OPERATIONS RESEARCH
Vol. 56, No. 5, September–October 2008, pp. 1146–1157
ISSN 0030-364X | EISSN 1526-5463 | 08 | 5605 | 1146

informs

DOI 10.1287/opre.1070.0498

Scoring Rules, Generalized Entropy, and Utility Maximization

Victor Richmond R. Jose, Robert F. Nau, Robert L. Winkler
The Fuqua School of Business, Duke University, Durham, North Carolina 27708
{vrj@duke.edu, rnau@duke.edu, rwinkler@duke.edu}

Information measures arise in many disciplines, including forecasting (where scoring rules are used to provide incentives for probability estimation), signal processing (where information gain is measured in physical units of relative entropy), decision analysis (where new information can lead to improved decisions), and finance (where investors optimize portfolios based on their private information and risk preferences). In this paper, we generalize the two most commonly used parametric families of scoring rules and demonstrate their relation to well-known generalized entropies and utility functions, shedding new light on the characteristics of alternative scoring rules as well as duality relationships between utility maximization and entropy minimization. In particular, we show that weighted forms of the pseudospherical and power scoring rules correspond exactly to measures of relative entropy (divergence) with convenient properties, and they also correspond exactly to the solutions of expected utility maximization problems in which a risk-averse decision maker whose utility function belongs to the linear-risk-tolerance family interacts with a risk-neutral betting opponent or a complete market for contingent claims in either a one-period or a two-period setting. When the market is incomplete, the corresponding problems of maximizing linear-risk-tolerance utility with the risk-tolerance coefficient β are the duals of the problems of minimizing the pseudospherical or power divergence of order β between the decision maker's subjective probability distribution and the set of risk-neutral distributions that support asset prices.

Subject classifications: decision analysis: theory; probability: entropy; utility/preference: theory; finance: portfolio.
Area of review: Decision Analysis.
History: Received September 2006; revision received March 2007; accepted August 2007. Published online in *Articles in Advance* August 21, 2008.

1. Introduction

Suppose that there is uncertainty concerning which of a set of n mutually exclusive and exhaustive events will occur, and the initial representation of that uncertainty consists of a "baseline" probability distribution $\mathbf{q} = (q_1, \ldots, q_n)$, which could be the subjective prior distribution of an individual or a distribution obtained from a statistical model or from market prices for contingent claims. If new information is subsequently received from an experiment or an expert's forecast, causing the baseline distribution to be revised to another distribution $\mathbf{p}$, how should the quantity or value of the information be measured?

The need for a quantitative measure of information—or more generally, a practical measure of the distance from one distribution $\mathbf{p}$ to some other distribution $\mathbf{q}$—arises in many fields, and the considerable literature on this topic includes (at least) three distinct but intertwined strands: scoring rules, entropy, and decision analysis. *Scoring rules* are reward functions for eliciting and evaluating probability forecasts, and the expected score associated with a forecast can be interpreted as a measure of the value of the forecaster's information. *Entropy* is a measure of the channel capacity required to communicate a stream of signals generated by a stationary process, and *relative entropy* measures the reduction in channel capacity that is possible when new information yields an updated signal distribution. *Decision analysis* provides a general framework for measuring information in terms of gains in expected utility as well for determining how to optimally use information to choose portfolios of financial assets.

These information-theoretic tools have been used for many decades, but new applications and theoretical developments have emerged during the last few years on several fronts, including experimental economics, Bayesian statistics, and financial engineering. The objective of this paper is to add to this recent stream of interdisciplinary literature by broadening the concept of a scoring rule to include a not-necessarily-uniform baseline distribution and to show that this leads immediately to tight connections with some well-known measures of divergence (relative entropy) as well as with models of utility maximization in markets under uncertainty. First, in §2 it is shown that the power and pseudospherical scoring rules (of which the quadratic and spherical rules are special cases) can be normalized so that they are continuous functions of their power parameter (denoted by β) on the entire real line and weighted by a baseline distribution $\mathbf{q}$ to reward updating of probabilities in relative rather than absolute terms. In §3, the forecaster's expected gains under these weighted scoring rules are

shown to correspond exactly to two well-known parametric families of generalized divergence that both reduce to the Kullback-Leibler divergence at $\beta = 1$. Section 4 introduces two canonical decision problems in which an individual with probability distribution $\mathbf{p}$ bets optimally against a nonstrategic, less well-informed opponent (or market) with distribution $\mathbf{q}$. The decision maker's utility function is assumed to belong to the normalized linear-risk-tolerance (LRT) family of utility functions, which includes the familiar exponential, logarithmic, and power functions and is indexed by a single parameter, namely, the risk-tolerance coefficient (also denoted by β). The solution of one canonical decision problem with LRT utility is shown to yield the weighted pseudospherical scoring rule and its associated relative entropy measure, with the same value of β, while the second canonical problem yields the weighted power scoring rule and its associated relative entropy measure. Section 5 generalizes the results of the earlier sections to the situation in which a decision maker with LRT utility optimally invests in an incomplete market for contingent claims, highlighting the duality between expected-utility maximization and relative-entropy minimization. Concluding comments are given in §6.

2. Weighted Scoring Rules

Scoring rules are reward functions for eliciting and evaluating probabilities, and they have played an important role in the foundations of subjective probability theory (de Finetti 1937, 1974; Good 1952; Winkler 1967, 1996; Savage 1971; Lindley 1982) as well as practical applications such as incentive schemes for paying weather forecasters (Brier 1950) and subjects in economic experiments (Selten 1998) and for evaluating the quality of forecasts used in risk analysis (Cooke 1991). Consider an individual (the "forecaster") who is asked to assess a probability distribution over a set of n mutually exclusive and collectively exhaustive events. Let $\mathbf{p}$ denote the forecaster's true distribution, let $\mathbf{r}$ denote her reported distribution (if different from $\mathbf{p}$), and let $\mathbf{e}_i$ denote the probability distribution that assigns probability one to event i and zero to all other events, i.e., the indicator vector for event i. A scoring rule is conventionally expressed as a function $S(\mathbf{r}, \mathbf{p})$, linear in its second argument, such that the score obtained if event i occurs is $S(\mathbf{r}, \mathbf{e}_i)$, and the forecaster's expected score for reporting $\mathbf{r}$ when her true distribution is $\mathbf{p}$ is $S(\mathbf{r}, \mathbf{p}) = \sum_i p_i S(\mathbf{r}, \mathbf{e}_i)$. It is assumed that the forecaster's objective is to maximize her expected score, which means that either she is risk neutral and $S(\mathbf{r}, \mathbf{e}_i)$ is measured in units of money or else she is not risk neutral and $S(\mathbf{r}, \mathbf{e}_i)$ is measured in units of utility.

The scoring rule is defined to be (*strictly*) *proper* if it encourages honest reporting in the sense that $S(\mathbf{p}, \mathbf{p}) \geqslant S(\mathbf{r}, \mathbf{p})$ for every $\mathbf{r}$ and $\mathbf{p}$ (with equality only when $\mathbf{r} = \mathbf{p}$), so that the forecaster whose true distribution is $\mathbf{p}$ maximizes her expected score by truthfully reporting $\mathbf{p}$ rather than some other distribution. The forecaster's optimal expected score that is obtained when her distribution is $\mathbf{p}$ will be denoted by merely suppressing the first argument: $S(\mathbf{p}) \equiv S(\mathbf{p}, \mathbf{p})$. A proper scoring rule is uniquely determined by its optimal-expected-score function, as noted by McCarthy (1956) and further elaborated by Hendrickson and Buehler (1971) and Savage (1971). In particular, if $S(\cdot)$ is a differentiable function, then $S(\cdot, \cdot)$ satisfies

$$S(\mathbf{r}, \mathbf{p}) = S(\mathbf{r}) + \nabla S(\mathbf{r}) \cdot (\mathbf{p} - \mathbf{r}), \tag{1}$$

where $\nabla S(\mathbf{r})$ denotes the gradient of $S(\cdot)$ evaluated at $\mathbf{r}$, and conversely every function S that is (strictly) convex and differentiable uniquely defines a (strictly) proper scoring rule.

The expected-score function of a proper scoring rule is closely related to a measure of distance between probability distributions known as a *Brègman divergence* (Brègman 1967), which generalizes the Kullback-Leibler divergence. Any strictly convex function F defines a Brègman divergence $B_F(\mathbf{p} \,\|\, \mathbf{r})$ as follows:

$$B_F(\mathbf{p} \,\|\, \mathbf{r}) = F(\mathbf{p}) - F(\mathbf{r}) - \nabla F(\mathbf{r}) \cdot (\mathbf{p} - \mathbf{r}). \tag{2}$$

Letting $F(\mathbf{p}) = S(\mathbf{p})$, it follows that for any strictly proper scoring rule, the function $S(\mathbf{p}) - S(\mathbf{r}, \mathbf{p})$, which represents the forecaster's expected loss for reporting $\mathbf{r}$ when the true distribution is $\mathbf{p}$, is a Brègman divergence, and vice versa. A Brègman divergence $B_F(\mathbf{p} \,\|\, \mathbf{r})$ is therefore a decision-theoretic measure of the "information deficit" that is faced by a decision maker who acts on the basis of the distribution $\mathbf{r}$ when the true distribution is $\mathbf{p}$. In this capacity, Brègman divergences (and their corresponding strictly proper scoring rules) provide a potentially rich class of loss functions that can be used for robust Bayesian inference, as discussed by Grünwald and Dawid (2004), Dawid (2007), and Gneiting and Raftery (2007). A problem of this kind can be framed as a game against nature in which nature chooses a true distribution $\mathbf{p}$ from some convex set $\mathcal{P}$, such as the set of distributions satisfying a mean value constraint. The robust Bayes problem for the decision maker is to determine the distribution $\mathbf{r}$ that minimizes her maximum expected loss over all $\mathbf{p} \in \mathcal{P}$, where the expected loss (in our terms) is the negative expected score $-S(\mathbf{r}, \mathbf{p})$. Grünwald and Dawid (2004) show that the optimal expected-loss function, $-S(\mathbf{p})$, is interpretable as a generalized entropy, and minimizing the maximum expected loss is equivalent to maximizing this entropy on the set $\mathcal{P}$. This scoring-rule entropy uniquely determines a corresponding Brègman divergence $B_S(\mathbf{p} \,\|\, \mathbf{r}) \equiv S(\mathbf{p}) - S(\mathbf{r}, \mathbf{p})$, and the distribution $\mathbf{r}$ that minimizes the maximum expected loss on $\mathcal{P}$ is also the distribution that minimizes this divergence with respect to an uninformative "reference" distribution $\mathbf{p}_0$ at which the entropy $-S(\mathbf{p})$ is maximized.

In this paper, we will consider a different kind of game and a correspondingly different decision-theoretic measure

of information, namely, we will suppose that the decision maker is in possession of the true distribution $\mathbf{p}$ (as seen from her own perspective), and a less well-informed opponent has a distribution $\mathbf{q}$ that is known to lie in some set $\mathcal{Q}$ that is disjoint from $\mathbf{p}$, thus providing the decision maker with an opportunity for profitable bets. The "information surplus" enjoyed by this decision maker will be shown to be measured by the minimum of a generalized divergence between $\mathbf{p}$ and all $\mathbf{q} \in \mathcal{Q}'$, and this divergence also corresponds to a strictly proper scoring rule, but it is a different kind of generalized divergence than a Brègman divergence.

