = \mathbf{g}$, which is a contradiction. □

When $z(\mathbf{f}) = \mathbf{f}$, this proposition demonstrates that scoring the normalizing constant guarantees that $\mathbf{f} = \mathbf{Xp}$ uniquely maximizes the expected score

among all possible forecasts consistent with a normalized probability distribution. Proposition 5.5 also provides useful insights into how to apply linear scoring rules.

1. One need not explicitly elicit a forecast for the normalizing constant of the forecaster's underlying distribution. Instead, one can ask only for the probabilities of interest, acquiring response $\mathbf{f}$, and then score $z(\mathbf{f}) = [1, \mathbf{f}]$.

2. One can elicit fewer probabilities than are actually scored. For example, complete marginal distributions can be rewarded in the score function, but only those probabilities needed to characterize the marginal distributions need be elicited.

3. Related to the last observation, invertible linear transformations of the obtained forecast will preserve strict propriety. For invertible matrix $\mathbf{G}$, we can obtain a proper score for $\mathbf{f} \in \mathcal{F}(\mathbf{X})$ by applying a linear scoring rule for $\mathbf{Gf}$ with $\mathbf{GX}$ in place of $\mathbf{X}$ in the construction.

   This property is particularly useful for comparing forecasts obtained with query matrices $\mathbf{X}$ and $\hat{\mathbf{X}}$ that have the same row-reduced forms. In practical terms, these matrices ask for the same forecasts in different manners. Proposition 5.5 allows the forecasts associated with these queries to be transformed (via matrix multiplication) into comparable forms without sacrificing propriety.

   For example, the two matrices:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

both lead to forecasts that fully characterize the C.D.F. of a distribution $\mathbf{r}$. However, two forecasts whose truth sets are equal might be scored differently depending on the form of the matrix. Thus, a forecast obtained using the first query matrix cannot be directly compared to a forecast obtained using the second. However, using elementary row operations, a forecast using the second query matrix can be converted into an equivalent forecast corresponding to the first matrix, and scored accordingly.

Proposition 5.5 also ties directly into the notion of elicitability complexity discussed in [57], as it provides a method for scoring forecasts for properties that, along with other information captured in $\mathbf{f}$, can only indirectly be rewarded under a linear scoring rule.

Having introduced a method for constructing linear scoring rules, we now describe additional properties and provide examples that demonstrate commonly used distributional scoring rules can be seen as special cases of linear scoring rules. We also derive linear scoring rules for practical applications.

5.3.2 Linear Scoring Rules: Properties and Examples

We now give a more detailed account of the structure of the linear scoring rule S_H^{lin} constructed in Definition 5.1. We can write:

$$S_H^{lin}(\omega, \mathbf{f}) = \boldsymbol{\delta}_\omega^\top \left(\mathbf{X}^\top \mathbf{v}(\mathbf{f})\right) = \sum_{i=1}^{p} \mathbf{X}(i)_\omega v(\mathbf{f})_i, \tag{5.4}$$

where $\mathbf{X}(i)_\omega$ is the component of the ith row of $\mathbf{X}$ corresponding to ω, and $v(\mathbf{f})_i$ is the ith component of $\mathbf{v}(\mathbf{f}) \in \partial H(\mathbf{f})$. The expected score for forecast $\mathbf{g}$ under the distribution $\mathbf{p}$ is $(\mathbf{Xp})^\top \mathbf{v}(\mathbf{g})$, and the divergence measure associated with S_H^{lin} is:

$$E_{\mathbf{p}}[S_H^{lin}(\mathbf{Xp})] - E_{\mathbf{p}}[S_H^{lin}(\mathbf{g})] = H(\mathbf{Xp}) - (\mathbf{Xp})^\top \mathbf{v}(\mathbf{g}),$$

which, when H is differentiable, is the Bregman Divergence between $\mathbf{f} = \mathbf{Xp}$ and $\mathbf{g}$ [32]. Like standard distributional scoring rules, linear scoring rules have the property that if αS_H^{lin} is strictly proper, then so is $\alpha S_H^{lin} + \beta$ for $\alpha > 0, \beta \in \mathbb{R}$.

As illustrated in Equation (5.4), the score for forecast $\mathbf{f}$, given an observation ω, is additive (however, not in the sense of Chapter 3). If $\mathbf{X}(i)_\omega > 0$, then we add $\mathbf{X}(i)_\omega v(\mathbf{f})_i$ to the total score. When we observe ω, we score the quality of the "answer" f_i when query $\mathbf{X}(i)$ collects information on probability p_ω. Our construction method easily allows extension of traditional scoring rules to linear forecasts, as in the following examples that use the Quadratic, Spherical, and Logarithmic scores.

Example 5.1. *Linear Quadratic Score*

Set

$$H(\mathbf{Xp}) = \frac{1}{\mathbf{1}^\top \mathbf{Xp}} \left\|\mathbf{Xp}\right\|_2^2 .$$

The associated distributional proper scoring rule is

$$S_H(\omega, \mathbf{p}) = (\mathbf{X}\boldsymbol{\delta}_\omega)^\top \left[2\frac{\mathbf{Xp}}{\mathbf{1}^\top \mathbf{Xp}} - \left\| \frac{\mathbf{Xp}}{\mathbf{1}^\top \mathbf{Xp}} \right\|_2^2 \mathbf{1} \right],$$

and the strictly proper linear scoring rule is

$$S_H^{lin}(\omega, \mathbf{f}) = (\mathbf{X}\boldsymbol{\delta}_\omega)^\top \left[2\frac{\mathbf{f}}{\mathbf{1}^\top \mathbf{f}} - \left\| \frac{\mathbf{f}}{\mathbf{1}^\top \mathbf{f}} \right\|_2^2 \mathbf{1} \right].$$

