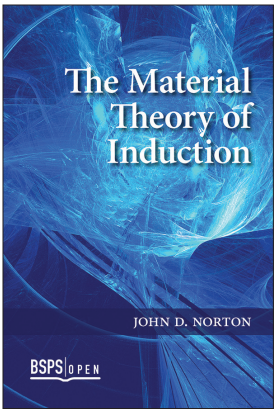

THE MATERIAL THEORY OF INDUCTION

by John D. Norton

ISBN 978-1-77385-254-6

11

Circularity in the Scoring Rule Vindication of Probabilities[1]

11.1. Introduction

The last chapter argued that all proofs of the necessity of probabilities fail. They are deductive arguments for a contingent conclusion. It is that probabilities must be used to represent inductive degrees of support or subjective degrees of belief. Thus, the proofs must employ premises that are deductively at least as strong as or even stronger than the conclusion sought. It follows that any proof of the necessity of probabilities can be undone merely by examining the premises of the proof and revealing the presence of the necessity of probability, in whatever congenial disguise it is hidden. The last chapter also predicted that any program of demonstration of the necessity of probabilities would be trapped forever in a cycle of near misses, corrections, and renewed attempts, none of which would ever succeed completely, for the program's goal is unattainable.

The present chapter offers an extended illustration of the claims of the last chapter through the recent literature that seeks to demonstrate the necessity of probabilities by means of considerations of accuracy alone, where accuracy means quantifiable closeness to the truth. This closeness is in turn measured by numerical scoring rules, which will become the major focus of what follows. If these scoring rule vindications succeed, they will have the potential to displace decision-theoretic approaches, for

1 I thank Joshua Fry, Lee Elkin, and Richard Pettigrew for helpful discussion that informed this chapter.

the scoring rule approach has no need to envisage elaborate scenarios with agents adapting beliefs to decisions that maximize utilities. Credences are chosen simply by the criterion of accuracy. The approach depends on an appealing dominance argument: if our credences are not probabilistic, then they will always be dominated by probabilistic credences in the sense that, whatever may be the case, we improve accuracy by shifting from non-probabilistic credences to probabilistic credences.

The discussion below will proceed within the framework routinely employed by the scoring rule literature. Its suppositions include

- that credences in any two propositions are always comparable; and
- that the relation of comparison can be captured by a real-valued degree in the interval 0 to 1.

Both of these suppositions, and others like them, also require justification; and attempts to justify them would in turn face just the same issues of circularity developed here.

The focus of attention in the analysis below will be the particular scoring rule employed to measure the accuracy of credences. We shall see that almost every slight change in the rule undoes the demonstration; and almost every larger change leads to a wide variety of alternative results. This shows that it is not the general notion of accuracy that drives the proof, for accuracy alone gives very little. Rather, everything depends on the delicate selection of an accuracy measure tailored to give the desired result. Herein lies the circularity. It is in this delicate fine-tuning that the probabilistic credences are presumed in disguised form.

The response to this threat of circularity has been a flourishing of attempts to make the choice of the fine-tuned scoring rule seem necessary or inevitable or perhaps just natural. We find a regress of reasons that never quite terminates successfully; or we find a proliferation of alternatives, each of which is replaced by another, without apparent end. This endless, frustrating dynamic is just what was predicted by the general argument against all proofs of the necessity of probabilities.

The exploration here of scoring rules will necessarily be partial. The literature on the topic is so large that a mere chapter can only scratch the

surface. The goal is not to review every demonstration. Rather, it is to display by example how the regress and proliferation of reasons comes about in this specific instance. In case after case, we shall see that plausible assumptions that initially appear independent of the assumption of the necessity of probabilities actually contain the assumption in covert form. An ardent vindicator will, no doubt, have further demonstrations that I have not discussed and may urge these as finally resolving all difficulties. I can only respond with some confidence as I would to a circle squarer or angle trisector: these further demonstrations would in turn succumb under scrutiny. For if they are to succeed, they must employ premises logically at least as strong as the conclusion sought.

The accuracy-driven demonstration of the necessity of probabilities draws on a much larger literature in meteorology, economics, and subjective Bayesianism that uses scoring rules for other purposes. These other uses will be sketched in Sections 11.2 and 11.3 below. They include the elicitation of true but secret probabilities from subjects who, we are to suppose, might otherwise not reveal them. In that context, the adaptation of scoring rules specifically to probabilities is benign, since these uses presume explicitly that credences are probabilistic. Use of these adapted rules in the newer context of the vindication of probabilities ceases to be benign, however, for there we are no longer allowed to presume that all credences are probabilities: the circularity of vindication lies precisely in that adaptation.

The original form of the accuracy-driven demonstration of the necessity of probabilities will be developed in Section 11.4. It employs a quadratic Brier scoring rule. This rule, we shall see, so favors probabilities that it rewards subjects with non-probabilistic credences for lying that their credences are probabilities. In Section 11.5, we will see that the success of the original accuracy-driven vindication depends on selection of exactly the Brier scoring rule and not on any other in its neighborhood. When we replace the power of 2 in the Brier score formula by a more general exponent n, the slightest change in the exponent—a shift from 2 to 2.01 or 1.99—is enough to undo the proof. Section 11.6 will reflect on how little in the original proof comes from the mere idea of accuracy, as opposed to the careful choice of scoring rule. Section 11.7 will review attempts to justify the restricted choice of scoring rule.

Section 11.8 will describe the "strictly proper" scoring rules that have been introduced into the larger literature with a different purpose. They are a generalization of the Brier scoring rule, contrived to preserve its key property of favoring probabilities. Hence, as we will see in Section 11.9, the success of strictly proper scoring rules in the dominance proof is to be expected. However, the contrived favoring of probabilities is precisely how the proof covertly assumes probabilities at the outset. Section 11.10 will review the inevitable failure of attempts to justify independently the restriction to strictly proper scoring rules in the dominance analysis. Section 11.11 will remind us once again of the pitfalls of "natural" criteria. Section 11.12 has a short conclusion.

11.2. Origins in Frequencies

The present literature on scoring rules has origins in considerations of frequencies. Identifying these considerations proves important in understanding what otherwise looks like arbitrariness in the systems now used.

In 1950, meteorologist and statistician Glenn Brier addressed a vexing problem in systems used to track the reliability of meteorologists' weather forecasts. The systems were leading meteorologists to deliver something other than their best forecasts in efforts to improve their ratings. They would, as Brier (1951, p. 10) put it, be "'hedging' or 'playing the system.'" For example, as Brier and Allen (1951, p. 843) noted, if a temperature forecast must be given as a single number, the forecaster may choose to report different temperatures according to the statistic that would be used to measure the forecaster's reliability. If it was measured by a count of how many predictions proved exactly right, the best strategy was to report the most probable temperature. If reliability was measured by mean absolute error, then the best strategy was to report the median temperature. If reliability was measured by the root-mean-square error, then the mean temperature was best. The forecaster's best judgment was overshadowed by a concern for the performance measure.

Brier's solution was to propose an assessment system that would not reward efforts to play the system: the forecasts are given as probabilities, and a "verification score"—later called the "Brier score"—is computed according to a scheme in which higher scores represent poorer performance.

If there are n possible, mutually exclusive weather conditions, the forecaster predicts them with probabilities $x_1, \ldots, x_n$. The best forecasts are to be given the lowest scores. So, if condition i does not occur, a term in x_i^2 is added to the score. The higher the probability x_i is, the more defective the prediction and thus the worse—that is—the higher the score. Correspondingly, if condition k arises, a larger associated probability x_k should contribute less to the score. This is achieved by adding a term $(1 - x_k)^2$ to the score. The final score P is recovered by averaging this sum over the N possible occasions over which the forecaster is scored.