The literature of scoring rules has focused mainly on a few strictly proper rules with particularly convenient parametric forms, axiomatic representations, and/or geometrical interpretations, namely, the *quadratic*, *logarithmic*, and *spherical* scoring rules. The quadratic rule (a.k.a. Brier score) is $-(\|\mathbf{e}_i - \mathbf{p}\|_2)^2$. Thus, under the quadratic rule, the forecast $\mathbf{p}$ is treated as an estimate of the indicator vector $\mathbf{e}_i$ of the uncertain event, and the forecaster is ultimately penalized in proportion to the squared Euclidean distance between $\mathbf{p}$ and the realized value of $\mathbf{e}_i$, in the tradition of least-squares estimation. The logarithmic scoring rule is $\ln(p_i)$, whose optimal expected score function is the negative entropy of the forecaster's true distribution, an issue to which we return below. The spherical scoring rule is $p_i/\|\mathbf{p}\|_2$, and it is generated by letting the set of feasible score vectors be the simplest strictly convex object in $\mathbb{R}^n$, namely, the unit sphere. Some additional properties of these rules have been studied recently by Bickel (2007).

The quadratic and spherical rules can be generalized into parametric families by replacing the 2-norm with the vector β-norm, $\|\mathbf{p}\|_\beta \equiv (\sum_{j=1}^n p_j^\beta)^{1/\beta}$. The generalized spherical rule is the *pseudospherical scoring rule*, $p_i/(\|\mathbf{p}\|_\beta)^{\beta-1}$, which was first proposed by Good (1971). The generalized quadratic rule is the *power scoring rule*, $\beta p_i^{\beta-1} - (\beta - 1)(\|\mathbf{p}\|_\beta)^\beta$. Written in this conventional fashion, these families of rules are well defined and proper only for $\beta > 1$ and the corresponding optimal expected-score functions that generate them via McCarthy's formula are $(\|\mathbf{p}\|_\beta)^\beta$ and $\|\mathbf{p}\|_\beta$, respectively. The logarithmic scoring rule is the limiting case of affine transformations of the pseudospherical and power scores as $\beta \longrightarrow 1$, but otherwise the two families do not intersect. A unifying perspective on these two families of rules, which might help to provide some guidance concerning appropriate values of β, has hitherto been lacking. Friedman (1983) attempted to identify scoring rules with metrics (rather than divergences) on the probability space, but most metrics turn out not to have associated scoring rules, and vice versa, as shown by Nau (1985). More recently, Selten (1998) has discussed the implications of different values of β in the power scoring rule, arguing against the logarithmic rule ($\beta = 1$) because of its hypersensitivity to the estimation of small probabilities and in favor of the quadratic rule ($\beta = 2$) because the latter uniquely satisfies a certain axiom of "neutrality," namely, that the expected loss for reporting $\mathbf{r}$ when the true distribution is $\mathbf{p}$ is the same as the expected loss for reporting $\mathbf{p}$ when the true distribution is $\mathbf{r}$, i.e., $S(\mathbf{p}, \mathbf{p}) - S(\mathbf{r}, \mathbf{p}) = S(\mathbf{r}, \mathbf{r}) - S(\mathbf{p}, \mathbf{r})$.

A key property of the aforementioned scoring rules is that they treat events symmetrically in the sense that if $p_i = (>)\ p_j$, then the score in event i is equal to (greater than) the score in event j, regardless of the descriptions of the events, and the forecaster's expected score is smallest when $\mathbf{p}$ is the uniform distribution. Thus, they implicitly reward the forecaster in proportion to some measure of the distance of $\mathbf{p}$ from a uniform distribution. However, in most real (and even hypothetical) applications, the relevant reference point is not a uniform distribution. For example, in weather forecasting, the events that are of interest are often known to have widely varying a priori probabilities, and baseline values for those probabilities, upon which the forecaster is supposed to improve, are obtainable from historical records (Winkler 1994) or alternative forecasting models. In predicting the outcomes of sporting events or movements of financial markets, there are public betting lines or posted prices for contingent claims that implicitly assign probabilities to events. Therefore, we disagree with Selten's (1998) additional axiom that scoring rules should not be "prejudiced" in favor of one hypothesis or another. Rather, we propose that scoring rules *ought* to be generalized so as to reward the forecaster in proportion to some measure of the distance of $\mathbf{p}$ from an appropriate baseline distribution $\mathbf{q}$. Such a scoring rule will be henceforth referred to as a *weighted* scoring rule; it will be expressed as a function of three arguments, $S(\mathbf{r}, \mathbf{p} \,\|\, \mathbf{q})$, and its associated optimal expected score will be expressed as a function of two arguments, $S(\mathbf{p} \,\|\, \mathbf{q})$.

There are various functional forms through which the dependence of the score on the baseline distribution could be modeled, and the one we find most compelling (for both practical and theoretical reasons) is that for fixed $\mathbf{p}$ and $\mathbf{q}$, the score in event i should depend on the ratio p_i/q_i; if $p_i/q_i = (>)\ p_j/q_j$, then the score in event i should be equal to (greater than) the score in event j. One simple rationale for this desideratum is that when bets may be placed on outcomes of events, *relative* rather than absolute differences in probabilities are what matter. Another rationale can be illustrated by a simple example: suppose that the state space consists of four states formed by the Cartesian product of two binary events E and F, and suppose it happens that the forecaster and client both agree on the probability of F and also agree that E and F are statistically independent. Then, it seems reasonable that the forecaster's payment should depend only on the outcome of E, not F, and this requires the payoff in each state to depend only on the ratio of the two agents' probabilities for that state.

The measurement of distance between two probability distributions in terms of ratios has a long history in statistics and information theory. It was noted above that under a strictly proper scoring rule, the forecaster's expected *loss*

for reporting a distribution $\mathbf{r}$ that is other than her true distribution $\mathbf{p}$ is a particular kind of divergence between $\mathbf{r}$ and $\mathbf{p}$, namely, a Brègman divergence. Under a weighted strictly proper scoring rule that bases the score on the ratio p_i/q_i, the forecaster's expected *gain* for possessing a distribution $\mathbf{p}$ that differs from $\mathbf{q}$ is a second kind of divergence, which is not a Brègman divergence. Rather, it turns out to be a special case (or a monotonic transformation) of another kind of generalized divergence known as an f-divergence (Csiszár 1967). If f is a strictly convex function, the corresponding f-divergence is defined as

$$D_f(\mathbf{p} \,\|\, \mathbf{q}) = E_{\mathbf{p}}[f(\mathbf{p}/\mathbf{q})]. \tag{3}$$

Divergences of this general form have been widely used in statistics for many years as (seemingly) utility-free measures of the value of the information; e.g., Goel (1983) uses f-divergence to define a "conditional amount of sample information" for measuring prior-to-posterior information gains in Bayesian hierarchical models. More recently, it has been recognized that f-divergences are interpretable as measures of expected utility gains that are available to decision makers who have opportunities to bet against less well-informed opponents or to invest in financial markets, as will be more fully discussed in later sections of this paper.

When the ratio p_i/q_i is substituted for p_i in the pseudospherical and power scoring rules, and they are affinely transformed to yield scores of zero when $\mathbf{p} = \mathbf{q}$, we obtain the *weighted power score*, denoted by $S_\beta^{\mathbf{P}}$, and the *weighted pseudospherical score*, denoted by $S_\beta^{\mathbf{S}}$, with the following parametric forms:

$$\begin{aligned} S_\beta^{\mathbf{P}}(\mathbf{p}, \mathbf{e}_i \,\|\, \mathbf{q}) &\equiv \frac{(p_i/q_i)^{\beta-1} - 1}{\beta - 1} - \frac{E_{\mathbf{q}}[(\mathbf{p}/\mathbf{q})^\beta] - 1}{\beta} \\ &= \frac{(p_i/q_i)^{\beta-1} - 1}{\beta - 1} - \frac{E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}] - 1}{\beta}, \end{aligned} \tag{4}$$

$$\begin{aligned} S_\beta^{\mathbf{S}}(\mathbf{p}, \mathbf{e}_i \,\|\, \mathbf{q}) &\equiv \frac{1}{\beta - 1}\left(\left(\frac{p_i/q_i}{(E_{\mathbf{q}}[(\mathbf{p}/\mathbf{q})^\beta])^{1/\beta}}\right)^{\beta-1} - 1\right) \\ &= \frac{1}{\beta - 1}\left(\left(\frac{p_i/q_i}{(E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}])^{1/\beta}}\right)^{\beta-1} - 1\right). \end{aligned} \tag{5}$$

The equivalence of the two forms of each rule follows from the identity $E_{\mathbf{q}}[(\mathbf{p}/\mathbf{q})^\beta] = E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]$. Table 1 highlights some important special cases.

Note that for any fixed values of $\mathbf{p}$, $\mathbf{q}$, and β, the pseudospherical score vector $(S_\beta^{\mathbf{S}}(\mathbf{p}, \mathbf{e}_1 \,\|\, \mathbf{q}), \ldots, S_\beta^{\mathbf{S}}(\mathbf{p}, \mathbf{e}_n \,\|\, \mathbf{q}))$ is a positive affine transformation of the power score vector $(S_\beta^{\mathbf{P}}(\mathbf{p}, \mathbf{e}_1 \,\|\, \mathbf{q}), \ldots, S_\beta^{\mathbf{P}}(\mathbf{p}, \mathbf{e}_n \,\|\, \mathbf{q}))$ because both vectors are affine transformations of $(\mathbf{p}/\mathbf{q})^{\beta-1}$, although the origins and scale factors of the transformations vary with $\mathbf{p}$, $\mathbf{q}$, and β. Thus, although the two rules yield different expected payoffs as a function of $\mathbf{p}$ (for the same $\mathbf{q}$ and β), and they create different incentives for information gathering and different penalties for dishonest reporting, they nevertheless present the same relative risk profile to a truthful forecaster whose $\mathbf{p}$ is already fixed. At $\beta = 1$, both rules converge to the weighted logarithmic score $\ln(p_i/q_i)$. At $\beta = 2$, weighted forms of the quadratic and spherical scoring rules are obtained. The cases $\beta = 0$ and $\beta = \frac{1}{2}$ have not received much (if any) attention in the antecedent literature, but it will be shown later that $\beta = 0$ corresponds to a decision model involving exponential utility, which is the utility function most commonly used in applied decision analysis; while $\beta = \frac{1}{2}$ arises from a decision model involving reciprocal utility, which has some appealing symmetry properties and is closely related to the Hellinger distance between $\mathbf{p}$ and $\mathbf{q}$. These special cases will be further explored in the next two sections.

Table 1. Weighted power and pseudospherical scores.