Example 5.2. *Linear Spherical Score*

Set

$$H(\mathbf{Xp}) = \left\|\mathbf{Xp}\right\|_2 .$$

The associated distributional proper scoring rule is

$$S_H(\omega, \mathbf{p}) = (\mathbf{X}\boldsymbol{\delta}_\omega)^\top \left[\frac{\mathbf{Xp}}{\|\mathbf{Xp}\|_2}\right],$$

and the strictly proper linear scoring rule is

$$S_H^{lin}(\omega, \mathbf{p}) = (\mathbf{X}\boldsymbol{\delta}_\omega)^\top \left[\frac{\mathbf{f}}{\|\mathbf{f}\|_2}\right].$$

Example 5.3. *Linear Logarithmic Score*

Set

$$H(\mathbf{Xp}) = (\mathbf{Xp})^\top \ln\left(\frac{\mathbf{Xp}}{\mathbf{1}^\top\mathbf{Xp}}\right).$$

The associated distributional proper scoring rule is

$$S_H(\omega, \mathbf{p}) = (\mathbf{X}\boldsymbol{\delta}_\omega)^\top \ln\left(\frac{\mathbf{Xp}}{\mathbf{1}^\top\mathbf{Xp}}\right),$$

and the strictly proper linear scoring rule is

$$S_H^{lin}(\omega, \mathbf{p}) = (\mathbf{X}\boldsymbol{\delta}_\omega)^\top \ln\left(\frac{\mathbf{f}}{\mathbf{1}^\top\mathbf{f}}\right).$$

Linear scoring rules remain proper, in the sense described below, when the matrix $\mathbf{X}$ is, from the perspective of the forecaster, random (in potentially both the dimension ℓ and the individual entries). We have

$$\begin{aligned}
&\max_{\mathbf{f}(\mathbf{X})\in\mathcal{F}(\mathbf{X})} E\left[S_{H^{\mathbf{X}}}^{lin}(\omega, \mathbf{f}(\mathbf{X}))\right] \\
=&E\left[\max_{\mathbf{f}\in\mathcal{F}(\mathbf{X})} E_{\mathbf{p}}[S_{H^{\mathbf{X}}}^{lin}(\omega, \mathbf{f})|\mathbf{X}]\right] \\
=&E\left[E_{\mathbf{p}}\left[S_{H^{\mathbf{X}}}^{lin}(\omega, \mathbf{Xp})\right]\right].
\end{aligned} \tag{5.5}$$

Here, the notation $H^{\mathbf{X}}$ denotes that the chosen entropy measure may depend on the realization of X (particularly, its row dimension ℓ). Two implications

follow. First, consider a sequential forecasting exercise where queries, in the form of a rows $\mathbf{X}(i)$ of $\mathbf{X}$, are answered by the forecaster in turn. Then the forecaster's optimal a priori strategy, assuming that the answer f_i to a given query does not influence future questions, is to tell the truth for each response. Second, viewing S_H^{lin} as a proper distributional scoring rule (as in Proposition 5.3), the forecaster's expected score is maximized by reporting $\mathbf{p}$ even when $\mathbf{X}$ (and thus the exact score function) is viewed as random by the forecaster.

We now give examples of linear scoring rules for different query matrices.

Example 5.4. *Marginal Information*

Suppose that $\mathbf{p}$ *is the joint distribution for a* J*-dimensional joint random variable. Let* $\mathbf{M}^j$ *be a suitably chosen* 0-1 *matrix such that* $\mathbf{M}^j\mathbf{p} = \mathbf{p}^j$*, where* $\mathbf{p}^j$ *is the marginal distribution for the* j*th dimension. Set*

$$\mathbf{X} = \begin{bmatrix} \mathbf{M}^1 \\ \mathbf{M}^2 \\ \vdots \\ \mathbf{M}^J \end{bmatrix}$$

so that $\mathbf{f} = \mathbf{X}\mathbf{p} = [\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^J]$*. This forecast provides the marginal information associated with the joint distribution* $\mathbf{p}$*.*

We can write

$$\mathbf{v}(\mathbf{f}) = [\mathbf{v}(\mathbf{p}^1), \mathbf{v}(\mathbf{p}^2), \ldots, \mathbf{v}(\mathbf{p}^1)], \tag{5.6}$$

i.e., the vector $\mathbf{v}(\mathbf{f})$ *has component blocks corresponding to each marginal forecast. Suppose that the joint event* $\omega = (\omega^1, \omega^2, \ldots, \omega^J)$ *is observed. This implies that*

$$S_H^{lin}(\omega, \mathbf{f}) = \sum_{j=1}^{J} \mathbf{v}(\mathbf{p}^j)_{\omega^j}.$$

The linear score decomposes into a sum where each term corresponds to an observed marginal outcome ω^j. We conclude from Proposition 5.3 that $S_H^{lin}(\omega, \mathbf{Xp})$ is not a strictly proper distributional scoring rule, because the matrix $\mathbf{X}$ is not invertible and two distributions with equivalent marginal decompositions will recieve identical scores.

Example 5.5. *Scoring Cumulative Distribution Forecasts*

Let $X : \Omega \to \mathbb{R}$ be a discrete random variable, with $|\Omega| = n$. Further assume, without loss of generality, that $X(\omega_1) \leq X(\omega_2) \leq \ldots \leq X(\omega_n)$. Let $\mathbf{X} \in \mathbb{R}^{n \times n}$, and set

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ \vdots & & & & \\ 1 & 1 & 1 & \ldots & 1 \end{bmatrix}. \tag{5.7}$$

The forecast $\mathbf{f} = \mathbf{Xp}$ gives the C.D.F for X, and the linear scoring rule

$$S_H^{lin}(\omega_i, \mathbf{f}) = \sum_{j \geq i} v(\mathbf{f})_j \ \mathbf{v}(\mathbf{f}) \in \partial H(\mathbf{f}) \tag{5.8}$$

for entropy measure H. Applying Proposition 5.3, this linear scoring rule for C.D.Fs is also strictly proper for distributional forecasts.