Write x_{ik} for the probability predicted on occasion i for condition k. The actual outcomes are encoded in the matrix E_{ik}, where $E_{ik} = 1$ encodes occurrence on occasion i of condition k; and $E_{ik} = 0$ encodes its failure to occur. The "verification score" Brier proposed is

$$P = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{n} \left(E_{ik} - x_{ik}\right)^2 . \tag{1}$$

At first, the choice of a reward $(1 - x)^2$ for correct predictions and a punishment of x^2 seems arbitrary. One might imagine that almost any decreasing or increasing functions of x, respectively, would serve equally well. This turns out not to be the case, for the score has an important property shared by relatively few other scores, as we shall see in Appendix 11.B below. The property appears in the case of N occurrences of some circumstance for which the same probability forecast x_k for condition k is appropriate for each occurrence. The frequency f_k of the kth condition among the N occurrences is given by $f_k = \sum_{i = 1,N} E_{ik}/N$. For this case, Brier (1951, p. 2) described the key property:[2]

> It is also easy to show that if $[f_1, \ldots, f_n]$ are the relative frequencies that the event occurred in classes 1, 2, …, $[n]$, then the minimum score that can be obtained by forecasting the same thing on every occasion is when
>
> $$[x_{ik} = f_n]$$

In this special case, Brier's verification score reduces to 2

2 The square brackets indicate minor changes from Brier's notation to mine.

$$
\begin{aligned}
P = {} & f_1 (1 - x_1)^2 + f_2 x_1^2 + f_3 x_1^2 + \ldots + f_n x_1^2 \\
& + f_1 x_2^2 + f_2 (1 - x_2)^2 + f_3 x_2^2 + \ldots + f_n x_2^2 \\
& + \ldots \\
& + f_1 x_n^2 + f_2 (1 - x_n)^2 + f_3 x_n^2 + \ldots + f_n (1 - x_n)^2. \qquad (2)
\end{aligned}
$$

The optimal (minimum) score arises when the derivative of P with respect to each of the $x_1, \ldots, x_n$ vanishes: $dP/dx_1 = \ldots = dP/dx_n = 0$. An easy calculation shows the minimum occurs when:

$$x_i = f_i \qquad \text{for } i = 1, \ldots, n \qquad (3)$$

Brier predicted the effect of the use of this score on a forecaster:

> A little experience with the use of the score P will soon convince him that he is fooling nobody but himself if he thinks he can beat the verification system by putting down only zeros and unities when his forecasting skill does not justify such statements of extreme confidence. And in the complete absence of any forecasting skill he is encouraged to predict the climatological probabilities instead of categorically forecasting the most frequent class on every occasion. (1950, p. 2)

Two features of Brier's verification score are noteworthy. First, Brier assumed at the outset that the forecasters' predictions, both private and public, are probabilities. There are no weights that do not normalize to unity and thus need correction to bring them into conformity with the probability calculus. Second, the score is designed to ensure that forecasters' probabilities are well calibrated in the sense that they are given the best scores when their forecast probabilities for the conditions match the frequencies of the conditions. In this calibration, the probabilities are calibrated to the *short-term* frequencies in N occurrences. These are not long-term, infinite limit frequencies, but the actual frequencies in a run of N occurrences, where N may be quite small.

11.3. Eliciting Credences

Brier used his score as a way of matching weather forecasts with short-term frequencies. Around the same time as Brier's work, a second literature sprang up in which the same devices were used for a different purpose.[3] The literature addressed a subject who harbored certain credences or subjective probabilities and the task was to elicit those credences. The means was to assign a score to probabilities announced by these subjects. The Brier score is most commonly used, but not exclusively so. For example, Brier's score formula (2) is used but its terms are interpreted differently. The quantities x_i are the subject's announced probabilities, and the quantities f_i are the subject's true beliefs. Replacing frequencies f_i by probabilities p_i, we have a penalty function:

$$\begin{aligned} P = {} & p_1 (1 - x_1)^2 + p_2 x_1^2 + p_3 x_1^2 + \ldots + p_n x_1^2 \\ & + p_1 x_2^2 + p_2 (1 - x_2)^2 + p_3 x_2^2 + \ldots + p_n x_2^2 \\ & + \ldots \\ & + p_1 x_n^2 + p_2 x_n^2 + p_3 x_n^2 + \ldots + p_n (1 - x_n)^2. \end{aligned} \tag{2a}$$

If the Brier score is a penalty that the subject seeks to minimize, the analog of (3) above shows that the subject does best by announcing the subject's true beliefs.

The literature presents different scenarios to motivate an interest in what otherwise looks like an arcane scenario of dissembling subjects who may not announce their true subjective probabilities. McCarthy (1956, p. 654) imagines a forecaster and a client. The client uses the penalty as a way to "keep the forecaster honest" (the scare quotes are McCarthy's). De Finetti (1965, §3; 1974, §5.5) is more detailed. He imagines scenarios in which an expert makes a probabilistic recommendation. A geologist, for example, may announce probabilities on the success of drilling an oil well at a particular site. We interest the geologist "*in giving an honest answer; in expressing his deep felt belief*" (De Finetti 1974, p. 193; emphasis in original) by associating the score with the fee to be paid to the geologist on completion of the drilling. In another scenario, probabilistic bets are

3 See, for example, McCarthy (1956), De Finetti (1965; 1974, chap. 5), Savage (1971).

made on the outcome of sporting events and the payoff is tied to the score. Finally, it is proposed that answers to multiple choice exam questions be given as probabilities and that the final score be computed as a Brier score.

For our purposes, however, minimizing the Brier score works *too* well. Our concern includes credences that may not be probabilities. Imagine that the true credences p_i of a subject are not probabilities. They are just a set of numbers $p_1, \ldots, p_n$ that do not sum to unity. The minimum of the penalty function P of (2a) occurs when the reported values $x_1, \ldots, x_n$ are not the true credences $p_1, \ldots, p_n$ but the true credences normalized to unity.

To see this, note that the minimum of (2a) with respect to varying x_i arises when we have $dP/dx_1 = \ldots = dP/dx_n = 0$. Thus we have

$$\begin{aligned} 0 = dP/dx_1 &= -2\, p_1\, (1 - x_1) + 2\, p_2\, x_1 + 2\, p_3\, x_1 + \ldots + 2\, p_n\, x_1 \\ &= -2\, p_1 + 2\, x_1\, (p_1 + p_2 + p_3 + \ldots + p_n) \end{aligned}$$

and similar conditions for the remaining $x_2, \ldots, x_n$. Rearranging these, we have

$$x_i = p_i/(p_1 + p_2 + p_3 + \ldots + p_n) \qquad \text{for } i = 1, \ldots, n. \qquad (3a)$$

The credences reported are the true credences renormalized, so they sum to unity.

Thus, elicitation of true credences by means of a Brier score rewards subjects for lying and saying that their credences are probabilities, when they are not. This is an indication that the scoring method is biased towards probabilities, for it rewards a shift to probabilities, even when they are not the quantities sought.

11.4. The Dominance Argument

What is distinctive about the last literature discussed above is that, first, the elicitation is governed by pragmatic factors. The students score the best on an exam or the geologist is paid the most if they reveal their true probabilistic credences. Second, the primary focus is the eliciting of credences,

which are already assumed to be probabilities. It is not offered as a way of demonstrating that one's credences must be probabilities.[4]

A more recent development in this literature sought to alter both features.[5] It produced an argument for the necessity of probabilities that is presently enjoying considerable popularity. The core idea is that credences should be distributed not on pragmatic grounds but in a way that optimizes the accuracy of the credences. The main result of this development is that the accuracy of a non-probabilistic credence can always be improved by switching to a probabilistic credence, no matter which outcome obtains

The simplest instantiation of the argument employs a Brier score. We have n mutually exclusive outcomes $E_1, \ldots, E_r$ over which credences $x_1, \ldots, x_r$ are distributed. All credences here and henceforth are restricted to the interval [0, 1]. The original Brier score formula (1) or (2), (2a) is broken up into r component loss functions L_i, $i = 1, \ldots, r$, according to which of outcome $E_1, \ldots, E_r$ obtains:

$$
\begin{aligned}
L_1 &= (1 - x_1)^2 + x_2^2 + x_3^2 + \ldots + x_r^2 \\
L_2 &= x_1^2 + (1 - x_2)^2 + x_3^2 + \ldots + x_r^2 \\
&\ldots \\
L_r &= x_1^2 + x_2^2 + x_3^2 + \ldots + (1 - x_r)^2
\end{aligned} \tag{4}
$$

The greatest accuracy is achieved by minimizing these scores. Hence, it is natural to characterize the quantities as "losses" to be minimized; and to think of an increasing loss score as a measure of increasing inaccuracy.

The association of loss with inaccuracy derives from the loss generating functions used. That is, each loss function L_k, associated with outcome E_k obtaining, is a sum of r terms:

$$
\begin{aligned}
g_1(x_i) &= (1 - x_i)^2 \quad \text{when } i = k \\
g_0(x_i) &= x_i^2 \quad \text{when } i \neq k.
\end{aligned} \tag{5}
$$

Generating function $g_1(x_i)$ assures that a larger x_i makes a smaller contribution to the loss, for the case in which E_i obtains. Generating function

4 For completeness, the devices needed are present. They are just not emphasized. The essential step of the dominance argument is mentioned in passing in the captions to Figure 1 and 2 of De Finetti (1965, p. 92) and Figure 5.3 of De Finetti (1974, p. 189).

5 See, for example, Rosenkrantz (1981, 2.2), Joyce (1989, 2009), (Pettigrew 2016).

$g_0(x_i)$ assures that a larger x_i makes a larger contribution to the loss in all the remaining cases.