	$S_\beta^{\mathbf{P}}(\mathbf{p}, \mathbf{e}_i \,\|\, \mathbf{q})$	$S_\beta^{\mathbf{S}}(\mathbf{p}, \mathbf{e}_i \,\|\, \mathbf{q})$
$\beta = -1$	$-\frac{1}{2}(1 + (q_i/p_i)^2) + E_{\mathbf{q}}[(\mathbf{q}/\mathbf{p})]$	$\frac{1}{2}(1 - ((q_i/p_i)/E_{\mathbf{q}}[(\mathbf{q}/\mathbf{p})])^2)$
$\beta = 0$	$1 - (q_i/p_i) + E_{\mathbf{q}}[\ln(\mathbf{q}/\mathbf{p})]$	$1 - (q_i/p_i)\exp(-E_{\mathbf{q}}[\ln(\mathbf{q}/\mathbf{p})])$
$\beta = \frac{1}{2}$	$2(2 - \sqrt{q_i/p_i} - E_{\mathbf{p}}[\sqrt{\mathbf{q}/\mathbf{p}}])$	$2(1 - \sqrt{q_i/p_i}E_{\mathbf{p}}[\sqrt{\mathbf{q}/\mathbf{p}}])$
$\beta = 1$	$\ln(p_i/q_i)$	$\ln(p_i/q_i)$
$\beta = 2$	$((p_i/q_i) - 1) - \frac{1}{2}(E_{\mathbf{p}}[\mathbf{p}/\mathbf{q}] - 1)$	$((p_i/q_i)/\sqrt{E_{\mathbf{p}}[\mathbf{p}/\mathbf{q}]}) - 1$

Figures 1 and 2 illustrate that even a uniform forecast can be informative when measured against a nonuniform baseline distribution. The figures show the scores for the three-event case when the forecast is $\mathbf{p} = (1/3, 1/3, 1/3)$ and the baseline is $\mathbf{q} = (1/12, 1/3, 7/12)$, as β varies over the range from -1 to $+2$. The vector of relative probabilities that determines the profile of scores is $\mathbf{p}/\mathbf{q} = (4, 1, 4/7)$. The two rules are qualitatively similar in that for $\beta \approx -1$, their scores mainly distinguish the lowest relative-probability event $(p_3/q_3 = 4/7)$ from all the others; and for $\beta \approx 2$, they mainly distinguish the highest relative-probability event $(p_1/q_1 = 4)$; while for β in the unit interval, they discriminate more finely among events with a wide range of relative probabilities.

Figure 1. Weighted power score vs. β.

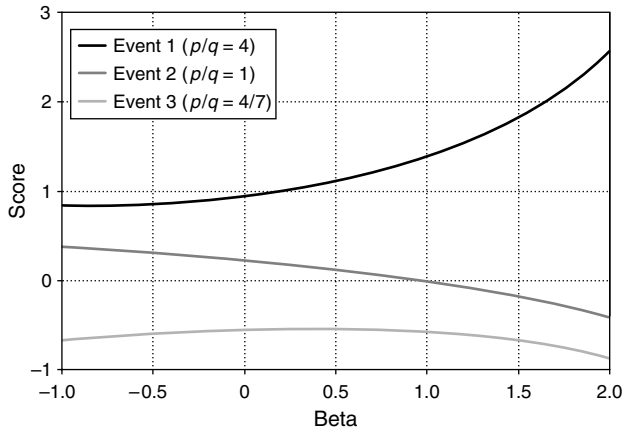
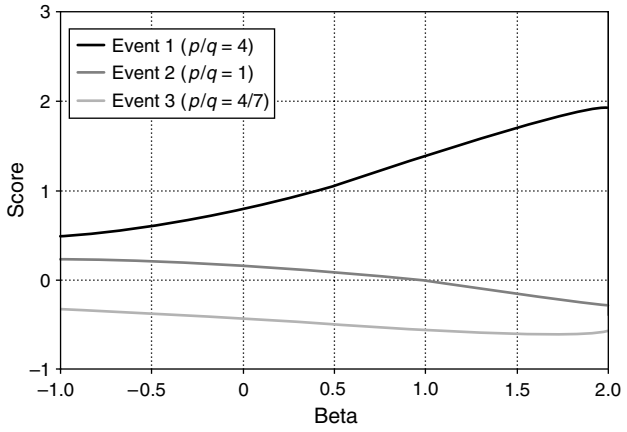


Figure 2. Weighted pseudospherical score vs. β.



The corresponding optimal expected-score functions for the two families of weighted scoring rules can also be expressed in terms of either an expectation over the baseline (prior) distribution $\mathbf{q}$ or over the true (posterior) distribution $\mathbf{p}$, and the optimal expected score under one rule is a monotonic transformation of the other:

$$S_\beta^{\mathrm{P}}(\mathbf{p}\,\|\,\mathbf{q}) = \frac{E_{\mathbf{q}}[(\mathbf{p}/\mathbf{q})^\beta]-1}{\beta(\beta-1)} = \frac{E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]-1}{\beta(\beta-1)}, \tag{6}$$

$$S_\beta^{\mathrm{S}}(\mathbf{p}\,\|\,\mathbf{q}) = \frac{(E_{\mathbf{q}}[(\mathbf{p}/\mathbf{q})^\beta])^{1/\beta}-1}{\beta-1} = \frac{(E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}])^{1/\beta}-1}{\beta-1}. \tag{7}$$

3. Generalized Measures of Entropy and Divergence

A second strand of literature considers information value from the perspective of an engineer who designs a communication channel to transmit the observations of a sequence of independent, identically distributed events. Shannon (1948) proved that under the most efficient coding scheme, the average number of bits needed to report the occurrence of an event whose relative frequency is p is proportional to $\ln(1/p) = -\ln(p)$, so the expected number of bits per event (which determines the required capacity of the channel) to encode events drawn from a distribution $\mathbf{p}$ is proportional to $H(\mathbf{p}) \equiv -\sum_i p_i \ln(p_i)$. This quantity is known as the *entropy* of the distribution $\mathbf{p}$ because up to a multiplicative constant (namely, Boltzmann's constant), it coincides exactly with the definition of the Gibbs entropy of a physical system whose distribution of internal states is $\mathbf{p}$, which in turn is the microscopic interpretation of the macroscopic concept of entropy from classical thermodynamics. Now suppose that an engineer who had optimized the encoding scheme on the assumption that the distribution was $\mathbf{q}$ now learns that it is $\mathbf{p}$ instead. A practical measure of the amount of information gained in updating $\mathbf{q}$ to $\mathbf{p}$ is the reduction in the expected number of bits needed to encode an event using the distribution now believed to be correct, which is known as the *Kullback-Leibler* (*KL*) *divergence* of $\mathbf{p}$ with respect to $\mathbf{q}$:

$$D_{KL}(\mathbf{p}\,\|\,\mathbf{q}) \equiv \sum_i p_i(\ln(1/q_i)-\ln(1/p_i)) = \mathbf{E}_{\mathbf{p}}[\ln(\mathbf{p}/\mathbf{q})]. \tag{8}$$

The KL divergence is measured in physical units, and it is seemingly utility free insofar as it does not involve the risk preferences of any individual.

The KL divergence has several very convenient and appealing properties that are often cited as reasons for adopting it as a universal measure of information gain. First, it is naturally *additive* with respect to independent experiments. Suppose that A and B are statistically independent partitions of the state space whose prior marginal probability distributions are $\mathbf{q}_A$ and $\mathbf{q}_B$, so their prior joint distribution is $\mathbf{q}_A \times \mathbf{q}_B$. Now suppose that independent experiments are performed, which result in the updating of $\mathbf{q}_A$ and $\mathbf{q}_B$ to $\mathbf{p}_A$ and $\mathbf{p}_B$, respectively, so that the posterior joint distribution is $\mathbf{p}_A \times \mathbf{p}_B$. Then, the total information gain of the two experiments is the sum of their separate KL divergences:

$$D_{KL}(\mathbf{p}_A \times \mathbf{p}_B \,\|\, \mathbf{q}_A \times \mathbf{q}_B) = D_{KL}(\mathbf{p}_A \,\|\, \mathbf{q}_A) + D_{KL}(\mathbf{p}_B \,\|\, \mathbf{q}_B). \tag{9}$$

Second, and even stronger, the KL divergence has the property of *recursivity* with respect to the splitting of events. Suppose that information is transmitted in a two-step process, in which two out of the n possible events—say, events 1 and 2—are not distinguished on the first step. If the realized event is neither 1 or 2, the process stops there, but otherwise a second signal is sent to report which of those two has occurred. The probabilities of events 1 and 2 are aggregated in the first step, so the information gain on that step is $D_{KL}(p_1+p_2, p_3, \ldots, p_n \,\|\, q_1+q_2, q_3, \ldots, q_n)$. On the second step, which occurs with probability (p_1+p_2), the additional gain is

$$D_{KL}\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2} \,\middle\|\, \frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2}\right).$$

The recursivity property of the KL divergence requires the expected total information gain of the two-step process to be the same as that of a one-step process:

$$\begin{aligned} &D_{KL}(\mathbf{p}\,\|\,\mathbf{q}) \\ &= D_{KL}(p_1+p_2, p_3, \ldots, p_n \,\|\, q_1+q_2, q_3, \ldots, q_n) \\ &\quad + (p_1+p_2) D_{KL}\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2} \,\middle\|\, \frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2}\right). \end{aligned} \tag{10}$$

The KL divergence is the only information measure that satisfies both additivity and recursivity; hence, it is the measure that is naturally obtained if those properties are embraced as axioms that an information measure should satisfy. However, it has been discovered that weakenings

of these axioms lead to several other interesting parametric families of generalized divergence with their own merits and their own applications. Havrda and Chavrát (1967) defined a quantity they called the *directed divergence of order β between* $\mathbf{p}$ *and* $\mathbf{q}$, and variants of this divergence, which are equivalent up to a scale factor, were discussed by Rathie and Kannappan (1972), Cressie and Read (1984), and Haussler and Opper (1997). Cressie and Read referred to this quantity as the *power divergence*, and that is the term adopted here for the following reason: the power divergence (as originally introduced by Havrda and Chavrát 1967) is defined for all $\beta \in \mathbb{R}$ by

$$D_\beta^{\mathbf{P}}(\mathbf{p} \,\|\, \mathbf{q}) \equiv \frac{E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}] - 1}{\beta(\beta-1)}, \tag{11}$$

which is identical to $S_\beta^{\mathbf{P}}(\mathbf{p} \,\|\, \mathbf{q})$ with the same power β, and it is an f-divergence with $f(x) = (x^{\beta-1} - 1)/(\beta(\beta-1))$. Hence, the power divergence is the information measure that implicitly underlies the weighted power scoring rule, and it has the same interesting special cases, namely, $\beta = -1, 0, \frac{1}{2}, 1$, and 2. In particular, at $\beta = 1$, the power divergence between $\mathbf{p}$ and $\mathbf{q}$ is equal to the KL divergence, while at $\beta = \frac{1}{2}$, the power divergence is

$$D_{1/2}^{\mathbf{P}}(\mathbf{p} \,\|\, \mathbf{q}) = 4\left(1 - \sum_{j=1}^{n} \sqrt{p_j q_j}\right), \tag{12}$$

which is proportional to the *squared Hellinger distance* between $\mathbf{p}$ and $\mathbf{q}$, as noted by Haussler and Opper (1997). The Hellinger distance $D_H(\mathbf{p}\|\mathbf{q})$ is widely used in statistics and is defined by

$$D_H(\mathbf{p} \,\|\, \mathbf{q}) \equiv \left(\sum_{j=1}^{n} (\sqrt{p_j} - \sqrt{q_j})^2\right)^{1/2}, \tag{13}$$

whence

$$D_{1/2}^{\mathbf{P}}(\mathbf{p} \,\|\, \mathbf{q}) = 2D_H(\mathbf{p} \,\|\, \mathbf{q})^2. \tag{14}$$

At $\beta = 2$, the power divergence reduces to (a multiple of) another well-known divergence, the *Chi-square divergence* (Pearson 1900):

$$D_2^{\mathbf{P}}(\mathbf{p} \,\|\, \mathbf{q}) = \tfrac{1}{2}(E_{\mathbf{p}}[\mathbf{p}/\mathbf{q}] - 1) = \tfrac{1}{2}\chi^2(\mathbf{p} \,\|\, \mathbf{q}), \tag{15}$$

while at $\beta = -1$, it is the reverse Chi-square divergence $\frac{1}{2}\chi^2(\mathbf{q} \,\|\, \mathbf{p})$.