Alternatively, one could define a linear scoring rule for C.D.Fs as follows. Let $f_i = P(X \leq X(\omega_i)$ and let $H^i : \mathcal{F}^2 \to \mathbb{R}, i = 1, \ldots, n$ be entropy functions and set

$$H(\mathbf{f}) = \sum_{i=1}^{n} H^i([f_i, 1 - f_i]). \tag{5.9}$$

H can easily be written as the entropy measure for a linear scoring rule. The resulting score, upon taking subgradients, can be seen to be equivalent to the CDF scoring rules in [59]. In fact, it is equivalent to a sum of strictly proper distributional scoring rules for binary distributions.

As demonstrated in Proposition 5.3, we can also construct strictly proper scoring rules by choosing $\mathbf{X}$ to be an invertible matrix. We give examples below.

Example 5.6. *Weighted Scoring Rules*

A weighted scoring rule incorporates a baseline distribution $\mathbf{q} \in \mathcal{P}^n,\ \mathbf{q} > 0$ *directly into the score function [54, 36]. This baseline function could represent a publicly available forecast or could come from empirical data.*

We construct a proper weighted scoring rule as follows. Choose an entropy measure H *with* $H(\mathbf{1}) = \min_{\mathbf{f} \in \mathcal{F}^n} H(\mathbf{f})$. *Set*

$$\mathbf{X} = \boldsymbol{Diag}\left(\left[\frac{1}{q_{\omega_1}}, \frac{1}{q_{\omega_2}}, \ldots, \frac{1}{q_{\omega_n}}\right]\right),$$

that is, $\mathbf{X}_{i,i} = 1/q_i$ *and* $\mathbf{X}_{i,j} = 0$.

Following Propositions 5.1 and 5.3, the scoring rule derived from $\hat{H}(\mathbf{p}) = H(\mathbf{X}\mathbf{p})$ *is*

$$S(\omega, \mathbf{p}) = \frac{1}{q_\omega} v\left(\left[\frac{p_{\omega_1}}{q_{\omega_1}}, \frac{p_{\omega_2}}{q_{\omega_2}}, \ldots, \frac{p_{\omega_n}}{q_{\omega_n}}\right]\right)_\omega,$$

where

$$\mathbf{v}\left(\left[\frac{p_{\omega_1}}{q_{\omega_1}}, \frac{p_{\omega_2}}{q_{\omega_2}}, \ldots, \frac{p_{\omega_n}}{q_{\omega_n}}\right]\right) \in \partial H\left(\left[\frac{p_{\omega_1}}{q_{\omega_1}}, \frac{p_{\omega_2}}{q_{\omega_2}}, \ldots, \frac{p_{\omega_n}}{q_{\omega_n}}\right]\right).$$

The optimal expected score for the forecaster is

$$H\left(\left[\frac{p_{\omega_1}}{q_{\omega_1}}, \frac{p_{\omega_2}}{q_{\omega_2}}, \ldots, \frac{p_{\omega_n}}{q_{\omega_n}}\right]\right),$$

which, due to our choice of H*, implies that the forecaster's optimal expected score is minimized at* $\mathbf{q}$.

Example 5.7. *Local Scoring Rule*

A local scoring rule rewards a distribution forecast $\mathbf{p} \in \mathcal{A}^{n+}$ *based on only some of its components. In [31], the authors present a general method for constructing these rules (and further show that all local rules are of the proposed form). These local rules can be viewed as linear scoring rules, as described below.*

Choose $\mathbf{X}^1, \ldots, \mathbf{X}^K$ *to be* 0-1 *matrices, where* $\mathbf{X}^k$ *is dimension* $\ell(k) \times n$ *with* $2 \leq \ell(k) \leq n$*. Each matrix* X^k *maps the distribution* $\mathbf{p}$ *to a subset of* $\ell(k)$ *of its components. We define the matrix:*

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \\ \vdots \\ \mathbf{X}^J \end{bmatrix}. \tag{5.10}$$

We derive a scoring rule from the 1*-homogeneous convex function*

$$H(\mathbf{Xp}) = \sum_{k=1}^{K} H^k(\mathbf{X}^k\mathbf{p}), \tag{5.11}$$

where each $H^k : \mathbb{R}^{\ell(k)} \to \mathbb{R}$ *is itself an entropy measure on the lower dimensional space that accords with the mapping* $\mathbf{p} \to \mathbf{X}^k\mathbf{p}$*. The resulting scoring rule rewards an observation* ω *according to*

$$S(\omega, \mathbf{p}) = \sum_{k=1}^{K} (\mathbf{X}^k\boldsymbol{\delta}_\omega)^T \mathbf{v}^k(\mathbf{X}^k\mathbf{p}).$$

If $\mathbf{X}^k\mathbf{p}$ *does not preserve information regarding* p_ω*, then the contribution to the sum corresponding with the index* k *is* 0*.*

We now identify those matrices $\mathbf{X}$ *for which the function* H *defined in Equation* (5.11) *generates a strictly proper scoring rule. By Proposition 5.3,*

the matrix $\mathbf{X}$ *must be full-row rank. However, even when this condition holds, and each function* H^k *is strictly convex on* $\mathcal{P}^{l(k)}$*, the function* $H(\mathbf{Xp})$ *might not be strictly convex on* $\mathcal{P}^n$ *(in turn implying that the associated scoring rule is proper but not strictly proper). The following proposition provides an algebraic condition for determining whether a local rule is strictly proper.*