With the loss functions (4), no matter which of $E_1, \ldots, E_r$ obtains, we always improve accuracy by replacing a non-probabilistic credence with a probabilistic credence. The argument is represented graphically in the simplest case of two outcomes E_1, E_2, with credences x_1, x_2. Figure 11.1 shows the space of credences with individual points $<x_1, x_2>$, where both credences are restricted to values in [0, 1]. On the left, the figure shows curves of constant loss L_1. They are circular arcs, centered on the corner point, $<x_1, x_2> = <1, 0>$, On the right, the figure shows the corresponding curves of constant loss L_2. The diagonal dashed line represents those credences conforming with the additivity of the probability calculus. That is, $x_1 + x_2 = 1$.

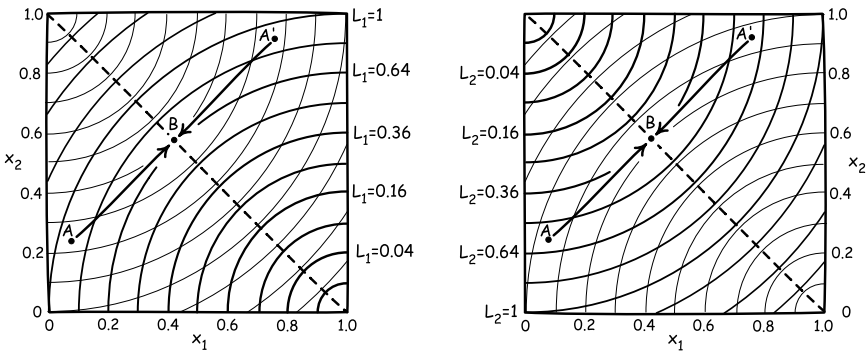

Figure 11.1. Dominance of probabilistic credences using a Brier score.

Pick any point in the space not on the diagonal, such as point A. This represents credences that violate the additivity axiom of the probability calculus. If we move along line AB, perpendicular to the diagonal, to the point B on the probabilistic diagonal, we replace the non-probabilistic credences at A with the probabilistic credences at B. We see in the figure on the left, that replacing credences at A by those at B reduces the loss L_1. The same is true if we approach probabilistic credence B from a corresponding non-probabilistic credence A', on the other side of the diagonal. That is, among all credences on the line AA', the probabilistic credence at B has

the lowest loss L_1. In other words, it is the most accurate among them if E_1 occurs. The same lines AB, $A'B$ are shown on the right. Once again, among all credences on the line AA', the probabilistic credence at B has the lowest loss L_2. It is the most accurate among them if E_2 occurs. This means that whichever of E_1 or E_2 occur, the probabilistic credence at B is the most accurate among all credences on the line AA'. Probabilistic credence B dominates: we achieve greater accuracy by replacing any non-probabilistic credence in AA' with a probabilistic credence B.

In both cases, what is key is the concavity of the curves[6] of constant loss towards the direction of smaller loss. Thus, moving towards the diagonal of probabilistic credences moves us to credences of smaller loss.

The result generalizes to the case of r outcomes, $E_1, \ldots, E_r$. The easy way to see it is to identify a differential condition that expresses the dominance. In the case of two outcomes E_1 or E_2, each probabilistic credence $<x_1, x_2>$ on the diagonal $x_1 + x_2 = 1$ dominates a set of non-probabilistic credences $\{<x_1 + k, x_2 + k>\}$ where k can have any value, both positive and negative, that generates points within the space. Each such set forms a line, such as AA' of Figure 11.1, that is perpendicular to the diagonal of probabilistic credences and will intersect it at one dominating point. For the case of L_1 and L_2 restricted just to the set $\{<x_1 + k, x_2 + k>\}$, the dominating point satisfies:

$$\frac{dL_1}{dk} = \frac{dL_2}{dk} = 0 .$$

We now give the same analysis for the case of r outcomes, $E_1, \ldots, E_r$. The hypersurface in the space of $x_1, x_2, \ldots, x_r$, corresponding to probabilistic credences is

$$x_1 + x_2 + \ldots + x_r = 1 .$$

Each such point $<x_1, x_2, \ldots, x_r>$ dominates points in the set $\{<x_1 + k, x_2 + k, \ldots, x_r + k>\}$, where k is both positive and negative as before.

6 To avoid confusion, "concavity" here simply reports that the curves of constant L_1 are geometrically concave towards the point that represents certainty of E_1's occurrence. The same property is described in Section 11.7 below, by standard convention, as the "convexity" of the function L_1. This usage presumably reflects geometrical convexity in the direction of increasing L_1.

The dominating point will satisfy an extension of the differential condition above:

$$\frac{dL_1}{dk} = \frac{dL_2}{dk} = \ldots = \frac{dL_r}{dk} = 0. \tag{6}$$

To find the dominating point, we start with some point $<x_1, x_2, \ldots, x_r>$ in the set that is not necessarily the dominating point, and we seek the value of k that satisfies condition (6). L_1 expressed as a function of k is

$$L_1(k) = (1 - x_1 - k)^2 + (x_2 + k)^2 + (x_3 + k)^2 + \ldots + (x_r + k)^2.$$

A short computation shows that condition (6) for L_1 is satisfied when

$$k = (1 - (x_1 + x_2 + \ldots + x_r))/r.$$

By the obvious symmetry in the formulae, the same value of k leads to satisfaction of condition (6) for the remaining loss functions.[7]

Thus the dominating point in the set has credences

$$X_i = x_i + (1 - (x_1 + x_2 + \ldots + x_r))/r.$$

For $i = 1, \ldots, r$. It is easy to confirm that these dominating credences satisfy the additivity condition

$$X_1 + X_2 + \ldots + X_r = 1.$$

That is, the dominating credence point $<X_1, X_2, \ldots, X_r>$ is probabilistic.

11.5. The Problem: Sensitivity to the Scoring Rule Chosen

The analysis as laid out in the last section shows a dominance argument that appears at once elegant and compelling. This impression fades, however, when we realize that the dominance of probabilistic credences depends delicately on the scoring rule or inaccuracy measure chosen. Most

7 Based on geometric intuitions, the tacit assumption above was that the set of points $\{<x_1 + k, x_2 + k, \ldots, x_r + k>\}$ is dominated by a single point. This assumption is now vindicated, since a single value of k produces a unique optimum for all loss functions. For completeness, the second derivative of all loss functions with respect to k is everywhere positive, so the optima computed are true minima.

scoring rules do not return the dominance of probabilities. Even rules that differ minutely from the Brier score are enough to undo the dominance.

To illustrate this, replace the power of 2 used in the Brier score with a different exponent n. That is, the generating functions for what I shall call the "n-power" scoring rule are now

$$g_1(x_i) = (1 - x_i)^n \quad \text{when } i = k$$

$$g_0(x_i) = x_i^n \quad \text{when } i \neq k, \tag{5a}$$

where, as before, outcome E_k is the one that obtains.

For $n > 0$, these will lead to what are, intuitively, accuracy measures. The function $g_1(x_i)$ is strictly decreasing, so it rewards a higher credence x_i in the result that obtains with a smaller loss. The function $g_0(x_i)$ is strictly increasing, so it punishes a higher credence in a result that does not obtain with a greater loss. The loss functions become

$$L_1 = (1 - x_1)^n + x_2^n + x_3^n + \ldots + x_r^n$$

$$L_2 = x_1^n + (1 - x_2)^n + x_3^n + \ldots + x_r^n$$

$$\ldots$$

$$L_r = x_1^n + x_2^n + x_3^n + \ldots + (1 - x_r)^n. \tag{4a}$$

Among all values of $n > 0$, the only value that supports the dominance of probabilistic credences is $n = 2$. The slightest deviation from it undoes the dominance. Choosing different values of n allows us to generate results of considerable variety, as we shall now see.

11.5.1. Scoring Rules with $n > 1$

We begin exploring the dominance relations by considering loss functions with $n > 1$. They exhibit dominance relations qualitatively similar to those of the Brier score. Their curves of constant loss are concave towards the region of lower loss, so that dominating points in the space arise in the same way, qualitatively, as in the case of the Brier score. However, the credences that dominate are not probabilistic. Loss functions with $1 < n < 2$ lead to superadditive credences. Loss functions with $n > 2$ lead to subadditive credences.