Unlike the KL divergence, the power divergence is generally neither additive nor recursive, but it satisfies two slightly weaker properties for all values of β. First, it satisfies the following *pseudoadditivity property* with respect to independent partitions A and B:

$$\begin{aligned} D_\beta^{\mathbf{P}}(\mathbf{p}_A \times \mathbf{p}_B \,\|\, \mathbf{q}_A \times \mathbf{q}_B) &= D_\beta^{\mathbf{P}}(\mathbf{p}_A \,\|\, \mathbf{q}_A) + D_\beta^{\mathbf{P}}(\mathbf{p}_B \,\|\, \mathbf{q}_B) \\ &\quad + \beta(\beta-1) D_\beta^{\mathbf{P}}(\mathbf{p}_A \,\|\, \mathbf{q}_A) D_\beta^{\mathbf{P}}(\mathbf{p}_B \,\|\, \mathbf{q}_B). \end{aligned} \tag{16}$$

Second, it satisfies the following *pseudorecursivity property* with respect to the splitting of events (Rathie and Kannappan 1972, Cressie and Read 1984):

$$\begin{aligned} D_\beta^{\mathbf{P}}(\mathbf{p} \,\|\, \mathbf{q}) &= D_\beta^{\mathbf{P}}(p_1+p_2, p_3, \ldots, p_n \,\|\, q_1+q_2, q_3, \ldots, q_n) \\ &\quad + (p_1+p_2)\left(\frac{p_1+p_2}{q_1+q_2}\right)^{\beta-1} \\ &\quad \times D_\beta^{\mathbf{P}}\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2} \,\middle\|\, \frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2}\right). \end{aligned} \tag{17}$$

Pseudoadditivity reduces to additivity in both of the special cases $\beta = 0$ and $\beta = 1$ (thus, the "zero power" divergence $D_0^{\mathbf{P}}(\mathbf{p} \,\|\, \mathbf{q})$ is additive, along with the KL divergence), while pseudorecursitivity reduces to recursivity only in the special case $\beta = 1$ (the KL divergence). Also note that for $\beta \in (0, 1)$, the power divergence is *subadditive*, i.e., $D_\beta^{\mathbf{P}}(\mathbf{p}_A \times \mathbf{p}_B \,\|\, \mathbf{q}_A \times \mathbf{q}_B) \leqslant D_\beta^{\mathbf{P}}(\mathbf{p}_A \,\|\, \mathbf{q}_A) + D_\beta^{\mathbf{P}}(\mathbf{p}_B \,\|\, \mathbf{q}_B)$, while for $\beta < 0$ or $\beta > 1$, it is *superadditive*, i.e., $D_\beta^{\mathbf{P}}(\mathbf{p}_A \times \mathbf{p}_B \,\|\, \mathbf{q}_A \times \mathbf{q}_B) \geqslant D_\beta^{\mathbf{P}}(\mathbf{p}_A \,\|\, \mathbf{q}_A) + D_\beta^{\mathbf{P}}(\mathbf{p}_B \,\|\, \mathbf{q}_B)$.

A different form of generalized entropy was introduced by Arimoto (1971) and further elaborated by Sharma and Mittal (1975), Boekee and Van der Lubbe (1980), and Lavenda and Dunning-Davies (2003). Arimoto's generalized entropy of order β is defined for $\beta > 0$ as follows:

$$\frac{\beta}{\beta-1}(E_{\mathbf{p}}[\mathbf{p}^{\beta-1}]^{1/\beta} - 1). \tag{18}$$

(Here, β corresponds to the term $1/\beta$ in Arimoto's 1971 original presentation and to the term R in Boekee and Van der Lubbe's 1980 presentation.) The factor of β in the numerator plays no essential role when β is restricted to be positive, and without it the measure is actually valid for all real β and closely related to the pseudospherical scoring rule.

The corresponding relative entropy measure, which we henceforth call the *pseudospherical divergence of order β* between $\mathbf{p}$ and $\mathbf{q}$, is obtained by introducing a baseline distribution $\mathbf{q}$ and dividing out the problematic factor of β,

$$D_\beta^{\mathbf{S}}(\mathbf{p} \,\|\, \mathbf{q}) \equiv \frac{(E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}])^{1/\beta} - 1}{\beta-1}. \tag{19}$$

This is seen to be identical to the weighted pseudospherical optimal expected-score function $S_\beta^{\mathbf{S}}(\mathbf{p} \,\|\, \mathbf{q})$, and it is a nonlinear transformation of the power divergence. Hence, it can also be expressed as a function of other well-known divergences for special cases of β, as summarized in Table 2 (which highlights the symmetry of the power divergence around $\beta = \frac{1}{2}$).

Like the power divergence, the pseudospherical divergence satisfies a pseudoadditivity property:

$$\begin{aligned} D_\beta^{\mathbf{S}}(\mathbf{p}_A \times \mathbf{p}_B \,\|\, \mathbf{q}_A \times \mathbf{q}_B) &= D_\beta^{\mathbf{S}}(\mathbf{p}_A \,\|\, \mathbf{q}_A) + D_\beta^{\mathbf{S}}(\mathbf{p}_B \,\|\, \mathbf{q}_B) \\ &\quad + (\beta-1) D_\beta^{\mathbf{S}}(\mathbf{p}_A \,\|\, \mathbf{q}_A) D_\beta^{\mathbf{S}}(\mathbf{p}_B \,\|\, \mathbf{q}_B). \end{aligned} \tag{20}$$

Table 2. Weighted expected scores and corresponding generalized divergences.

	$S^{\mathbf{P}}_\beta(\mathbf{p}\,\|\,\mathbf{q}) = D^{\mathbf{P}}_\beta(\mathbf{p}\,\|\,\mathbf{q})$	$S^{\mathbf{S}}_\beta(\mathbf{p}\,\|\,\mathbf{q}) = D^{\mathbf{S}}_\beta(\mathbf{p}\,\|\,\mathbf{q})$
$\beta=-1$	$\frac{1}{2}\chi^2(\mathbf{q}\,\|\,\mathbf{p})$	$\frac{1}{2}(1-(\chi^2(\mathbf{q}\,\|\,\mathbf{p})+1)^{-1})$
$\beta=0$	$D_{KL}(\mathbf{q}\,\|\,\mathbf{p})$	$1-\exp(-D_{KL}(\mathbf{q}\,\|\,\mathbf{p}))$
$\beta=\frac{1}{2}$	$2D_H(\mathbf{p}\,\|\,\mathbf{q})^2 = 2D_H(\mathbf{q}\,\|\,\mathbf{p})^2$	$2(1-(1-\frac{1}{2}D_H(\mathbf{p}\,\|\,\mathbf{q})^2)^2)$
$\beta=1$	$D_{KL}(\mathbf{p}\,\|\,\mathbf{q})$	$D_{KL}(\mathbf{p}\,\|\,\mathbf{q})$
$\beta=2$	$\frac{1}{2}\chi^2(\mathbf{p}\,\|\,\mathbf{q})$	$\sqrt{\chi^2(\mathbf{p}\,\|\,\mathbf{q})+1}-1$

The coefficient of the cross-term in this case is $\beta-1$, not $\beta(\beta-1)$, and hence $D^{\mathbf{S}}_\beta(\mathbf{p}\,\|\,\mathbf{q})$ is subadditive for $\beta<1$ and superadditive for $\beta>1$. However, the pseudospherical divergence is generally not pseudorecursive, and it is not an f-divergence, although it is monotonically related to one.

4. Decision Models and Information Measures

We now consider two generic optimization problems in which a risk-averse decision maker with probability distribution $\mathbf{p}$ bets against a nonstrategic risk-neutral opponent with distribution $\mathbf{q}$, or equivalently, invests in a complete market for contingent claims whose supporting risk-neutral distribution is $\mathbf{q}$. (The assumption of a risk-neutral opponent is without loss of generality: as long as one party is more risk averse than the other is risk seeking, the decision problems can be converted into this form.) Let $\mathbf{x}\in\mathbb{R}^n$ denote the vector of monetary payoffs to the decision maker, and let $u(\mathbf{x})\equiv(u(x_1),\ldots,u(x_n))$ denote the vector of utilities that the function u yields when applied to $\mathbf{x}$, and similarly for other functions with vector arguments.