Proposition 5.6. *Define* $\mathbf{X}$ *as in Equation* (5.10) *and* $H(\mathbf{Xp})$ *as in Equation* (5.11)*. Then the scoring rule:*

$$S(\omega, \mathbf{p}) = \sum_{k=1}^{K} (\mathbf{X}^k \boldsymbol{\delta}_\omega)^T \mathbf{v}^k(\mathbf{X}^k \mathbf{p}), \tag{5.12}$$

with $\mathbf{v}^k(\mathbf{X}^k\mathbf{p}) \in \partial H^k(\mathbf{X}^k\mathbf{p})$ *for* $k = 1, \ldots, K$*, is strictly proper if* $\mathbf{X}$ *satisfies*

$$\begin{bmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \\ \vdots \\ \mathbf{X}^J \end{bmatrix} \mathbf{f} = \begin{bmatrix} c^1\mathbf{1}^1 \\ c^2\mathbf{1}^2 \\ \vdots \\ c^j\mathbf{1}^J \end{bmatrix} \implies \mathbf{f} = c\mathbf{1}, \tag{5.13}$$

where $\mathbf{f} \in \mathcal{P}^n$*,* $c^1, \ldots, c^K, c$ *are arbitrary positive constants, and* $\mathbf{1}^k$ *is a vector of* 1*s of length* $\ell(k)$*.*

Proof. Choose $\mathbf{r}, \mathbf{p} \in \mathcal{P}^+$. It follows immediately from the subgradient inequality that $E_\mathbf{p}[S(\omega, \mathbf{r})] \leq E_\mathbf{p}[S(\omega, \mathbf{p})]$, for S defined in Equation (5.12). We will show that $E_\mathbf{p}[S(\omega, \mathbf{r})] = E_\mathbf{p}[S(\omega, \mathbf{p})]$ implies that $\mathbf{r} = \mathbf{p}$.

The equality

$$E_\mathbf{p}[S(\omega, \mathbf{r})] = \sum_{k=1}^{K} (\mathbf{X}^k\mathbf{p})^T \mathbf{v}^k(\mathbf{X}^k\mathbf{r}) = \sum_{k=1}^{K} (\mathbf{X}^k\mathbf{p})^T \mathbf{v}^k(\mathbf{X}^k\mathbf{p}) = E_\mathbf{p}[S(\omega, \mathbf{p})]$$

implies that $\mathbf{X}^k\mathbf{p} = c^k\mathbf{X}^k\mathbf{r}$ for all k, due to the assumption that the functions H^k are strictly convex on $\mathcal{P}^{\ell(k)}$. This condition can be expressed as

$$\begin{bmatrix}\mathbf{X}^1\\ \mathbf{X}^2\\ \vdots\\ \mathbf{X}^J\end{bmatrix}\left[\frac{p_1}{r_1},\ldots,\frac{p_n}{r_n}\right] = \begin{bmatrix}c^1\mathbf{1}^1\\ c^2\mathbf{1}^2\\ \vdots\\ c^j\mathbf{1}^J\end{bmatrix}, \tag{5.14}$$

which implies, by Equation (5.13) and because $\mathbf{r}$ and $\mathbf{p}$ are normalized probability distributions, that $\mathbf{r} = \mathbf{p}$. □

Condition (5.13) *implies that* $\mathbf{Xf} = \mathbf{0} \implies \mathbf{f} = \mathbf{0}$, *consistent with Proposition 5.3.*

Example 5.8. *Scoring Conditional Distributions and Additive Scoring Rules*

This example constructs linear scoring rules for conditional distributions. For the collection of random variables $Y^j, j = 1,\ldots,J$ *and joint random variable* Y, *let* $\mathbf{r}$ *represent the joint distribution, which we assume is in* $\mathcal{A}^+$, *so that all conditional distributions are well defined. Analogous to previous chapters,* $\mathbf{r}^{j|y^{i(1)},y^{i(2)},\ldots,y^{i(K)}}$ *denotes the conditional distribution, associated with* $\mathbf{r}$, *for the random variable* Y^j *given an observation* $Y^{i(k)} = y^{i(k)}$ *for* K *of the other random variables.*

The conditional distribution can be expressed in terms of matrix multiplication:

$$\mathbf{r}^{j|y^{i(1)},y^{i(2)},\ldots,y^{i(K)}} = \frac{\mathbf{C}^j(y^{i(1)},y^{i(2)},\ldots,y^{i(K)})\mathbf{r}}{\chi(\{Y^{i(1)} = y^{i(1)}, Y^{i(2)} = y^{i(2)},\ldots,Y^{i(K)} = y^{i(K)}\})}, \tag{5.15}$$

where $\mathbf{C}^j(y^{i(1)},y^{i(2)},\ldots,y^{i(K)}) \in \mathbb{R}^{n^j\times n}$, *and where row* i *of the matrix is*

$$\chi(\{Y^j = y_i^j, Y^{i(1)} = y^{i(1)}, Y^{i(2)} = y^{i(2)},\ldots,Y^{i(K)} = y^{i(K)}\}.$$

Thus, a forecast for the conditional distribution $\mathbf{r}^{j|y^{i(1)},y^{i(2)},\ldots,y^{i(K)}}$ *can be obtained indirectly using the query matrix*

$$\mathbf{X} = \begin{bmatrix} \mathbf{C}^j(y^{i(1)}, y^{i(2)}, \ldots, y^{i(K)}) \\ \chi(\{Y^j = y_i^j, Y^{i(1)} = y^{i(1)}, Y^{i(2)} = y^{i(2)}, \ldots, Y^{i(K)} = y^{i(K)}\} \end{bmatrix} \tag{5.16}$$

and scored using a linear scoring rule.