To recall the definitions: if credences $x(A)$ and $x(B)$ for mutually exclusive outcomes A and B are subadditive, then the credence $x(A \vee B)$ elicited

for their disjunction satisfies $x(A \vee B) < x(A) + x(B)$. If the credences are superadditive, then we have for this last case that $x(A \vee B) > x(A) + x(B)$. In the analysis that follows, we will identify subadditive and superadditive behavior in relation to the credence in the full outcome set to which credence 1 is assigned:

$$x_1 + x_2 + \ldots + x_r > 1 \qquad \text{(subadditive)}$$

$$x_1 + x_2 + \ldots + x_r < 1 \qquad \text{(superadditive)}$$

To see with least effort how these deviations from additivity arise, we calculate the dominating credence for the "diagonal" set of points:

$$\{<x_1, x_2, \ldots, x_r>: x_1 = x_2 = \ldots = x_r = x, 0 \leq x \leq 1\} \qquad (7)$$

This is just the diagonal that runs from the origin <0, 0, …, 0> to <1, 1, …, 1> of the r-dimensional hypercubic space. The dominating point in the set is identified once again by condition (6). In this set, each loss function is the same function of x:

$$L_1 = L_2 = \ldots = L_1 = L(x) = (1 - x)^n + (r - 1)\, x^n$$

A short calculation that sets $dL/dx = 0$ in accord with condition (6) shows that the minimum loss for all the loss functions occurs when[8]

$$x_{dom} = \frac{(1/r)^{1/(n-1)}}{(1/r)^{1/(n-1)} + (1 - 1/r)^{1/(n-1)}} = \frac{(1/r)^{1/(n-1)}}{(1/r)^{1/(n-1)} + (r-1)^{1/(n-1)}(1/r)^{1/(n-1)}}\,. \qquad (8)$$

That is, $<x_1, x_2, \ldots, x_r> = <x_{dom}, x_{dom}, \ldots, x_{dom}>$ dominates this diagonal set as the point of smallest loss.

To conform with the probability calculus, the r credences of this dominating point must be $x_{dom} = 1/r$, so that their sum for the r outcomes, ($r \times 1/r$), equals unity. This will happen only in two cases. First is the case of $r = 2$; that is, of two outcomes only. Then $(r - 1)^{1/(n-1)} = (1)^{1/(n-1)} = 1$, and we have, for all n, that

$$x_1 = x_2 = x_{dom} = 1/2.$$

8 For $n > 1$, the second derivative $d^2L/dx^2 > 0$, everywhere, so the turning point is a minimum.

Second is the case of the Brier score, $n = 2$. For then $1/(n - 1) = 1$, so that $(r - 1)^{1/(n-1)} = (r - 1)$; and we have for the dominating point

$$x_1 = x_2 = \ldots = x_r = x_{dom} = 1/r.$$

In all other cases, additivity fails.

For $r > 2$ and $n > 2$, the exponent in (8) satisfies $0 < 1/(n - 1) < 1$, and we have

$$(r - 1)^{1/(n-1)} < (r - 1).$$

It follows from (8) that:

$$x_{dom} > \frac{(1/r)^{1/(n-1)}}{(1/r)^{1/(n-1)} + (r-1)(1/r)^{1/(n-1)}} = \frac{(1/r)^{1/(n-1)}}{r \cdot (1/r)^{1/(n-1)}} = \frac{1}{r}$$

This entails that the r credences x_{dom} sum to greater than unity (subadditivity):

$$x_1 + x_2 + \ldots + x_r = rx_{dom} > 1$$

For $r > 2$ and $1 < n < 2$, the exponent in (8) satisfies $1/(n - 1) > 1$, and we have

$$(r - 1)^{1/(n-1)} > (r - 1).$$

By analogous reasoning to the previous case, the r credences x_{dom} sum to less than unity (superadditivity):

$$x_1 + x_2 + \ldots + x_r = rx_{dom} < 1.$$

The failure of additivity arises with the slightest deviation from the Brier score exponent 2. That is, the dominance argument fails to returns probabilities if the exponent is 2.01 or 1.99. In these cases, the deviations from additivity of the dominating credences will be small. The deviations can be made as large as we please simply by selecting suitably large or small values of n.

For example, for $r = 28$ and $n = 4$, we find $x_{dom} = 1/4$. Then the credences sum to

$$x_1 + x_2 + \ldots + x_{28} = 28(1/4) = 7$$

If we set $r = 11$ and $n = 11/10$, we find $x_{dom} \approx 10^{-10}$. Then the credences sum to

$$x_1 + x_2 + \ldots + x_{11} \approx 11 \times 10^{-10}$$

A more general sense of the range of possibilities is provided by a plot in Figure 11.2 of the sum $S = r \cdot x_{dom}$ against n, for various values of $r > 2$. Additivity is respected just when $S = 1$. This arises only when $n = 2$. All the curves intersect at $S = 1$, $n = 2$.

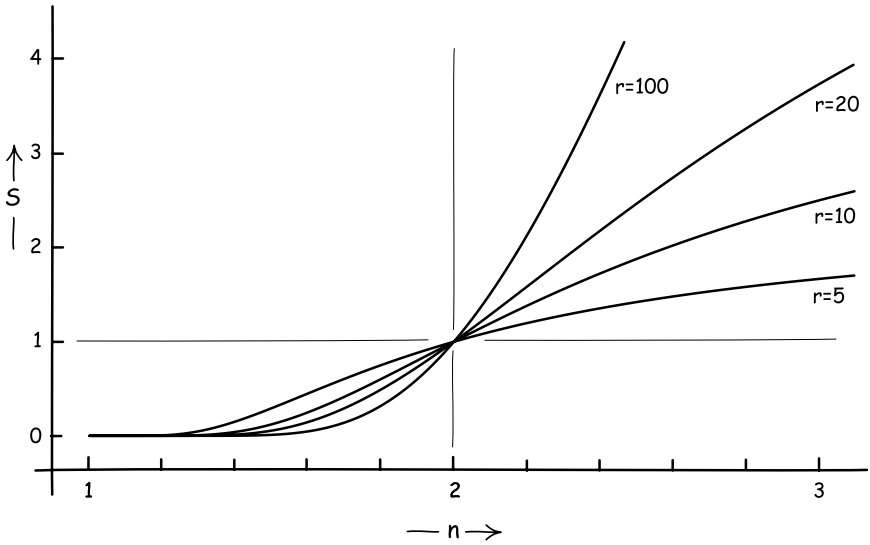

Figure 11.2. Failure of additivity for *n*-power scoring rules.

These results are a special case of the general result demonstrated in Appendix 11.A. That is, for $n > 1$, the dominating points in the space of r credences $x_1, x_2, \ldots, x_r$ lie on an $r - 1$ dimensional hypersurface in the space of credences, satisfying

$$1 = \frac{x_1^{\,n-1}}{\left[x_1^{\,n-1} + (1-x_1)^{n-1}\right]} + \ldots + \frac{x_i^{\,n-1}}{\left[x_i^{\,n-1} + (1-x_i)^{n-1}\right]} + \ldots + \frac{x_r^{\,n-1}}{\left[x_r^{\,n-1} + (1-x_r)^{n-1}\right]}. \qquad (9)$$

For $r > 2$, this surface coincides with the surface of additive probabilities

$$1 = x_1 + x_2 + \ldots + x_r.$$

only when $n = 2$. Otherwise, for $n > 2$, the surface lies above this additivity surface, and the credences are subadditive. For $n < 2$, the surface lies below this additivity surface, and the credences are superadditive.[9]

11.5.2. Scoring Rules with $0 < n < 1$

We now consider the case of loss functions (4a) with exponent n satisfying $0 < n < 1$. This case exhibits behavior that is qualitatively different from the case of $n > 1$. For now, the surfaces of constant loss are convex towards the direction of smaller loss. This inclines credences to move to extreme values to secure smaller losses. This effect can be seen in the case of two outcomes, $r = 2$, and a square root loss function, $n = 1/2$. Then we have two loss functions:

$$L_1 = \sqrt{1 - x_1} + \sqrt{x_2}$$

$$L_2 = \sqrt{x_1} + \sqrt{1 - x_2}\ .$$

Curves of constant loss are plotted in Figure 11.3. Those for loss L_1 are on the left, and those for loss L_2 are on the right. Probabilistic credences satisfying $x_1 + x_2 = 1$ lie on the dashed diagonal.

9 Equation (8) picks out a point on this surface. It is recoved by substituting $x_1 = \ldots = x_r = x$ into (12) and solving for x.