In the first problem (Problem **S**), there is a single time period in which consumption occurs, the decision maker has a single-attribute vNM utility function $u(x)$, and her objective is to find the payoff vector $\mathbf{x}$ that maximizes her subjective expected utility subject to the self-financing constraint $E_{\mathbf{q}}[\mathbf{x}]\leqslant 0$. The decision maker's optimal expected utility, denoted $U^{\mathbf{S}}(\mathbf{p}\,\|\,\mathbf{q})$, is determined by

$$\textbf{Problem S:}\quad U^{\mathbf{S}}(\mathbf{p}\,\|\,\mathbf{q})\equiv\max_{\mathbf{x}\in\mathbf{R}^n}\ E_{\mathbf{p}}[u(\mathbf{x})]\quad\text{s.t. } E_{\mathbf{q}}[\mathbf{x}]\leqslant 0. \tag{21}$$

In the second problem (Problem **P**), there are two periods in which consumption occurs, and the decision maker with probability distribution $\mathbf{p}$ has a quasilinear utility function $u(a,b)=a+u(b)$, where a is money consumed at time 0 and b is money consumed at time 1. The decision maker's objective is to choose a vector $\mathbf{x}$ of time-1 payoffs to be purchased from time-0 funds at market prices to maximize the expected utility of consumption in both periods. The time-0 cost of purchasing $\mathbf{x}$ is $E_{\mathbf{q}}[\mathbf{x}]$, so the optimal expected utility, denoted $U^{\mathbf{P}}(\mathbf{p}\,\|\,\mathbf{q})$, is the solution of

$$\textbf{Problem P:}\quad U^{\mathbf{P}}(\mathbf{p}\,\|\,\mathbf{q})\equiv\max_{\mathbf{x}\in\mathbf{R}^n}E_{\mathbf{p}}[u(\mathbf{x})]-E_{\mathbf{q}}[\mathbf{x}]. \tag{22}$$

The preceding optimization problems will next be given a more specific form by letting u be a utility function from the general exponential/logarithmic/power family, which will be parameterized as

$$u_\beta(x)\equiv\frac{1}{\beta-1}((1+\beta x)^{(\beta-1)/\beta}-1)\quad\text{if }\beta x>-1,\qquad u_\beta(x)\equiv-\infty\quad\text{otherwise,} \tag{23}$$

for all $\beta\in\mathbb{R}$, and the corresponding optimal expected utilities in Problems **S** and **P** will be denoted $U^{\mathbf{S}}_\beta(\mathbf{p}\,\|\,\mathbf{q})$ and $U^{\mathbf{P}}_\beta(\mathbf{p}\,\|\,\mathbf{q})$, respectively. This parameterization has two key properties. First, $u_\beta(0)=0$ and $u'_\beta(0)=1$, so for every β the graph of u_β passes through the origin and has the same slope of one there, and the marginal rate of substitution between time-0 consumption and time-1 consumption is unity at $x=0$ for the decision maker in Problem **P**. Second, the corresponding *risk-tolerance function* $\tau_\beta(x)$, which is the reciprocal of the Pratt-Arrow risk-aversion measure, is a linear function of wealth with slope equal to β and intercept equal to one: $\tau_\beta(x)\equiv-u'_\beta(x)/u''_\beta(x)=1+\beta x$. Thus, risk tolerance as well as marginal utility is normalized to a value of one at $x=0$. The LRT utility functions are also known as hyperbolic absolute risk-aversion (HARA) utility functions in the literature of financial economics, although parameterizing them in terms of their risk-tolerance coefficients rather than their risk-aversion coefficients is more useful for our purposes. Some important special cases of $u_\beta(x)$ are given in Table 3.

The utility functions $\{u_\beta\}$ also exhibit a convenient symmetry around $\beta=\frac{1}{2}$, namely, that $u_{1-\beta}(x)=-u_\beta(-x)$, or equivalently, $u_\beta(-u_{1-\beta}(-x))=x$. In other words, the graph of $u_{1-\beta}(x)$ is obtained from the graph of $u_\beta(x)$ by reflecting it around the line $y=-x$. The power (exponent) in u_β is the term $(\beta-1)/\beta$, which has the property that $((\beta-1)/\beta)^{-1}=((1-\beta)-1)/(1-\beta)$, so that swapping β for $1-\beta$ results in another power utility function whose power is the reciprocal of the original. Thus, under this parameterization, the reciprocal utility function ($\beta=\frac{1}{2}$) is its own reflection around the line $y=-x$, the exponential and logarithmic utility functions ($\beta=0$ and $\beta=1$) are reflections of each other, and the power utility function with exponent δ is the reflection of the power utility function with exponent $1/\delta$ for any positive or negative δ other than zero or one.

Table 3. Examples of normalized linear-risk-tolerance utility functions.

$\beta=-1$	Quadratic utility	$u_{-1}(x)=-\frac{1}{2}((1-x)^2-1)$
$\beta=0$	Exponential utility	$u_0(x)=1-\exp(-x)$
$\beta=\frac{1}{2}$	Reciprocal utility	$u_{1/2}(x)=2\left(1-\frac{1}{1+x/2}\right)$
$\beta=1$	Logarithmic utility	$u_1(x)=\ln(1+x)$
$\beta=2$	Square-root utility	$u_2(x)=\sqrt{1+2x}-1$

Henceforth, let $\mathbf{x}^{\mathbf{S}}_{\beta}(\mathbf{p}\,\|\,\mathbf{q})$ and $\mathbf{x}^{\mathbf{P}}_{\beta}(\mathbf{p}\,\|\,\mathbf{q})$ denote the solutions of Problems **S** and **P**, with ith elements $x^{\mathbf{S}}_{\beta,i}(\mathbf{p}\,\|\,\mathbf{q})$ and $x^{\mathbf{P}}_{\beta,i}(\mathbf{p}\,\|\,\mathbf{q})$, respectively. Our first main result is that the utility gains to the decision maker under Problems **S** and **P** are precisely the pseudospherical and power scores for the same **p**, **q**, and β, and the weighted power expected score is always greater than or equal to the corresponding weighted pseudospherical expected score.

THEOREM 1. (a) $S^{\mathbf{S}}_{\beta}(\mathbf{p}, \mathbf{e}_i\,\|\,\mathbf{q}) = u_{\beta}(x^{\mathbf{S}}_{\beta,i}(\mathbf{p}\,\|\,\mathbf{q}))$ *and* $S^{\mathbf{S}}_{\beta}(\mathbf{p}\,\|\,\mathbf{q}) = U^{\mathbf{S}}_{\beta}(\mathbf{p}\,\|\,\mathbf{q})$.

(b) $S^{\mathbf{P}}_{\beta}(\mathbf{p}, \mathbf{e}_i\,\|\,\mathbf{q}) = u_{\beta}(x^{\mathbf{P}}_{\beta,i}(\mathbf{p}\,\|\,\mathbf{q})) - E_{\mathbf{q}}[\mathbf{x}^{\mathbf{P}}_{\beta}(\mathbf{p}\,\|\,\mathbf{q})]$ *and* $S^{\mathbf{P}}_{\beta}(\mathbf{p}\,\|\,\mathbf{q}) = U^{\mathbf{P}}_{\beta}(\mathbf{p}\,\|\,\mathbf{q})$.

(c) $S^{\mathbf{P}}_{\beta}(\mathbf{p}\|\mathbf{q}) \geqslant S^{\mathbf{S}}_{\beta}(\mathbf{p}\|\mathbf{q})$ *for all* **p**, **q**, *and* β.

PROOF. For part (a), note that with utility function u_{β} the formulation of Problem **S** becomes

$$\begin{aligned} \max_{\mathbf{x}\in\mathbf{R}^n} \quad & \frac{1}{\beta-1}\sum_{j=1}^{n} p_j((1+\beta x_j)^{(\beta-1)/\beta}-1) \\ \text{s.t.} \quad & \sum_{j=1}^{n} q_j x_j \leqslant 0. \end{aligned} \tag{24}$$

Introducing a Lagrange multiplier λ, the constrained maximization can be rewritten in unconstrained form:

$$\min_{\lambda\in\mathbb{R}^+}\max_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{\beta-1}\sum_{j=1}^{n} p_j((1+\beta x_j)^{(\beta-1)/\beta}-1) - \lambda\sum_{j=1}^{n} q_j x_j. \tag{25}$$

One first-order condition is $p_i(1+\beta x_i)^{-1/\beta} = \lambda q_i$ for all i, whence $1+\beta x_i = (p_i/\lambda q_i)^{\beta}$, or equivalently, $x_i = \frac{1}{\beta}((p_i/\lambda q_i)^{\beta}-1)$. The other first-order condition is $\sum_{j=1}^{n} q_j x_j = 0$, which yields $\lambda^* \equiv (\sum_{j=1}^{n} q_j(p_j/q_j)^{\beta})^{1/\beta}$ as the optimal value of λ (which is the marginal utility of wealth at the optimum), so the optimal monetary payoff in event i is

$$x^{\mathbf{S}}_{\beta,i}(\mathbf{p}\,\|\,\mathbf{q}) \equiv \frac{1}{\beta}\left(\left(\frac{p_i/q_i}{(\sum_{j=1}^{n} q_j(p_j/q_j)^{\beta})^{1/\beta}}\right)^{\beta}-1\right), \tag{26}$$

whose utility for the risk-averse decision maker with utility function $u_{\beta}(x)$ is

$$\begin{aligned} u_{\beta}(x^{\mathbf{S}}_{\beta,i}(\mathbf{p}\,\|\,\mathbf{q})) &= \frac{1}{\beta-1}((1+\beta x^{\mathbf{S}}_{\beta,i}(\mathbf{p}\,\|\,\mathbf{q}))^{(\beta-1)/\beta}-1) \\ &= S^{\mathbf{S}}_{\beta}(\mathbf{p}, \mathbf{e}_i\,\|\,\mathbf{q}). \end{aligned} \tag{27}$$

For part (b), with utility function u_{β} the formulation of Problem **P** becomes

$$\max_{\mathbf{x}\in\mathbf{R}^n} \frac{1}{\beta-1}\sum_{j=1}^{n} p_j((1+\beta x_j)^{(\beta-1)/\beta}-1) - \sum_{j=1}^{n} q_j x_j. \tag{28}$$

The first-order condition for an optimal solution is $p_i(1+\beta x_i)^{-1/\beta} = q_i$ for all i, so that $1+\beta x_i = (p_i/q_i)^{\beta}$, which yields $x^{\mathbf{P}}_{\beta,i}(\mathbf{p}\,\|\,\mathbf{q}) = \frac{1}{\beta}((p_i/q_i)^{\beta}-1)$, whence

$$\begin{aligned} &u_{\beta}(x^{\mathbf{P}}_{\beta,i}(\mathbf{p}\,\|\,\mathbf{q})) - E_{\mathbf{q}}[\mathbf{x}^{\mathbf{P}}_{\beta}(\mathbf{p}\,\|\,\mathbf{q})] \\ &\quad= \frac{1}{\beta-1}((1+\beta x^{\mathbf{P}}_{\beta,i}(\mathbf{p}\,\|\,\mathbf{q}))^{(\beta-1)/\beta}-1) - \sum_{j=1}^{n} q_j x^{\mathbf{P}}_{\beta,j}(\mathbf{p}\,\|\,\mathbf{q}) \\ &\quad= \frac{(p_i/q_i)^{\beta-1}-1}{\beta-1} - \frac{(\sum_{j=1}^{n} q_j(p_j/q_j)^{\beta})-1}{\beta} \\ &\quad= S^{\mathbf{P}}_{\beta}(\mathbf{p}, \mathbf{e}_i\,\|\,\mathbf{q}). \end{aligned} \tag{29}$$

Part (c) merely follows from the fact that both expected scores are obtained by maximizing the quantity $E_{\mathbf{p}}[u_{\beta}(\mathbf{x})] + \lambda E_{\mathbf{q}}[-\mathbf{x}]$ over all **x**, but in the case of the power score λ is set equal to one, while in the case of the pseudospherical score λ is set equal to the value that minimizes the maximum. The optimal values are equal for $\mathbf{p} \neq \mathbf{q}$ only in the case $\beta = 1$, which yields $\lambda^* = 1$. □

5. Utility/Entropy Duality in Incomplete Markets

In situations where the decision maker invests in an incomplete market for contingent claims, the relevant baseline distribution **q** is not a singleton but rather a convex set of risk-neutral distributions determined by asset prices. The problem of expected utility maximization in incomplete markets has been widely studied in the mathematical finance literature in recent years, and it has been shown that there is a duality relationship between maximization of expected utility and minimization of an appropriate divergence (e.g., Frittelli 2000, Rouge and El Karoui 2000, Goll and Rüschendorf 2001, Delbaen et al. 2002, Slomczyński and Zastawniak 2004, Ilhan et al. 2008, Samperi 2005). Most of this literature has focused on the case of exponential utility, for which the dual problem is the minimization of the reverse KL divergence $D_{KL}(\mathbf{q}\,\|\,\mathbf{p})$, as well as on issues that arise in multiperiod or continuous-time markets. In this section, we will show that in a single-period or two-period market, the duality relationship applies to the entire spectrum of LRT utility and pseudospherical divergence or power divergence.