Next, we consider a collection of proper distributional scoring rules, with respect to distributions in $\mathcal{A}^{n+}$*, that are built by scoring a sequence of conditional distributions. These rules appear in* [31] *and can be derived using the matrix-based construction in Definition 5.1. Let* $H^j : \mathcal{A}^{n^j+} \to \mathbb{R}$ *be entropy measures for* $j = 1, \ldots, J$*, and consider the function*

$$\begin{aligned} &\chi(\{Y^{i(1)} = y^{i(1)}, Y^{i(2)} = y^{i(2)}, \ldots, Y^{i(K)} = y^{i(K)}\} H^j(\mathbf{r}^{j|y^{i(1)},y^{i(2)},\ldots,y^{i(K)}}) \\ &= H^j\left(\mathbf{C}^j(y^{i(1)}, y^{i(2)}, \ldots, y^{i(K)})\mathbf{p}\right), \end{aligned}$$

where the equality follows from representation (5.15) *and the fact that* H^j *is* 1*-homogeneous. The function on the right side of the equality is once again an entropy measure composed with a query matrix, meaning that it produces a distributional scoring rule of the form given in Proposition 5.3. We now consider the entropy measure defined as*

$$H(\mathbf{p}) = E_{\mathbf{p}}\left[\sum_{j=1}^{J} H^j\left(\mathbf{p}^{j|Y^{i(1)},Y^{i(2)},\ldots,Y^{i(K)}}\right)\right] \tag{5.17}$$

$$= \sum_{j=1}^{J}\left[\sum_{y^{i(1)},y^{i(2)},\ldots,y^{i(K)} \in \Omega^{i(1)} \times \Omega^{i(2)} \times \ldots \times \Omega^{i(K)}} H^j\left(\mathbf{C}^j(y^{i(1)}, y^{i(2)}, \ldots, y^{i(K)})\mathbf{p}\right)\right]. \tag{5.18}$$

The entropy measure defined in Equation (5.17) *is the sum of expected entropies for the conditional distributions for* $Y^1, \ldots, Y^J$*, where for each* j*, we*

condition on a subset of the remaining random variables. The expectations are taken with respect to the conditional distributions, which are random prior to a realization of $Y^{i(1)}, Y^{i(2)}, \ldots, Y^{i(K)}$ *for each* j*. Equality* (5.18) *follows from* 1*-homogeneity of the* H^j *after writing out the expectations.*

The entropy measure in Equation (5.18) *is the entropy function associated with a local scoring rule, as described in Example 5.7. Taking subgradients, we can form the linear scoring rule (for distributional forecasts)*

$$S(\omega, \mathbf{r}) = \sum_{j=1}^{J} \boldsymbol{\delta}_{\omega^j}^{\top} \mathbf{v}^j(\mathbf{C}^j(Y^{i(1)}(\omega^{i(1)}), Y^{i(2)}(\omega^{i(2)}), \ldots, Y^{i(K)}(\omega^{i(K)})\mathbf{r}).$$

Now, employing the 0*-homegeneity of the subgradient* $\mathbf{v}^j, j = 1, \ldots, J$*, and Theorem 2.1 yields*

$$S(\omega, \mathbf{r}) = \sum_{j=1}^{J} S^j \left(\omega^j, \mathbf{r}^{j|Y^{i(1)}(\omega^{i(1)}), Y^{i(2)}(\omega^{i(2)}) \ldots, Y^{i(K)}(\omega^{i(K)})}\right), \tag{5.19}$$

where the S^j *are the strictly proper scoring rules associated with the entropy functions* H^j*. This constitutes scoring a joint forecast* $\mathbf{r}$ *by adding the proper scores associated with a sequence of conditional distributions corresponding to the observed joint outcome* ω*. By construction, the rule is proper over joint distributional forecasts. The class of proper distributional scoring rules given in Equation* (5.19) *generalizes the class of additive scoring rules constructed in Chapter 3. Recall that these rules have the property that the joint score is equal to a sum of marginal scores whenever distribution* $\mathbf{r}$ *implies independence between the random variables* $Y^1, \ldots, Y^J$*. In addition, additive scoring rules have the intuitive property that when the joint forecast implies independence, better marginal forecasts (as measured by the rule* S^j*) imply a higher joint score.*

Chapter 3 implicitly applied Proposition 5.6 in showing that the rule (5.19) *is strictly proper in the special case where each variable Y^j is conditioned on realization of all the other random variables. In other cases, rule* (5.19) *will clearly not be strictly proper. For example, one could score the joint distribution* $\mathbf{r}$ *by summing scores for the sequence of associated marginal distributions* $\mathbf{r}^j, j = 1, \ldots, J$. *In this case, as in Example 5.4, the scoring rule will not distinguish joint distributions that have the same marginal decomposition, and thus will not be strictly proper.*

5.4 Applications

This section describes two decision analysis applications where linear scoring rules could be beneficial.