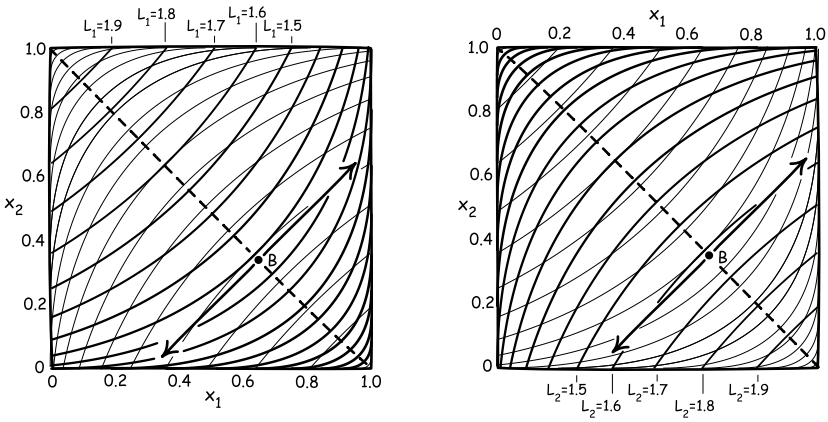

Figure 11.3. Dominance of extremes with n = 1/2.

Repeating the analysis of Figure 11.1, we find in this case that moving credences away from the diagonal decreases both loss functions L_1 and L_2 and thus increases accuracy. An arbitrarily chosen additive credence at B is dominated by non-additive credences to which we arrive by following the arrows towards the extremes. Most striking is that the additive credences at $x_1 = x_2 = 0.5$ are dominated by the credences $x_1 = x_2 = 0$ and $x_1 = x_2 = 1$.

This striking behavior of the dominance of probabilistic credences by both subadditive and superadditive credences is an artifact of having just two outcomes, $r = 2$. For the case of more than two outcomes, the dominating credences all have lower values and are superadditive. This is easy to see in the case of the diagonal set (7). All the loss functions for it are the same for the case of $n = 1/2$:

$$L_1 = L_2 = \ldots = L_r = L(x) = \sqrt{1-x} + (r-1)\sqrt{x}\ .$$

More generally, for all $0 < n < 1$, the loss functions are

$$L_1 = L_2 = \ldots = L_r = L(x) = (1-x)^n + (r-1)x^n\ .$$

For all of these cases, the loss functions have a dominating minimum at the origin only:

$$x_1 = x_2 = \ldots x_r = x = 0,$$

where $L = 1$.[10] When $x_1 = x_2 = \ldots x_r = x = 1$, $L = r - 1$, which is greater than one for $r > 2$.

11.5.3. Scoring Rules with $n = 1$

The final case uses the absolute norm. That is, the generating functions are now[11]

$$g_1(x_i) = (1 - x_i) \text{ when } i = k$$

$$g_0(x_i) = x_i \qquad \text{when } i \neq k, \tag{5b}$$

where, as before, E_k is the outcome that obtains. In the case of two outcomes, this scoring rule exhibits qualitatively different behavior again. The two loss functions are

$$L_1 = (1 - x_1) + x_2 = 1 - (x_1 - x_2)$$

$$L_2 = x_1 + (1 - x_2) = 1 + (x_1 - x_2).$$

The curves of constant loss for both are the same

$$x_1 - x_2 = \text{constant}.$$

They differ only in the values assigned to the curves. Since $L_2 = 2 - L_1$, the curves differ in the direction of increasing loss. These curves are plotted in Figure 11.4, with curves of constant L_1 on the left and curves of constant L_2 on the right.

10 Write, $L(x, n) = (1 - x)^n + (r - 1)\, x^n$. We have $L(0, n) = 1$. Also $L(x, 1) = 1 + (r - 2)\, x > 1$, for all $x > 0$, $r > 2$. But $L(x, n) > L(x, 1)$, for all $0 < n < 1$ and $x > 0$, since then $(1 - x)^n > (1 - x)$ and $x^n > x$.

11 This case is often presented as the absolute norm, writing $g_1(x_i) = |1 - x_i|$. Since $0 \leq x_i \leq 1$, the absolute operator $|.|$ is superfluous.

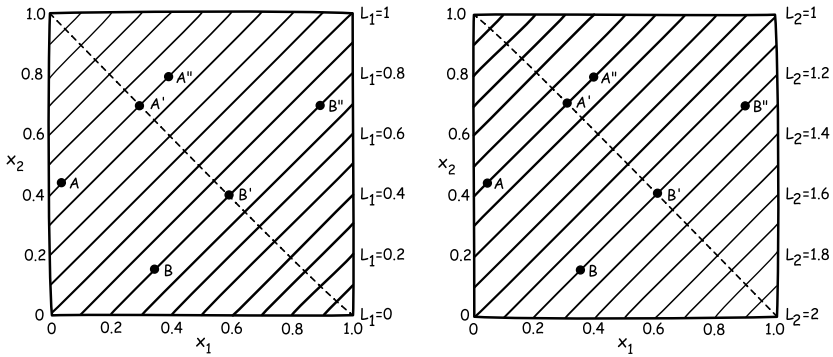

Figure 11.4. Degeneracy of dominance with n = 1.

In this degenerate case, dominance fails, since both loss functions are constant along the curves shown. Thus, as far as the accuracy measure is concerned, all the credences A, A', A'', … are equally accurate; and all the credences B, B', B'', … are equally accurate.

This degeneracy is not specific to the absolute norm $n = 1$, but is recoverable in the case of two outcomes, $r = 2$. For example, take the generating functions

$$g_1(x_i) = 1 - h(x_i) \quad \text{when } i = k$$

$$g_0(x_i) = h(x_i) \quad \text{when } i \neq k, \tag{5d}$$

where, as before, outcome E_k is the one that obtains. Then, as above, curves of constant loss for both L_1 and L_2 are the same:

$$h(x_1) - h(x_2) = \text{constant}$$

Instead of a dominance relation, we find all credences on each of the curves to have the same loss L_1 and L_2 and thus to be equally accurate. We can take many increasing functions for $h(x)$, such as $h(x) = x^2$. For this case, these curves are hyperbolas with an asymptote of $x_1 = x_2$.

The degeneracy of the absolute norm rule does not persist when we move to more than two outcomes, $r > 2$. Then, smaller-valued credences dominate. The loss functions are

$$L_1 = 1 - x_1 + x_2 + x_3 + \ldots + x_r$$

$$L_2 = x_1 + 1 - x_2 + x_3 + \ldots + x_r$$

$$\ldots$$

$$L_r = x_1 + x_2 + x_3 + \ldots + 1 - x_r.$$

For the diagonal set of credences (7), all the loss functions are equal

$$L_1 = L_2 = \ldots = L_1 = L(x) = 1 + (r - 2)x.$$

The dominating credence is

$$x_1 = x_2 = \ldots = x_r = x = 0.$$

More generally, uniformly reducing credences in such a way that we remain within the space $0 < x_i < 1$ ($i = 1, \ldots, r$), uniformly decreases all the loss functions and thus increases accuracy. For example, we start at $\mathbf{x} = < x_1, x_2, \ldots, x_r >$ in this space and move to a new point:

$$\mathbf{x} - \boldsymbol{\varepsilon} = < x_1 - \varepsilon, x_2 - \varepsilon, \ldots, x_r - \varepsilon >$$

for some increment $\varepsilon > 0$ sufficiently small to keep us in the space. Then we have for all $i = 1, \ldots r$,

$$L_i(\mathbf{x} - \boldsymbol{\varepsilon}) = L_i(\mathbf{x}) - (r - 2)\,\varepsilon.$$

Thus the credence $\mathbf{x}$ is dominated by the uniformly smaller credence $\mathbf{x} - \boldsymbol{\varepsilon}$. We can continue descending to smaller credences until we finally strike the origin $\mathbf{x} = \mathbf{0}$ or end up on one of the two-dimensional edges of the hypercubic space (in which case, the above degeneracy replaces the dominance relations).

11.6. Accuracy Gives Very Little

In sum, the above exploration shows that the accuracy dominance of probabilistic credences is fragile. It depends critically on choosing exactly the right scoring rule. The Brier score belongs to a larger family of power rule scores (4a) and (5a), characterized by the exponent n. The case of $n = 2$ is the only case among them that returns the dominance of probabilistic credences. Other values of n give widely varying results. For $n > 2$,

the dominating credences are subadditive. For $1 < n < 2$, the dominating credences are superadditive. Scoring rules with $0 < n \leq 1$ generally exhibit dominance by the lower values of credence in the space. Cases of equal credence, such as the probabilistic $x_i = 1/r$, ($i = 1, \ldots, r$) are dominated by all-zero credences $x_1 = x_2 = \ldots x_r = 0$, for example. We also saw anomalous cases of dominance by small and large credences and failures of dominance, in favor of equality of accuracy over some sets of credences.

If one is not antecedently committed to probabilistic credences, there is nothing especially troublesome in these results. We learn from them that a requirement of accuracy does not have univocal import. It must balance rewards for credence in the outcome that obtains with punishments for credences in those that do not. There are, it turns out, many ways to effect this balance. There is no obviously right way to do it.