An incomplete, single-period market can either be parameterized in terms of an $m \times n$ matrix **A**, whose rows are the (net) payoff vectors of available assets (i.e., $\mathbf{A} = \{a_{ij}\}$, where a_{ij} is the net payoff to the decision maker of one unit of the ith asset in event j), or in terms of a $k \times n$ matrix **Q**, whose rows are risk-neutral probability distributions that support the asset prices (i.e., $\mathbf{Q} = \{q_{ij}\}$, where q_{ij} is the probability of event j under the ith risk-neutral distribution). The rows of **Q** are the extremal risk-neutral probability distributions assigning nonpositive expectation to all the rows of **A**, i.e., the rows of $-\mathbf{Q}$ are the dual cone of the rows of **A**. The parameterization in terms of **Q** will be adopted here. Let **x** denote an arbitrary n-vector of monetary payoffs to the decision maker (an element of $\mathbb{R}^n$), and let **z** denote an arbitrary k-vector of nonnegative weights summing to one (an element of $\triangle^k$, the unit simplex in $\mathbb{R}^k$). As before, let **p** denote the decision maker's subjective probability distribution, and henceforth let **q** denote one of many possible probability distributions attributable to a risk-neutral trading opponent or market. Then, the decision problem can summarized in terms of **p**, **Q**, and β.

In the incomplete market generalization of Problem **S**, the problem of finding the maximum expected utility, which will be denoted as $U_\beta^{\mathbf{S}}(\mathbf{p}\,\|\,\mathbf{Q})$, is dual to the problem of finding the minimum pseudospherical divergence of order β between $\mathbf{p}$ and all $\mathbf{q}$ in the convex hull of the rows of $\mathbf{Q}$, which will be denoted as $D_\beta^{\mathbf{S}}(\mathbf{p}\,\|\,\mathbf{Q})$.

Primal Problem S:

$$U_\beta^{\mathbf{S}}(\mathbf{p}\,\|\,\mathbf{Q}) \equiv \max_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] \quad \text{subject to } \mathbf{Q}\mathbf{x} \leqslant 0.$$

Dual Problem S: $\quad D_\beta^{\mathbf{S}}(\mathbf{p}\,\|\,\mathbf{Q}) \equiv \min_{\mathbf{z}\in\triangle^k} D_\beta^{\mathbf{S}}(\mathbf{p}\,\|\,\mathbf{z}^T\mathbf{Q})$.

Note that $-\mathbf{Q}\mathbf{x}$ is the k-vector of the opponent's expected values for payoff vector $\mathbf{x}$ under all the extremal risk-neutral distributions, hence the condition $\mathbf{Q}\mathbf{x} \leqslant \mathbf{0}$ means that $\mathbf{x}$ yields nonnegative expected value to the opponent under all those distributions. In the dual problem, $\mathbf{z}^T\mathbf{Q}$ is a mixture of the rows of $\mathbf{Q}$ using weights $\mathbf{z}$, i.e., it is an element of the convex polytope of risk-neutral distributions. The special case $\beta = 1$ corresponds to logarithmic utility in the primal problem and KL divergence in the dual problem, while $\beta = 0$ corresponds to exponential utility in the primal problem and reverse KL divergence in the dual problem, and the cases $\beta = 1/2$ and $\beta = 2$ are related to the squared Hellinger distance and the Chi-square divergence, as shown in the right-hand column of Table 2.

In the incomplete market generalization of Problem **P**, the decision maker's objective is to determine an amount z to be spent at time 0 to finance consumption in period 1. For the period-1 payoff vector $\mathbf{x}$ that the decision maker wishes to purchase, the risk-neutral expected value of $\mathbf{x}$ must be less than or equal to z for all the extremal risk-neutral distributions. The corresponding primal and dual problems are

Primal Problem P:

$$U_\beta^{\mathbf{P}}(\mathbf{p}\,\|\,\mathbf{Q}) \equiv \max_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] - z \quad \text{subject to } \mathbf{Q}\mathbf{x} \leqslant z\mathbf{1}.$$

Dual Problem P: $\quad D_\beta^{\mathbf{P}}(\mathbf{p}\,\|\,\mathbf{Q}) \equiv \min_{\mathbf{z}\in\triangle^k} D_\beta^{\mathbf{P}}(\mathbf{p}\,\|\,\mathbf{z}^T\mathbf{Q})$.

Note that because the pseudospherical divergence is a monotonic transformation of the power divergence, the distribution $\mathbf{q}$ $(=\mathbf{z}^T\mathbf{Q})$ that solves Dual Problem **S** is the same one that solves Dual Problem **P**, although the objective values and the primal payoff vectors are generally different. The formal statements and proofs of these duality relationships are as follows.

THEOREM 2. (a) *In an incomplete, single-period market, maximization of expected LRT utility with risk-tolerance coefficient* β (*Primal Problem* **S**) *is equivalent to minimization of the pseudospherical divergence of order* β *between the decision maker's subjective distribution* $\mathbf{p}$ *and a risk-neutral distribution* $\mathbf{q}$ *consistent with asset prices* (*Dual Problem* **S**). *Their optimal objective values are the same, and the optimal values of the decision variables in one problem are equal to the normalized optimal values of the Lagrange multipliers in the other.*

(b) *In an incomplete, two-period market, maximization of expected quasilinear LRT utility with second-period risk-tolerance coefficient* β (*Primal Problem* **P**) *is equivalent to minimization of the power divergence of order* β *between the decision maker's subjective distribution* $\mathbf{p}$ *and a risk-neutral distribution* $\mathbf{q}$ *consistent with asset prices* (*Dual Problem* **P**). *Their optimal objective values are the same and the optimal values of the decision variables in one problem are equal to the normalized optimal values of the Lagrange multipliers in the other.*

PROOF. For part (a), Lagrangian relaxation is applicable because the primal problem has a strictly concave, continuously differentiable objective function and linear constraints. Let $\boldsymbol{\lambda}$ denote the vector of Lagrange multipliers associated with the constraints $\mathbf{Q}\mathbf{x} \leqslant \mathbf{0}$. The Lagrangian relaxation of Primal Problem **S**, which generalizes (25), is then $\min_{\boldsymbol{\lambda}\in\mathbb{R}^{k+}} L(\boldsymbol{\lambda})$, where

$$L(\boldsymbol{\lambda}) = \max_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] - \boldsymbol{\lambda}^T\mathbf{Q}\mathbf{x}. \tag{30}$$

The Lagrangian $L(\boldsymbol{\lambda})$ is an unconstrained maximum of a continuously differentiable concave function, so it can be solved for x in terms of $\boldsymbol{\lambda}$ by setting $\nabla(E_{\mathbf{p}}[u_\beta(x)] - \boldsymbol{\lambda}^T\mathbf{Q}x) = \mathbf{0}$, which yields

$$\mathbf{x} = \frac{1}{\beta}\left(\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^T\mathbf{Q}}\right)^{\beta} - 1\right), \tag{31}$$

whence

$$\begin{aligned} L(\boldsymbol{\lambda}) &= E_{\mathbf{p}}\left[\frac{1}{\beta-1}\left(\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^T\mathbf{Q}}\right)^{\beta-1} - 1\right)\right] \\ &\quad -\boldsymbol{\lambda}^T\mathbf{Q}\left(\frac{1}{\beta}\left(\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^T\mathbf{Q}}\right)^{\beta} - 1\right)\right) \\ &= \frac{1}{\beta-1}\left(E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^T\mathbf{Q}}\right)^{\beta-1}\right] - 1\right) \\ &\quad -\frac{1}{\beta}\left(E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^T\mathbf{Q}}\right)^{\beta-1}\right] - \mathbf{1}^T(\boldsymbol{\lambda}^T\mathbf{Q})\right). \end{aligned} \tag{32}$$

In the optimal solution $\boldsymbol{\lambda}^*$, where the constraints are satisfied, the second term will be zero, which implies

$$\mathbf{1}^T(\boldsymbol{\lambda}^{*T}\mathbf{Q}) = E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^{*T}\mathbf{Q}}\right)^{\beta-1}\right], \tag{33}$$

and consequently,

$$L(\boldsymbol{\lambda}^*) = \frac{1}{\beta-1}\left(E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^{*T}\mathbf{Q}}\right)^{\beta-1}\right] - 1\right). \tag{34}$$

Now let $\mathbf{z}^* = \boldsymbol{\lambda}^*/\mathbf{1}^T\boldsymbol{\lambda}^*$ be the probability distribution that is obtained by normalization of the optimal Lagrange multipliers $\boldsymbol{\lambda}^*$. Then, it follows from (33) that

$$\mathbf{z}^{*T}\mathbf{Q} = \frac{\boldsymbol{\lambda}^{*T}\mathbf{Q}}{E_{\mathbf{p}}[(\mathbf{p}/\boldsymbol{\lambda}^{*T}\mathbf{Q})^{\beta-1}]}. \tag{35}$$