5.4.1 Tailored Scoring Rules for Distributionally Robust Optimization

Once again, let Ω be a finite state space, and consider the generic stochastic optimization problem

$$f(\mathbf{p}) = \sup_{\mathbf{x} \in \mathcal{X}} E_{\mathbf{p}}[u(\mathbf{x}, \omega)], \tag{5.20}$$

where u is a concave function (usually a utility function), and $\mathbf{x} \in \mathcal{P}$ is a tractable constraint set. Recall (from Chapter 4) that this type of decision problem produces a *tailored* proper distributional scoring rule of the form

$$S(\omega, \mathbf{r}) = u(\mathbf{x}^*(\mathbf{r}), \omega), \ \mathbf{x}^*(\mathbf{r}) \in \arg\max_{\mathbf{x} \in \mathcal{X}} E_{\mathbf{r}}[u(\mathbf{x}, \omega)]. \tag{5.21}$$

Here, we consider the following distributionally robust version. Consider the decision problem

$$f_{\mathrm{DR}}(\mathbf{f}) = \sup_{\mathbf{x}\in\mathcal{X}} \left\{ \inf_{\mathbf{p}:\mathbf{X}\mathbf{p}=\mathbf{f}} E_{\mathbf{p}}[u(\mathbf{x},\omega)] \right\}. \tag{5.22}$$

This decision maker seeks to maximize his worst-case expected utility against the uncertainty set $\{\mathbf{p} : \mathbf{X}\mathbf{p} = \mathbf{f}\}$ comprised of all discrete distributions satisfying a linear equality constraint.

We construct a tailored linear scoring rule for the problem (5.22) in the case where the information $\mathbf{f}$ is obtained from a forecast. A tailored rule for this scenario is useful if it is infeasible for the expert to provide an accurate assessment for the entire distribution. For example, the expert might be able to accurately assess only the marginal distributions, or be able to accurately give information on only some of the uncertainties. In the case of partial information, assuming that the forecast can be expressed as $\mathbf{f} = \mathbf{X}\mathbf{p}$, such a rule will align the incentives of the forecaster and those of the ambiguity-averse decision maker, as in the case of traditional tailored distributional scoring rules.

To derive the form of the tailored scoring rule for problem (5.22), we take the dual of the inner linear program, which can be written:

$$\sup_{\boldsymbol{\pi}\in\mathbb{R}^p} \left\{ \mathbf{f}^\top \boldsymbol{\pi} : \mathbf{X}^\top \boldsymbol{\pi} \leq u(\mathbf{x}) \right\} \tag{5.23}$$

for each $\mathbf{x} \in \mathcal{X}$, where $u(\mathbf{x}) := [u(\mathbf{x},\omega_1), u(\mathbf{x},\omega_2), \ldots, u(\mathbf{x},\omega_n)]$. Combining the outer maximization with the inner gives

$$f_{\mathrm{DR}}(\mathbf{f}) = \sup_{\mathbf{x}\in\mathcal{X},\boldsymbol{\pi}\in\mathbb{R}^p} \left\{ \mathbf{f}^\top \boldsymbol{\pi} :\ \mathbf{X}^\top \boldsymbol{\pi} - u(\mathbf{x}) \leq 0,\ \mathbf{x} \in \mathcal{X} \right\}. \tag{5.24}$$

Function $f_{\mathrm{DR}}(\mathbf{f})$ is clearly 1-homogeneous, and is also convex as the supremum of affine functions of $\mathbf{f}$.

When $\mathbf{f}$ is obtained from a forecast, we can write $\mathbf{f} = \mathbf{X}\mathbf{p}'$, where $\mathbf{p}'$ is the forecaster's underlying subjective probability distribution. Making this

substitution, we obtain

$$f_{\mathrm{DR}}(\mathbf{f}) = \sup_{\mathbf{x}\in\mathcal{X},\boldsymbol{\pi}\in\mathbb{R}^p} \left\{ E_{\mathbf{p}'}\left[(\mathbf{X}^\top\boldsymbol{\pi})_\omega\right] :\ \mathbf{X}^\top\boldsymbol{\pi} - u(\mathbf{x}) \leq 0,\ \mathbf{x}\in\mathcal{X} \right\}, \tag{5.25}$$

which is a concave optimization problem and a special case of problem (4.3). The proper linear scoring rule for forecast $\mathbf{f}$ is equivalent to the standard tailored scoring rule for problem (5.25) and has the form

$$S(\omega, \mathbf{f}) = \boldsymbol{\delta}_\omega^\top \mathbf{X}^\top \boldsymbol{\pi}^*,$$

$$(\mathbf{x}^*, \boldsymbol{\pi}^*) \in \operatorname*{arg\,max}_{\mathbf{x}\in\mathcal{X},\boldsymbol{\pi}\in\mathbb{R}^p} \left\{ \mathbf{f}^\top\boldsymbol{\pi} :\ \mathbf{X}^\top\boldsymbol{\pi} - u(\mathbf{x}) \leq 0,\ \mathbf{x}\in\mathcal{X} \right\}.$$

5.4.2 Efficient probability assessments in a decision problem under uncertainty

Consider a risk-averse decision maker with utility function u over wealth. Further, suppose that the decision maker must choose between two competing alternatives, o^1 and o^2, to solve the optimization problem:

$$\max_{o\in\{o^1,o^2\}} E_{\mathbf{p}}[u(\omega, o)]. \tag{5.26}$$

Eliciting $\mathbf{p}$ can be onerous, particularly when the distribution $\mathbf{p}$ required to solve (5.26) is a large joint distribution over multiple dependent uncertainties. To address this, the assessment process can be decomposed into a sequence of simpler queries on marginal probabilities, conditional probabilities, and cumulative probabilities, which can be represented in terms of the query matrix $\mathbf{X}^k$ after the kth assessment.

The distribution $\mathbf{p}$ need not be fully specified: the assessment procedure can be stopped after the Kth assessment when the expected utility of one option dominates the other on the set $\{\mathbf{p} : \mathbf{X}^K\mathbf{p} = \mathbf{f}\}$. The final forecast,

$\mathbf{f} \in \mathcal{F}(\mathbf{X}^K)$, can be scored using a linear scoring rule, which leads to the following ex-ante incentives for the forecaster.