Some rules, such as those with $n > 1$, encourage prudence and direct credences towards intermediate values, while generally still not favoring probabilities. Others (such as $n = 1/2$, $r = 2$) effect the balance so that rashness is rewarded. All unit credences dominate in the equal credence case, since the reward for assigning unit credences to the outcome that obtains exceeds the punishment for assigning unit credences to the outcome that does not obtain. Still other rules encourage timidity. Accordingly, assigning all-zero credences is most accurate, since the reward for a higher credence on the outcome that obtains is overwhelmed by the punishment for higher credences in outcomes that do not obtain.

These are widely varying results and we should accept them. To do otherwise and select among them for those we prefer is simply to invalidate the whole accuracy-based method. We would not be using the method to inform our understanding and correct our prejudices. We would be using our prejudices to overturn what our method tells us.

11.7. Attempts to Justify the Choice of Scoring Rule

If one is antecedently committed to probabilistic credences, matters look very different. The results are troublesome. One has to find some way to impugn virtually all the accuracy measures employed in favor of the very few that return the desired result. In effect, one must work backwards from the probabilistic result desired to a condition that will deliver it. When

working backwards is done well, the resulting conditions will be congenial to those who already conceive of credences as probabilities. To those who are not antecedently committed to probabilistic credences, however, these conditions will appear as arbitrary as the original commitment to probabilistic credences. This, in my view, is what the following review of these attempts shows.

Rosenkrantz (1981, 2.2) presented an early attempt to justify the Brier score independently within the context of a dominance-based vindication of probabilities. He noted that when the Brier score is used for elicitation of credences, it has the property that a subject with non-probabilistic credences minimizes the loss by reporting credences that are *proportional* to the "true probabilities." This, he called "absolutely non-distorting." Rosenkrantz conjectured but did not show that the Brier score is uniquely selected by this property, supplemented by other, weaker properties. The analysis seems hasty, since all strictly proper scoring rules (to be discussed below) share this property. Moreover, the property does not seem praiseworthy, since it is just the result reported above in Section 11.3, namely that a Brier score elicitation rewards subjects for lying about their non-probabilistic credences by rescaling them to probabilities with a constant multiplicative factor.

Joyce's (1998) proposal for restricting scoring rules is more definite and more confident. His "main theorem" (pp. 587–88) shows that probabilistic credences dominate if we use a scoring rule that satisfies six conditions that he names: "Structure," "Extensionality," "Normality," "Dominance," "Weak Convexity," and "Symmetry." None of these conditions is a logical necessity. Each is merely natural for probabilists. Each introduces into the proof a contingent presupposition congenial to probabilists. As a result, each contributes to the circularity. Lest the analysis grow too lengthy, we consider only two of the strongest conditions: Weak Convexity and Symmetry.

If two credences **c** and **c'** have the same score on some outcome, then Weak Convexity requires that the score assigned to their midpoint, (**c** + **c'**)/2 is strictly less, unless **c** = **c'**. Considered abstractly, the requirement seems natural enough. "Weak Convexity is motivated by the intuition that extremism in the pursuit of accuracy is no virtue," Joyce (p. 596) assures us. However, Weak Convexity is violated by power scoring rules

with $0 < n < 1$. As we saw above in Section 11.6, that does not make them defective, but just different ways of balancing the rewards for true beliefs and punishments for false beliefs. To preclude them is not to learn from what accuracy measures tell us, but to tell accuracy measures what they should be doing to accord with our other notions. It is part of the artificial adjustment of the premises needed if the demonstration is to yield the predetermined result—that is, the necessity of probabilities.

Weak Convexity alone, however, does not restrict power scoring rules with $n > 1$. The further restriction needed in the main theorem is Symmetry. If two credences **c** and **c'** have the same score on some outcome i, then the distribution of scores over the intermediate credences is symmetric in the sense that, for any $0 \leq \lambda \leq 1$

$$L_i(\lambda \mathbf{c} + (1 - \lambda)\mathbf{c'}) = L_i((1 - \lambda)\mathbf{c} + \lambda \mathbf{c'}).$$

This condition picks out just the quadratic Brier score from all n-power scoring rules as required.[12] Thus, if we are working backwards to a predetermined result, the condition will seem apposite. However, it is difficult to see any independent justification for it. Joyce's rationale (p. 597) merely restates what the formula says in words and suggests that Symmetry somehow precludes an improper favoring of one credence over another.

About a decade later, Joyce (2009) had presumably recognized the fragility of positing these conditions unequivocally. They were, he conceded, "not all well justified" (p. 264), and a reappraisal was undertaken. Indeed, at times the commitment to the overall project is equivocal. The decline predicted earlier seems well underway. He writes: "Readers will be left to decide for themselves which of the properties discussed below conform to their intuitions about what makes a system of beliefs better or worse from the purely epistemic perspective" (p. 266). A proof has scant foundations if acceptance of its premises depends on the intuitions of individual readers. My intuitions about angles and lines are immaterial to

12 An easy way to see this is to consider credences $(x_{\text{dom}} + \varepsilon)$ among the diagonal set (7) in the immediate vicinity of the dominating point x_{dom}, for $n > 1$. The symmetry of scoring rule L_i will manifest in the vanishing of the cubic term in e^3 in the power series expansion

$$L_i(x_{\text{dom}} + \varepsilon) = L_i(x_{\text{dom}}) + \varepsilon\, L_i'(x_{\text{dom}}) + e^2/2\, L_i''(x_{\text{dom}}) + e^3/6\, L_i'''(x_{\text{dom}}) + \ldots$$

However, $L_i'''(x_{\text{dom}}) = 0$ only in the case of $n = 2$.

the proof of Pythagoras' theorem or the impossibility of duplicating the cube. In a notable compromise of the entire program of providing quantitative, normative guides to credences, he notes that "epistemic goodness or badness for partial beliefs can be made sufficiently precise and determinate to admit of quantification" is merely a "useful fiction." Further, he reports on a newly named condition, "admissibility," which "is not a substantive claim about epistemic rationality" but a way to "capture one's sense of what is valuable about beliefs from a purely epistemic perspective" (p. 267). Nonetheless, it is used to restrict the choice of scoring rules, although apparently on rather infirm ground.

One should not fear that Joyce (2009) has abandoned the original project entirely. For eventually Joyce settles on what is offered as the "least restrictive" of the theorems that employ dominance ideas to demonstrate the necessity of probabilities. The theorem—whose details are found in Joyce (2009, pp. 287–88)—depends, among other things, on the condition of "Coherent Admissibility" (p. 280). This condition dismisses a scoring rule as "unreasonable" if it assigns a worse score to a probabilistic credence than to a non-probabilistic one in the case of all outcomes.

Hannes Leitgeb and Richard Pettigrew (2010, p. 246) seem to me to give the correct appraisal. As they put it, Coherent Admissibility is far from benign since "it accords a privileged status to probability functions." They add,

> We are inclined to ask: Why is it that we are justified in demanding that every probability function is admissible? Why are we not justified in demanding the same of a belief function that lies outside that class? And, of course, we must not make this demand of any nonprobability function.

Just this sort of privileging of probabilities seems quite benign if one is working backwards from the predetermined conclusion that credences must be probabilities, for the condition says that a scoring rule cannot preclude probabilities, as Joyce says, "a priori" (2009, p. 280). It does not appear benign to those who have not already prejudged the outcome.

A real difficulty for probabilists is that once one becomes convinced that credences have to be probabilities, it is hard to conceive of

how alternatives could be cogent. This may be behind Joyce's (2009, p, 283) concerns that the all-zero-valued credences that can dominate with power scoring rules when $0 < n \leq 1$. His assessment is severe. He calls them "logically inconsistent," since "the believer minimizes [the] expected inaccuracy by being absolutely certain that every [proposition] is false even though logic dictates that one of them must be true." This accusation of logical inconsistency will be unwelcome to proponents of the Shafer-Dempster theory of belief functions. Complete ignorance is represented there by assigning zero-valued belief functions "Bel" to all outcome sets except the universal set. We see here that Joyce's assessments are driven by a prior commitment to interpreting credences as probabilities, so that zero credence coincides with certain falsity.[13] In the Shafer-Dempster theory, a zero-belief function can be interpreted as demarcating an interval of belief stretching from zero to one.

In my view, the most promising avenue for restriction of scoring rules is through the class of "strictly proper" scoring rules that are much used elsewhere. Joyce (2009, §8) discusses and defends them. Let us first review them.

11.8. Strictly Proper Scoring Rules

Strictly proper scoring rules arose in the context of scoring a predictor's performance and of the elicitation of subjective probabilities. It addresses the problem that most alternatives to the Brier rule do not deliver probabilistic credences at their minima.