The pseudospherical divergence between $\mathbf{p}$ and $\mathbf{z}^{*T}\mathbf{Q}$ can therefore be expressed in terms of $\boldsymbol{\lambda}^*$ as

$$\begin{aligned} D_\beta^{\mathbf{S}}(\mathbf{p} \,\|\, \mathbf{z}^{*T}\mathbf{Q}) &= \frac{(E_{\mathbf{p}}[(\mathbf{p}/\mathbf{z}^{*T}\mathbf{Q})^{\beta-1}])^{1/\beta} - 1}{\beta - 1} \\ &= \frac{(E_{\mathbf{p}}[(E_{\mathbf{p}}[(\mathbf{p}/\boldsymbol{\lambda}^{*T}\mathbf{Q})^{\beta-1}](\mathbf{p}/\boldsymbol{\lambda}^{*T}\mathbf{Q}))^{\beta-1}])^{1/\beta} - 1}{\beta - 1} \\ &= \frac{(E_{\mathbf{p}}[(\mathbf{p}/\boldsymbol{\lambda}^{*T}\mathbf{Q})^{\beta-1}])^{(\beta-1)/\beta}(E_{\mathbf{p}}[(\mathbf{p}/\boldsymbol{\lambda}^{*T}\mathbf{Q})^{\beta-1}])^{1/\beta} - 1}{\beta - 1} \\ &= \frac{1}{\beta - 1}\left(E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^{*T}\mathbf{Q}}\right)^{\beta-1}\right] - 1\right) = L(\boldsymbol{\lambda}^*), \end{aligned} \tag{36}$$

which is the optimal objective value of the primal problem. Furthermore, $\mathbf{z}^* = \boldsymbol{\lambda}^*/\mathbf{1}^T\boldsymbol{\lambda}^*$ must also minimize $D_\beta^{\mathbf{S}}(\mathbf{p} \,\|\, \mathbf{z}^T\mathbf{Q})$ over all $\mathbf{z} \in \triangle^k$, because if there were some other $\mathbf{z}^{**} \in \triangle^k$ such that $D_\beta^{\mathbf{S}}(\mathbf{p} \,\|\, \mathbf{z}^{**T}\mathbf{Q}) < D_\beta^{\mathbf{S}}(\mathbf{p} \,\|\, \mathbf{z}^{*T}\mathbf{Q})$, then it would be possible to find some $\boldsymbol{\lambda}^{**} \in \mathbb{R}^{k+}$ proportional to $\mathbf{z}^{**}$ such that $\mathbf{z}^{**T}\mathbf{Q} = \boldsymbol{\lambda}^{**T}\mathbf{Q}/(\mathbf{E}_{\mathbf{p}}[(\mathbf{p}/(\boldsymbol{\lambda}^{**T}\mathbf{Q}))^{\beta-1}])$. By construction, this $\boldsymbol{\lambda}^{**}$ would satisfy $E_{\mathbf{p}}[(\mathbf{p}/\boldsymbol{\lambda}^{**T}\mathbf{Q})^{\beta-1}] - \mathbf{1}^T(\boldsymbol{\lambda}^{**T}\mathbf{Q}) = 0$, implying $L(\boldsymbol{\lambda}^{**}) = D_\beta^{\mathbf{S}}(\mathbf{p} \,\|\, \mathbf{z}^{**T}\mathbf{Q})$, and it would follow that $L(\boldsymbol{\lambda}^{**}) < L(\boldsymbol{\lambda}^*)$, contradicting the assumption that $\boldsymbol{\lambda}^*$ was optimal. (In the optimal solution of Dual Problem **P**, there is a single Lagrange multiplier for the constraint $\mathbf{1}^T\mathbf{q} = 1$ as well as m Lagrange multipliers for the constraints $\mathbf{A}\mathbf{q} \geqslant \mathbf{0}$. The latter divided by the former are equal to the optimal values of the decision variables in Primal Problem **P** multiplied by $-\beta$.)

For part (b), the problem of finding the feasible risk-neutral distribution that minimizes the power divergence of order β,

$$\min_{\mathbf{z} \in \triangle^k} D_\beta^{\mathbf{P}}(\mathbf{p} \,\|\, \mathbf{z}^T\mathbf{Q}), \tag{37}$$

is equivalent to the Lagrangian problem $\min_{\boldsymbol{\lambda} \in \triangle^k} L(\boldsymbol{\lambda})$, where $L(\boldsymbol{\lambda}) = \max_{\mathbf{x} \in \mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] - \boldsymbol{\lambda}^T\mathbf{Q}\mathbf{x}$ is the same Lagrangian that was used in the proof of part (a) to minimize the pseudospherical divergence, except that here $\boldsymbol{\lambda}$ is constrained to be in the simplex, not just the nonnegative orthant ($\boldsymbol{\lambda} \in \triangle^k$ rather than $\boldsymbol{\lambda} \in \mathbb{R}^{k+}$). The optimal value of $\boldsymbol{\lambda}$ is a unit vector selecting the largest element of $\mathbf{Q}\mathbf{x}$. Let z denote this largest element. Then, $\min_{\boldsymbol{\lambda} \in \triangle^k} \max_{\mathbf{x} \in \mathbb{R}^n} E_{\mathbf{p}}[u_\beta(x)] - \boldsymbol{\lambda}^T\mathbf{Q}x$ is equivalent to $\max_{\mathbf{x} \in \mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] - z$ subject to $\mathbf{Q}\mathbf{x} \leqslant z\mathbf{1}$. □

The power divergence is always strictly greater than the pseudospherical divergence ($D_\beta^{\mathbf{P}}(\mathbf{p} \,\|\, \mathbf{q}) > D_\beta^{\mathbf{S}}(\mathbf{p} \,\|\, \mathbf{q})$), except at $\beta = 1$, as pointed out earlier, but this inequality is further illuminated by a comparison of the corresponding Lagrangian relaxation problems: the minimization of $L(\boldsymbol{\lambda})$ over $\boldsymbol{\lambda} \in \triangle^k$ must yield a result greater than or equal to its minimization over the larger set $\boldsymbol{\lambda} \in \mathbb{R}^{k+}$, whether or not the market is complete.

Versions of the duality relation of Theorem 2 have been discussed in the mathematical finance literature, as noted above, although the full spectrum of LRT utility has not previously been characterized. In particular, Goll and Rüschendorf (2001) consider the analog of Problem **S** in which the state space is $\mathbb{R}$ and the decision maker has a risk-averse utility function u with initial wealth w. In general, the solution of this problem does not have a closed form, but it can be characterized as follows. Translating from their continuous setting into our discrete setting (i.e., replacing dP and dQ with p_i and q_i, etc.), the primal utility-maximization problem with respect to a particular risk-neutral distribution $\mathbf{q}$ is

$$U_w(\mathbf{p} \,\|\, \mathbf{q}) \equiv \max_{\mathbf{x} \in \mathbb{R}^n} E_{\mathbf{p}}[u(\mathbf{x})] \quad \text{subject to } \mathbf{x}^T\mathbf{q} \leqslant w. \tag{38}$$

Let $\lambda_w(\mathbf{p} \,\|\, \mathbf{q}) \equiv \partial U_w(\mathbf{p} \,\|\, \mathbf{q})/\partial w$ denote the marginal utility of wealth at the optimum, and let $I(y) \equiv (u'(y))^{-1}$. Then, $\lambda_w(\mathbf{p} \,\|\, \mathbf{q})$ is implicitly uniquely determined by

$$E_{\mathbf{p}}\left[I\left(\lambda\frac{\mathbf{q}}{\mathbf{p}}\right)\right] = w, \tag{39}$$

in terms of which the solution of the primal problem, denoted $\mathbf{x}_w(\mathbf{p} \,\|\, \mathbf{q})$, is given by

$$\mathbf{x}_w(\mathbf{p} \,\|\, \mathbf{q}) = I\left(\lambda_w(\mathbf{p} \,\|\, \mathbf{q})\frac{\mathbf{q}}{\mathbf{p}}\right). \tag{40}$$

It follows that $\lambda_w(\mathbf{p} \,\|\, \mathbf{q})$ and $\mathbf{x}_w(\mathbf{p} \,\|\, \mathbf{q})$ are the solutions of the Lagrangian relaxation

$$\min_{\lambda > 0} \max_{\mathbf{x} \in \mathbb{R}^n} E_{\mathbf{p}}[u(\mathbf{x})] - \lambda(\mathbf{x}^T\mathbf{q} - w).$$

In the incomplete market case, let $\mathcal{Q}$ denote the set of risk-neutral distributions, i.e., the convex hull of the rows of $\mathbf{Q}$. Then, the maximum expected utility obtainable with wealth w, denoted here as $U_w(\mathbf{p} \,\|\, \mathbf{Q})$, is also the minimum of $U_w(\mathbf{p} \,\|\, \mathbf{q})$ over all $\mathbf{q}$ in $\mathcal{Q}$:

$$\begin{aligned} U_w(\mathbf{p} \,\|\, \mathbf{Q}) &\equiv \max_{\mathbf{x} \in \mathbb{R}^n} E_{\mathbf{p}}[u(\mathbf{x})] \\ &\qquad \text{subject to } \mathbf{x}^T\mathbf{Q} \leqslant w\mathbf{1} = \min_{\mathbf{q} \in \mathcal{Q}} U_w(\mathbf{p} \,\|\, \mathbf{q}). \end{aligned} \tag{41}$$

Let $\mathbf{q}^* \equiv \arg\min_{\mathbf{q} \in \mathcal{Q}} U_w(\mathbf{p} \,\|\, \mathbf{q})$, which is the *minimax measure* with respect to u, and let $\lambda_w(\mathbf{p} \,\|\, \mathbf{Q}) \equiv \lambda_w(\mathbf{p} \,\|\, \mathbf{q}^*)$ denote the marginal utility of wealth at the optimum.

To construct the dual problem, let u^* denote the convex conjugate of u:

$$u^*(y) \equiv \sup_{x \in \mathbb{R}}\{u(x) - xy\} = u(I(y)) - yI(y), \tag{42}$$

which is strictly convex if u is strictly convex. For any positive constant λ, a corresponding f-divergence can be defined by setting $f(y) = u^*(\lambda y)$:

$$\begin{aligned} u_\lambda^*(\mathbf{p} \,\|\, \mathbf{q}) &\equiv E_{\mathbf{p}}[u^*(\lambda\mathbf{q}/\mathbf{p})] \\ &= E_{\mathbf{p}}[u(I(\lambda\mathbf{q}/\mathbf{p}))] - \lambda E_{\mathbf{q}}[I(\lambda\mathbf{q}/\mathbf{p})]. \end{aligned} \tag{43}$$

Now let

$$u_\lambda^*(\mathbf{p} \,\|\, \mathbf{Q}) \equiv \min_{\mathbf{q}\in\mathcal{Q}} u_\lambda^*(\mathbf{p} \,\|\, \mathbf{q}) \tag{44}$$

be the minimum u_λ^*-divergence between $\mathbf{p}$ and $\mathcal{Q}$, also known as the u_λ^*-*projection* of $\mathbf{p}$ on $\mathcal{Q}$, and let $\mathbf{q}^{**} \equiv \arg\min_{\mathbf{q}\in\mathcal{Q}} u_\lambda^*(\mathbf{p} \,\|\, \mathbf{q})$, which is the *minimal distance measure* with respect to u_λ^*. Then, when $\lambda = \lambda_w(\mathbf{p} \,\|\, \mathbf{Q})$, the maximum expected utility and minimum divergence are related by

$$u_\lambda^*(\mathbf{p} \,\|\, \mathbf{Q}) = U_w(\mathbf{p} \,\|\, \mathbf{Q}) - \lambda w, \tag{45}$$

and moreover, $\mathbf{q}^* = \mathbf{q}^{**}$, i.e., the minimax measure with respect to u is the minimal distance measure with respect to u_λ^*. (Goll and Rüschendorf 2001, Proposition 4.3 and Theorem 5.1) Note, however, that when λ is a function of $\mathbf{p}$ and $\mathbf{q}$, as it is in this result, u_λ^* is not an f-divergence.