Each row of the matrix $\mathbf{X}^K$ may depend on previous forecasts, as the queries may be dynamically sequenced to efficiently determine the optimal decision [1]. Further, the final dimension of the query matrix, K, will depend on the sequence of forecasted values, as these determine the set $\{\mathbf{p} : \mathbf{X}^K\mathbf{p} = \mathbf{f}\}$. Thus, a linear scoring rule will not be strictly proper, as the forecaster has some control over the matrix $\mathbf{X}^K$, which affects the linear score.

However, practically speaking, the algorithms used to sequence queries, as well as the methods used for determining when the forecasting process can be terminated, are likely to be opaque to the forecaster. In this case, the final query matrix $\mathbf{X}^K$ will appear random to the forecaster. The property (5.5) then guarantees that when a linear scoring rule is used to reward the final sequence of forecasts, the forecaster is a priori incentivized to report honestly.

5.5 Conclusion

This chapter has examined a matrix-based construction method for linear scoring rules based on the Hendrickson and Buehler characterization for traditional scoring rules. These rules can be used to score a sequence of queries and responses provided by a forecaster regarding the underlying probability distribution. Further, the construction method given here can be used to create scoring rules for distributional forecasts as well. A specific application area of interest in decision analysis relates to constructing tailored scoring rules for a distributionally robust utility maximization problem.

Chapter 6

Conclusion and Future Work

This dissertation extends the literature on strictly proper scoring rules. Chapter 3 describes additive and strongly additive scoring rules, and connects the additive properties of scoring rules with a general class of entropy measures. It also provides construction methods for additive scoring rules. Chapter 4 extends the literature connecting weighted scoring families with decision problems under uncertainty. Furthermore, it presents new connections between weighted scoring rules and convex risk measures. Chapter 5 gives a matrix-based construction method for scoring linear forecasts, along with applications.

One challenge when evaluating human forecasters is choosing a scoring rule that is simple enough to use, but well suited for the forecasting problem at hand. For example, the scoring rules described in this dissertation are designed for certain decision problems or applications. A trade-off in designing specialized rules is that these scores are often more complex and may be harder for a forecaster to understand. Thus, future research should develop methodology for applying more complex proper scoring rules in practical situations.

Another avenue for future research could involve development of scoring rules tailored to robust decision problems, like the one presented at the end of Chapter 5. These rules are likely to be difficult to present to a forecaster, but

could be useful in decision-making situations where the forecaster's probabilities are imprecise.

Bibliography

[1] Ali E Abbas. Entropy methods for adaptive utility elicitation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 34(2):169–178, 2004.

[2] Jacob D Abernethy and Rafael M Frongillo. A characterization of scoring rules for linear properties. In *Conference on Learning Theory*, pages 27.1–27.13, 2012.

[3] János Aczél and Zoltán Daróczy. *On measures of information and their characterizations*. Academic Press: New York, 1975.

[4] János Aczél and Johann Pfanzagl. Remarks on the measurement of subjective probability and information. *Metrika*, 11(1):91–105, 1967.

[5] Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

[6] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

[7] Güzin Bayraksan and David K Love. Data-driven stochastic programming using phi-divergences. In *The Operations Research Revolution*, pages 1–19. INFORMS, 2015.

[8] David E Bell and Peter C Fishburn. Strong one-switch utility. *Management Science*, 47(4):601–604, 2001.

[9] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[10] Aharon Ben-Tal and Marc Teboulle. Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Science*, 32(11):1445–1466, 1986.

[11] Aharon Ben-Tal and Marc Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

[12] Riccardo Benedetti. Scoring rules for forecast verification. *Monthly Weather Review*, 138(1):203–211, 2010.

[13] José M Bernardo. Expected information as expected utility. *the Annals of Statistics*, pages 686–690, 1979.

[14] J Eric Bickel. Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis*, 4(2):49–65, 2007.

[15] J Eric Bickel. Scoring rules and decision analysis education. *Decision Analysis*, 7(4):346–357, 2010.

[16] J Eric Bickel, Eric Floehr, and Seong Dae Kim. Comparing NWS PoP forecasts to third-party providers. *Monthly Weather Review*, 139(10):3304–3321, 2011.

[17] J Eric Bickel and Seong Dae Kim. Verification of The Weather Channel probability of precipitation forecasts. *Monthly Weather Review*, 136(12):4867–4881, 2008.

[18] J Eric Bickel and James E Smith. Optimal sequential exploration: A binary learning model. *Decision Analysis*, 3(1):16–32, 2006.

[19] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[20] Jochen Bröcker and Leonard A Smith. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22(2):382–388, 2007.

[21] Arthur Carvalho. An overview of applications of proper scoring rules. *Decision Analysis*, 13(4):223–242, 2016.

[22] Christopher P Chambers, Paul J Healy, and Nicolas S Lambert. Proper scoring rules with general preferences: A dual characterization of optimal reports. *Games and Economic Behavior*, 117:322–341, 2019.

[23] Robert T Clemen and Terence Reilly. Correlations and copulas for decision and risk analysis. *Management Science*, 45(2):208–224, 1999.

[24] Patrick L Combettes. Perspective functions: Properties, constructions, and examples. *Set-Valued and Variational Analysis*, 26(2):247–264, 2016.

[25] Anthony Costa Constantinou and Norman Elliott Fenton. Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, 8(1), 2012.

[26] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, second edition, 2012.

[27] Imre Csiszár. Information type measures of differences of probability distribution and indirect observations. *Studia Math. Hungarica*, 2:299–318, 1967.

[28] Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

[29] A Philip Dawid. Coherent measures of discrepancy, uncertainty and dependence, with applications to bayesian predictive experimental design. *Department of Statistical Science, University College London. http://www. ucl. ac. uk/Stats/research/abs94. html, Tech. Rep*, 139, 1998.