For example, we can generalize the Brier rule by replacing its exponent 2 by an arbitrarily selected n, as in the n-power rule of (5a) above. It is shown in Appendix 11.B below that the only value of n that gives a rule that correctly elicits probabilities is $n = 2$. For all $n > 2$ (and $r > 2$), the power rule (5a) elicits subadditive credences. Alternatively, if $1 < n < 2$, then the n-power rule elicits superadditive credences.

13 Of course, even for probabilists, zero probability does not coincide with certain falsity, but merely measure zero improbability. De Finetti's finitely additive treatment of the infinite lottery assigns zero probability to each outcome individually, even though one must obtain. That a dart strikes any particular point on the board is a probability zero outcome, even though some point must be struck.

These general n-power rule elicitations have an awkward property that is something like the reverse of the $n = 2$ Brier rule. We saw above in Section 11.3 that the Brier rule elicits an additive probability measure, even when the subject's true credences are not probabilistic. The n-power rule (for n not 2) elicits credences that are not probabilities, even when the subject's true credences are probabilities.

The upshot is that the formal properties of the credences elicited by the scoring rule method will only be probabilities if the rule used is very carefully tuned to give just that result. The standard response in the literature on elicitation and assessment of a predictor's performance is to restrict the scoring rules under consideration to "strictly proper" scoring rules.

As a background to the notion, we recall that a general scoring rule employs two functions: $g_1(x)$ to reward a credence x in what turns out to be the true outcome; and $g_0(x)$ to punish a credence x in an outcome that turns out not to be true. The loss score assigned to elicited credences $\mathbf{x} = \langle x_1, x_2, \ldots, x_r \rangle$ for true probabilistic credences or true frequencies $\mathbf{p} = \langle p_1, p_2, \ldots, p_r \rangle$ is

$$\begin{aligned} L(\mathbf{p}, \mathbf{x}) = {} & p_1\, g_1(x_1) + \ldots + p_1\, g_0(x_i) + \ldots + p_1\, g_0(x_r) \\ & + \ldots \\ & + p_i\, g_0(x_1) + \ldots + p_i\, g_1(x_i) + \ldots + p_i\, g_0(x_r) \\ & + \ldots \\ & + p_r\, g_0(x_1) + \ldots + p_r\, g_0(x_i) + \ldots + p_r\, g_1(x_r). \end{aligned} \tag{10a}$$

The most direct definition (such as given in Gneiting and Raftery 2007, p. 359) simply asserts that

Strictly Proper I

A scoring rule L is strictly proper just if $L(\mathbf{p}, \mathbf{x}) \geq L(\mathbf{p}, \mathbf{p})$, for all p_i in $0 \leq p_i \leq 1$, $i = 1, \ldots, r$, with equality only when $\mathbf{x} = \mathbf{p}$.

This definition explicitly rules out by fiat any scoring rule that fails to elicit $\mathbf{x}$ as a probability measure. Note that the definition is so strong that, like the Brier rule, a strictly proper scoring rule will elicit a probability even when the subject's true credences are not probabilities. To illustrate,

imagine that the subject's true credences are a non-probabilistic $\mathbf{q} = (q_1, q_2, \ldots, q_r)$. We can normalize them to a probability

$$\mathbf{p} = \langle p_1, p_2, \ldots, p_r \rangle = \mathbf{q}/Q = \langle q_1/Q, q_2/Q, \ldots, q_r/Q \rangle$$

by dividing by $Q = (q_1 + q_2 + \ldots + q_r)$. If the subject's true probability is $\mathbf{p}$, we know that the scoring rule will elicit $\mathbf{x} = \mathbf{p}$. By the definition of strictly proper scoring rules, $\mathbf{x} = \mathbf{p}$ is the unique value of $\mathbf{x}$ that minimizes $L(\mathbf{p}, \mathbf{x})$. However, $L(\mathbf{p}, \mathbf{x})$ is linear in $\mathbf{p}$ so that $L(\mathbf{p}, \mathbf{x}) = L(\mathbf{q}, \mathbf{x})/Q$. Hence, $\mathbf{x} = \mathbf{p}$ will also minimize $L(\mathbf{q}, \mathbf{x})$ uniquely. That is, if the subject's true credences are a non-probabilistic $\mathbf{q}$, a strictly proper scoring rule will reward the subject most if the subject lies and reports a probabilistic, normalized credence $\mathbf{p} = \mathbf{q}/Q$.

11.9. Strictly Proper Scoring Rules in the Dominance Argument

This favoring of probabilities by strictly proper scoring rules is unproblematic in the context in which the notion was introduced. For when the rules are used to elicit probabilities from a subject, we begin with the assumption that the subject's credences are already probabilities. Correspondingly, when we use the rules to assess the performance of a predictor against the actual frequencies of outcomes, these actual frequencies are also additive measures.

The use of strictly proper scoring rules ceases to be benign, however, when they are used as part of a vindication of probabilities. For the rules are engineered to favor probabilities and will yield them even when they are not the subject's credences. They exhibit the same favoring of probabilities if they are used as accuracy measures in the dominance arguments used to vindicate probabilities. A much-noted theorem in the scoring rule literature asserts exactly this: any non-probabilistic credence $\mathbf{q}$ is strongly dominated by a probabilistic credence $\mathbf{p}$, where "strongly dominated" means that $\mathbf{p}$ has a strictly lower score than $\mathbf{q}$ for all possible outcomes when the scoring rule used is strictly proper.[14]

14 See, for example, Predd et al. (2009, p. 4788).

A simpler but less transparent definition of a strictly proper scoring rule lets us display the dominance in an example.

Strictly Proper II[15]

A scoring rule L is strictly proper just if $pg_1(x) + (1 - p)g_0(x)$ is uniquely minimized at $x = p$ for all $0 \leq p \leq 1$.

This definition is equivalent to the definition *Strictly Proper I*.[16]

This simpler form of the definition lets us see quickly how probabilistic credences dominate in a special case, that of the "diagonal" set (7) of credences above. For the general scoring rule, the generalization of the r loss functions (4) and (4a) above is the following:

$$L_1 = g_1(x_1) + g_0(x_2) + g_0(x_3) + \ldots + g_0(x_r)$$

$$L_2 = g_0(x_1) + g_1(x_2) + g_0(x_3) + \ldots + g_0(x_r)$$

$$\ldots$$

$$L_r = g_0(x_1) + g_0(x_2) + g_0(x_3) + \ldots + g_1(x_r). \tag{4a}$$

For the diagonal set (7) of credences, all of these loss functions reduce to the same expression:

$$L = L_1 = L_2 = \ldots = L_r = g_1(x) + (r - 1)\, g_0(x) = r \cdot [\, (1/r)\, g_1(x) + (1 - 1/r)\, g_0(x)\,]$$

The second definition of strict propriety tells us directly that all of these loss functions are uniquely minimized when

$$x = x_1 = x_2 = \ldots = x_r = 1/r.$$

That is, all credences in the set are strongly dominated by this probabilistic credence.

The selection of a strictly proper scoring rule in the accuracy-driven vindication of probability amounts to a delicate fine-tuning of the analysis to give just the probabilistic result antecedently desired. The extent of the fine-tuning depends on just how sparsely the strictly proper scoring rules

15 Predd et al. (p. 4787) also include the requirement that the functions $g_0(x)$ and $g_1(x)$ are continuous. Schervish, Seidenfeld, and Kadane (2009, p. 205) relax the condition of continuity. Some of my analysis assumes differentiability of these functions, however.

16 For a demonstration of the equivalence, see Appendix 11.D.

are distributed among scoring rules that we would intuitively judge to be admissible measures of accuracy.

In short, the strictly proper rules are very sparsely distributed among this larger class of rules. This is already suggested by theorems such as those of Schervish (1989), which show how all strictly proper scoring rules can be generated from the selection of a small class of functions. We can more directly gauge the sparseness by means of the second definition above. In brief, we have considerable freedom in selecting either of the functions $g_0(x)$ or $g_1(x)$. But once one is fixed, then so is the other; and we can generate an arbitrary number of scoring rules that are not strictly proper simply by selecting different functions for the second.