In the special case of LRT utility, the cast of characters has the following parametric forms:

$$I(y) = \frac{y^{-\beta} - 1}{\beta}, \tag{46}$$

$$u^*(y) = \frac{y^{1-\beta}}{\beta(\beta-1)} + \frac{y}{\beta} - \frac{1}{\beta-1}, \tag{47}$$

$$\begin{aligned} u_\lambda^*(\mathbf{p} \,\|\, \mathbf{q}) &= \sum_{i=1}^{n} E_{\mathbf{p}}\left[\frac{(\lambda\mathbf{q}/\mathbf{p})^{1-\beta}}{\beta(\beta-1)} + \frac{(\lambda\mathbf{q}/\mathbf{p})}{\beta} - \frac{1}{\beta-1}\right] \\ &= \beta D_\beta^{\mathbf{P}}(\mathbf{p} \,\|\, \lambda\mathbf{q}) \quad \forall\, \lambda > 0, \end{aligned} \tag{48}$$

$$\lambda_w(\mathbf{p} \,\|\, \mathbf{q}) = \left(\frac{E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]}{1+\beta w}\right)^{1/\beta}. \tag{49}$$

Finally, setting $\lambda = \lambda_w(\mathbf{p} \,\|\, \mathbf{q})$ yields

$$U_w(\mathbf{p} \,\|\, \mathbf{q}) = \frac{((1+\beta w)^{\beta-1} E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}])^{1/\beta} - 1}{\beta - 1}, \tag{50}$$

which in general is an affine transformation of the pseudospherical divergence, and it is precisely the pseudospherical divergence when $w = 0$. Because w enters this expression only as part of a scale factor, the minimax measure and minimal distance measure do not depend on it, which is a convenient property of LRT utility.

6. Concluding Comments

The families of weighted power and pseudospherical scoring rules developed in this paper constitute a rich set of strictly proper scoring rules. They include but are not limited to commonly encountered rules such as the quadratic, logarithmic, and spherical rules, and they can measure information relative to any baseline distribution, not just a uniform distribution. Furthermore, the flexibility of the weighted power and pseudospherical rules and their connections with well-studied generalized divergence measures and utility functions have the potential to provide further insight.

Various arguments have been put forth previously in favor of power or pseudospherical rules with particular values of β. For example, the logarithmic scoring rule is unique in that the score depends only on the probability assigned to the event that actually occurs, consistent with the likelihood principle (Winkler 1969). Also, the fact that the logarithmic rule corresponds exactly to the KL divergence and is the limiting case of both the pseudospherical and power rules at $\beta = 1$ might be viewed as a point in its favor. Selten (1998) has argued that the quadratic rule (the power rule with $\beta = 2$) is superior to the logarithmic rule precisely because (1) it is less sensitive to errors in judging small probabilities, (2) its neutrality is between true and reported distributions, and (3) it follows squarely in the tradition of least-squares estimation. Meanwhile, the spherical rule (the pseudospherical rule with $\beta = 2$) has its own appealing geometrical properties (Jose 2008). However, the effect of increasing β above one in the unweighted pseudospherical and power rule is precisely to dampen the differences in scores among all events *except the one that was judged the most probable* in absolute terms, as illustrated in Figures 1 and 2. This has the side-effect of reducing sensitivity to forecasts for low-probability events in situations where $n \geqslant 3$, merely because events deemed unlikely in absolute terms tend to receive similar scores. But in many applications, the state space is deliberately constructed to distinguish events whose a priori probabilities are small but whose consequences are large (hurricanes, economic shocks, nuclear accidents, terrorist strikes, etc.), and interest could center on accurately assessing those risks.

When some events are a priori more likely or less likely than others, this fact will presumably be known in advance to the client as well as to the forecaster, in which case a nonuniform baseline distribution is appropriate. The expected utility analysis makes clear that the usual one-parameter forms of *all* the major scoring rules are implicitly predicated on the assumption of a uniform baseline distribution. If a uniform distribution is not the appropriate straw man against which to compare the forecast, then those rules should not be expected to work well, and fiddling with β instead of $\mathbf{q}$ is not the right solution. In particular, the weighted logarithmic rule responds to Selten's objection by basing the score on the ratio p_i/q_i, i.e., the *relative* magnitude of the forecaster's probability of event i in comparison to the baseline probability.

We therefore conclude that (1) the choice between the pseudospherical rule and the power rule depends to some extent on whether a one-period market or a two-period market is the best analogy for the decision problem at hand; (2) the most appropriate values of β for either rule appear to be those in the interval from zero to one, and the cases $\beta = 0$ and $\beta = \frac{1}{2}$, which have received little (if any)

discussion in the scoring rule literature, are of interest because of their associations with the exponential utility function and the Hellinger distance, respectively; and (3) a well-chosen and not-necessarily-uniform baseline distribution is the most important parameter of the scoring rule in any case. With a nonuniform baseline, the expected scores yielded by both rules are equivalent to well-known generalized divergences, providing a natural bridge from decision analysis to information theory.

Finally, we have shown that when the decision maker invests in an incomplete market for contingent claims, characterized by a set of risk-neutral distributions rather than a single baseline distribution, there is a natural duality between maximizing LRT utility and minimizing pseudo-spherical or power divergence with the same value of β. In particular, when the decision maker has subjective probability distribution $\mathbf{p}$, maximization of logarithmic utility ($\beta = 1$) corresponds to finding the feasible risk-neutral distribution $\mathbf{q}$ that minimizes the KL divergence $D_{KL}(\mathbf{p} \,\|\, \mathbf{q})$, maximization of exponential utility ($\beta = 0$) corresponds to minimizing the reverse KL divergence $D_{KL}(\mathbf{q} \,\|\, \mathbf{p})$, and maximization of reciprocal utility ($\beta = \frac{1}{2}$) or square-root utility ($\beta = 2$) corresponds to minimization of the Hellinger distance $D_H(\mathbf{p} \,\|\, \mathbf{q})$ or the Chi-square divergence $\chi^2(\mathbf{p} \,\|\, \mathbf{q})$, respectively.

Acknowledgments

The authors are grateful to the associate editor and four referees for their many constructive comments.

References

Arimoto, S. 1971. Information-theoretical considerations on estimation problems. *Inform. Contr.* **19** 181–194.

Bickel, E. 2007. Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Anal.* **4**(2) 49–65.

Boekee, D. E., J. C. A. Van der Lubbe. 1980. The *R*-norm information measure. *Inform. Contr.* **45** 136–155.

Brègman, L. 1967. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **7** 200–217.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* **78**(1) 1–3.

Cooke, R. 1991. *Experts in Uncertainty*: *Opinion and Subjective Probability in Science*. Oxford University Press, Oxford, UK.

Cressie, N., T. R. C. Read. 1984. Multinomial goodness of fit. *J. Roy. Statist. Soc. B* **46**(3) 440–464.

Csiszár, I. 1967. Information type measures of differences of probability distribution and indirect observations. *Studia Math. Hungarica* **2** 299–318.

Dawid, A. P. 2007. The geometry of proper scoring rules. *Ann. Inst. Statist. Math.* **59** 77–93.

de Finetti, B. 1937. La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Ann. Inst. Henri Poincaré* **7** 1–68. Translation reprinted in H. E. Kyburg, H. E. Smokler, eds. 1980. *Studies in Subjective Probability*, 2nd ed. Robert Krieger, New York, 53–118.

de Finetti, B. 1974. *Theory of Probability*, Vol. 1. Wiley, New York.

Delbaen, F., P. Grandits, T. Rheinländer, D. Samperi, M. Schweizer, C. Stricker. 2002. Exponential hedging and entropy penalties. *Math. Finance* **12** 99–123.

Friedman, D. 1983. Effective scoring rules for probabilistic forecasts. *Management Sci.* **29**(4) 447–454.

Frittelli, M. 2000. The minimal entropy martingale measure and the valuation problem in incomplete markets. *Math. Finance* **10** 39–52.

Gneiting, T., A. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378.

Goel, P. 1983. Information measures and Bayesian hierarchical models. *J. Amer. Statist. Assoc.* **78** 408–410.

Goll, T., L. Rüschendorf. 2001. Minimax and minimal distance martingale measures and their relationship to portfolio optimization. *Finance Stochastics* **5** 557–581.

Good, I. J. 1952. Rational decisions. *J. Roy. Statist. Soc. B.* **14** 107–114.

Good, I. J. 1971. Comment on paper by Buehler. Godambe, Sprott, eds. *Foundations of Statistical Inference*. Holt, Reinhart, and Winston, Toronto, 337–339.

Grünwald, P. D., A. P. Dawid. 2004. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Ann. Statist.* **32** 1367–1433.

Haussler, D., M. Opper. 1997. Mutual information, metric entropy, and cumulative relative entropy risk. *Ann. Statist.* **25** 2451–2492.

Havrda, J., F. Chavrát. 1967. Quantification method of classification processes: The concept of structural α-entropy. *Kybernetika* **3** 30–35.

Hendrickson, A. D., R. J. Buehler. 1971. Proper scores for probability forecasters. *Ann. Math. Statist.* **42** 1916–1921.

Ilhan, A., M. Jonsson, R. Sircar. 2008. Portfolio optimization with derivatives and indifference pricing. R. Carmona, ed. *Volume in Indifference Pricing*. Princeton University Press, Princeton, NJ. Forthcoming.

Jose, V. R. R. 2008. A characterization for the spherical scoring rule. *Theory and Decision*. ePub ahead of print. http://dx.doi.org/10.1007/S11238-007-9067-x.

Lavenda, B. H., J. Dunning-Davies. 2003. Qualms concerning Tsallis's condition of pseudo-additivity as a definition of non-extensivity. http://arxiv.org/abs/cond-mat/0312132.

Lindley, D. 1982. Scoring rules and the inevitability of probability. *Internat Statist. Rev.* **50**(1) 1–26.

McCarthy, J. 1956. Measures of the value of information. *Proc. Nat. Acad. Sci. USA* **42** 654–655.

Nau, R. F. 1985. Should scoring rules be "effective"? *Management Sci.* **31**(5) 527–535.

Pearson, K. 1900. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London Edinburgh Dublin Philos. Mag. J. Sci. Ser. 5* **50** 157–175.

Rathie, P. N., P. Kannappan. 1972. A directed-divergence function of type β. *Inform. Contr.* **20** 38–45.

Rouge, R., N. El Karoui. 2000. Pricing via utility maximization and entropy. *Math. Finance* **10** 259–276.

Samperi, D. 2005. Model selection using entropy and geometry: Complements to the six-author paper. Working paper, Decision Synergy, New York.

Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783–801.

Selten, R. 1998. Axiomatic characterization of the quadratic scoring rule. *Experiment. Econom.* **1** 43–62.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell Sys. Tech. J.* **27** 379–423.

Sharma, B. D., D. P. Mittal. 1975. New non-additive measures of entropy for discrete probability distributions *J. Math. Sci.* **10** 28–40.

Slomczyński, W., T. Zastawniak. 2004. Utility maximizing entropy and the second law of thermodynamics. *Ann. Probab.* **32** 2261–2285.

Winkler, R. L. 1967. The quantification of judgment: Some methodological suggestions. *J. Amer. Statist. Assoc.* **62** 1105–1120.

Winkler, R. L. 1969. Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.* **64** 1073–1078.

Winkler, R. L. 1994. Evaluating probabilities: Asymmetric scoring rules. *Management Sci.* **40**(11) 1395–1405.

Winkler, R. L. 1996. Scoring rules and the evaluation of probabilities. *Test* **5**(1) 1–60.