[30] A Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

[31] A Philip Dawid, Steffen Lauritzen, Matthew Parry, et al. Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40(1):593–608, 2012.

[32] Alexander Philip Dawid and Monica Musio. Theory and applications of proper scoring rules. *Metron*, 72(2):169–183, 2014.

[33] Werner Ehm. Unbiased risk estimation and scoring rules. *Comptes Rendus Mathematique*, 349(11-12):699–702, 2011.

[34] Fang Fang, Maxwell B Stinchcombe, and Andrew B Whinston. Proper scoring rules with arbitrary value functions. *Journal of Mathematical Economics*, 46(6):1200–1210, 2010.

[35] Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447, 2002.

[36] Peter GM Forbes. Compatible weighted proper scoring rules. *Biometrika*, 99(4):989–994, 2012.

[37] Marco Frittelli and Emanuela Rosazza Gianin. Putting order in risk measures. *Journal of Banking & Finance*, 26(7):1473–1486, 2002.

[38] Rafael Frongillo and Ian A Kash. Vector-valued property elicitation. In *Conference on Learning Theory*, pages 710–727, 2015.

[39] Sebastian Geissel, Jörn Sass, and Frank Thomas Seifried. Optimal expected utility risk measures. *Statistics & Risk Modeling*, 35(1-2):73–87, 2018.

[40] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[41] Thomas Goll and Ludger Rüschendorf. Minimax and minimal distance martingale measures and their relationship to portfolio optimization. *Finance and Stochastics*, 5(4):557–581, 2001.

[42] Irving John Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114, 1952.

[43] Andrew Grant, David Johnstone, and Oh Kang Kwon. A probability scoring rule for simultaneous events. *Decision Analysis*, To appear, 2019.

[44] Peter D Grünwald, A Philip Dawid, et al. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

[45] Yun He, A Ben Hamza, and Hamid Krim. A generalized divergence measure for robust image registration. *IEEE Transactions on Signal Processing*, 51(5):1211–1220, 2003.

[46] Arlo D Hendrickson and Robert J Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42(6):1916–1921, 1971.

[47] Hajo Holzmann and Bernhard Klar. Weighted scoring rules and hypothesis testing. *arXiv preprint arXiv:1611.07345*, 2016.

[48] Hajo Holzmann, Bernhard Klar, et al. Focusing on regions of interest in forecast evaluation. *The Annals of Applied Statistics*, 11(4):2404–2431, 2017.

[49] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

[50] David Johnstone and Yan-Xia Lin. Fitting probability forecasting models by scoring rules and maximum likelihood. *Journal of Statistical Planning and Inference*, 141(5):1832–1837, 2011.

[51] David J Johnstone. The parimutuel Kelly probability scoring rule. *Decision Analysis*, 4(2):66–75, 2007.

[52] David J Johnstone, Victor Richmond R Jose, and Robert L Winkler. Tailored scoring rules for probabilities. *Decision Analysis*, 8(4):256–268, 2011.

[53] Victor Richmond Jose. A characterization for the spherical scoring rule. *Theory and Decision*, 66(3):263–281, 2009.

[54] Victor Richmond R Jose, Robert F Nau, and Robert L Winkler. Scoring rules, generalized entropy, and utility maximization. *Operations Research*, 56(5):1146–1157, 2008.

[55] Victor Richmond R Jose and Robert L Winkler. Evaluating quantile assessments. *Operations Research*, 57(5):1287–1297, 2009.

[56] Craig W Kirkwood. Approximating risk aversion in decision analysis applications. *Decision Analysis*, 1(1):51–67, 2004.

[57] Nicolas S Lambert, David M Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138. ACM, 2008.

[58] Kenneth C Lichtendahl Jr and Robert L Winkler. Probability elicitation, scoring rules, and competition among forecasters. *Management Science*, 53(11):1745–1755, 2007.

[59] James E Matheson and Robert L Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

[60] John McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9):654–655, 1956.

[61] Robert Nau, Victor Richmond Jose, and Robert Winkler. Scoring rules, entropy, and imprecise probabilities. In *Proc. of ISIPTA*, volume 7, pages 307–315, 2007.

[62] Evgeni Y Ovcharov. Existence and uniqueness of proper scoring rules. *Journal of Machine Learning Research*, 16:2207–2230, 2015.

[63] Evgeni Y Ovcharov. Proper scoring rules and bregman divergence. *Bernoulli*, 24(1):53–79, 2018.

[64] Matthew Parry. Extensive scoring rules. *Electronic Journal of Statistics*, 10(1):1098–1108, 2016.

[65] R. J. Plemmons and R. E. Cline. The generalized inverse of a nonnegative matrix. *Proceedings of the American Mathematical Society*, 31(1):46–50, 1972.

[66] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[67] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

[68] R Tyrrell Rockafellar. Coherent approaches to risk in optimization under uncertainty. *Tutorials in Operations Research*, 3:38–61, 2007.

[69] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

[70] Mark S Roulston and Leonard A Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653–1660, 2002.

[71] Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[72] Reinhard Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1):43–61, 1998.

[73] Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[74] Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[75] Carl-Axel S Staël Von Holstein. Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance*, 8(1):139–158, 1972.

[76] Robert L Winkler, Javier Munoz, José L Cervera, José M Bernardo, Gail Blattenberger, Joseph B Kadane, Dennis V Lindley, Allan H Murphy, Robert M Oliver, and David Ríos-Insua. Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60, 1996.

[77] Robert L Winkler and Allan H Murphy. "Good" probability assessors. *Journal of applied Meteorology*, 7(5):751–758, 1968.