To see this, assume that $g_0(x)$ is fixed at some function suitable for penalizing a credence x on an outcome that does not obtain. We have from the second definition that $pg_1(x) + (1 - p)g_0(x)$ has a unique minimum, for fixed p, when $x = p$. This minimum arises when the derivative with respect to x vanishes

$$p\frac{dg_1(x)}{dx} + (1-p)\frac{dg_0(x)}{dx} = 0 \cdot$$

Substituting $x = p$ at this minimum, we have

$$x\frac{dg_1(x)}{dx} + (1-x)\frac{dg_0(x)}{dx} = 0 \cdot$$

Since p can have any value in $0 \leq p \leq 1$, this relation is a restriction on the functions $g_0(x)$ and $g_1(x)$ for any x in the same range. It follows that

$$g_1(x) - g_1(0) = -\int_0^x \left(\frac{1-y}{y}\right)\frac{dg_o(y)}{dy}dy \,. \tag{11}$$

Reading from right to left in this formula, fixing $g_0(x)$ fixes $g_1(x)$ up to the additive constant $g_1(0)$. Selecting any other function for $g_1(x)$ will yield a scoring rule that is not strictly proper. For example, if we fix $g_0(x) = x^n$ for $n > 1$, then a short calculation shows that $g_1(x)$ must be

$$g_1(x) = x^n - \left(\tfrac{n}{n-1}\right)x + 1$$

up to the additive constant $g_1(0) = 1$. Any other choice of function for

$g_1(x)$, such as the apparently "natural" n-power rule (5a), fails to be strictly proper.

11.10. Justifying Strict Propriety

A dominance-accuracy argument for probabilities that employs strictly proper scoring rules must provide independent grounds for the restriction to strictly proper scoring rules. That these rules are popular in the broader elicitation literature provides no such grounds. Indeed, it is quite the reverse. Since strictly proper scoring rules have been designed explicitly to favor probabilities, using them to preclude non-probabilistic credences is prima facie circular. Their favoring is so strong that, used as a means of elicitation, they will reward a subject with non-probabilistic credences who lies and declares probabilistic credences.

All that can now prevent the analysis from collapsing into circularity is some independent justification of the use of strictly proper scoring rules. Joyce (2009, pp. 277–79) attempts such a justification by means of the notion of "immodesty." The quantity $L(\mathbf{p}, \mathbf{x})$ of (10a) is the probabilistically expected score using rule L of a credence $\mathbf{x}$, according to the expectations of probabilistic credence $\mathbf{p}$. A "modest" credence will judge $L(\mathbf{p}, \mathbf{x}) < L(\mathbf{p}, \mathbf{p})$. That is, it will judge some other credence $\mathbf{x}$ to have a lower expected score and thus to be more accurate than $\mathbf{p}$ itself. This is a poor situation for credence $\mathbf{p}$, since considerations of expected accuracy indicate that, by $\mathbf{p}$'s own assessment, credence $\mathbf{x}$ is the better one. The credences we should seek, therefore, are "immodest." They are such that, by their own lights, they are the most accurate.

This favoring of immodest credences is, in effect, a guide for selecting scoring rules, for a credence can only be immodest or modest relative to a scoring rule. This guide leads us directly to strictly proper scoring rules. We are asking for rules in which $L(\mathbf{p}, \mathbf{p})$ takes the minimum value in comparison with all other $L(\mathbf{p}, \mathbf{x})$. But just this property of a scoring rule is strict propriety in the form of definition I of Section 11.8 above.

The justification of a restriction just to strictly proper scoring rules is still not complete. For nothing so far precludes another scoring rule that might render some non-probabilistic credence immodest. The analysis stalls at this point since we have no precise characterization of this last

sort of scoring rule. Note that the score $L(\mathbf{p}, \mathbf{x})$ of a strictly proper rule is the expected score for credence $\mathbf{x}$ according to probability $\mathbf{p}$. If we seek an immodest, non-probabilistic credence $\mathbf{y}$, then we would replace $\mathbf{p}$ in the score by $\mathbf{y}$. But then $L(\mathbf{y}, \mathbf{x})$ is no longer an expectation. It is unclear how the quantity should be interpreted.[17] We have no clear way to characterize an immodest, non-probabilistic credence.

The regress of reasons must continue. In an attempt to complete the justification, Joyce considers cases of physical chances in which we naturally choose probabilistic credences. What credence can we have in the each of the six outcomes of a fair die throw, other than a probability of one sixth? Thus we should demand the hospitality condition of "Minimal Coherence" of our scoring rules: they should not preclude in advance probabilistic credences. This way credences concerning physical chance can be accommodated. If, however, we require both immodesty and the possibility of rules that favor probabilistic credences in their expectations, then we are led to strictly proper scoring rules. They are, by their definition, the only rules that can serve.

As we have already seen, this latest step in the regress of reasons will seem quite compelling to someone who antecedently favors probabilities. It is surely benign, they might think, to demand that we use scoring rules that are minimally hospital to probabilities in the sense that they do not automatically preclude them. To someone who has not prejudged the outcome, the demand is anything but benign.[18] For the burden of the analysis shows that this demand is enough to force probabilistic credences in all cases.

If our earnest desire is not to prejudge, then should we not ask that our scoring rules be hospitable to more than just probabilistic credences? Once we demand hospitality for one favored type of credence, no others are sustainable.

---

17 For example, expectation-like quantities computed using a non-probabilistic $\mathbf{y}$ fail to meet minimal conditions of an expectation. For example, the expectation for a quantity $\mathbf{Q} = <Q_1, Q_2, \ldots, Q_r>$ in the special case in which $Q_1 = Q_2 = \ldots = Q_r = Q$, should be Q. However the sum $\Sigma_i\, y_i Q_i = \Sigma_i\, y_i Q$ is equal to Q only when $\Sigma_i\, y_i = 1$, which is the case of probabilistic credence $\mathbf{y}$.

18 Let us set aside the quibble that considerations of strict dominance in accuracy have been replaced by considerations of expected accuracy. That weakens the whole argument since maximizing expectations is not automatically always the best.

If this last vindication is unsatisfactory, might we find another? Pettigrew (2016, chap. 4) offers another vindication of strictly proper scoring rules. The analysis depends on positing several conditions on an inaccuracy measure that include what he calls "Divergence Additivity," "Divergence Continuity," and "Decomposition." We find once again that these conditions are congenial for a probabilist who knows that they will yield the required result. They appear arbitrary, however, to someone not antecedently committed to probabilities.

Divergence Additivity requires that the inaccuracy of some set of credences $<x_1, x_2, \ldots, x_r>$ is measured by taking the arithmetic sum of the inaccuracies of the individual credences, using $g_1(x_i)$ or $g_0(x_i)$, according to whether the credence x_i is in the true state or not. Summation seems, initially, to be an innocent requirement. Pettigrew (p. 49) calls the summation "the natural thing to do." But it is far from innocent, for it represents a particular rule for determining the import of variation among individual inaccuracy measures. Take the case of five credences $r = 5$ and assume that we have two different sets of inaccuracies provided by the functions $g_1(x_i)$ or $g_0(x_i)$:

$$0.1, 0.1, 0.1, 01, 0.1 \text{ and } 0.01, 0.01, 0.01, 0.01, 0.46.$$

How are we to summarize the combined inaccuracy in each case? Is the combined inaccuracy of the first the same as the second? Or does the presence of the large inaccuracy 0.46 in the second render the second case more inaccurate than the first? Or is this second case less inaccurate since four of its five components are very small, 0.01? Divergence Additivity measures the combined inaccuracy by summing the components. Since the components in both cases sum to 0.5, this condition judges them equal in combined inaccuracy. This is quite a specific way to trade off the import of non-uniformities of the second case. Since it competes with many other possible ways of trading off non-uniformities, merely finding it "natural" falls well short of the independent justification needed.

Similar arbitrariness troubles the other two conditions. Briefly, Divergence Continuity requires the analogs of the functions $g_1(x)$ or $g_0(x)$ to be continuous in x. In the abstract, the requirement seems innocent. However, requirements of continuity can be far from innocent. In

geometry, we might think it innocent to require that some two-dimensional surface be covered continuously by the familiar $<x, y>$ coordinate system. However, this condition restricts us to surfaces that are topologically "R^2," precluding surfaces of spheres and toruses, even though both are, in a geometric sense, everywhere continuous. Finally, Decomposition arises from two further conditions, Calibration and Truth-Directedness, each of which, independently, looks quite natural. The difficulty is that these two conditions turn out to be incompatible, so that at least one is wrong. Once again, naturalness proves to be a poor guide. Decomposition is a compromise condition that attempts to mediate between them. We may well wonder why it is a good idea to mediate between two conditions, one or both of which might be wrong. The mediation uses a formula that in turn appears arbitrary, unless one knows that it will enable a demonstration of the result sought.

All of these efforts end up offering no escape from the problem that has dogged the accuracy-based vindication of probabilities from the start. We are trapped in an endless regress of reasons. The requirement of accuracy alone, it turns out, gives us very little. What really determines the outcome is our choice of scoring rule. Even among n-power scoring rules, we can select any desired extent of superadditivity or subadditivity of our credences just by choosing a suitable n. If we are to vindicate a restriction to probabilistic credences, we must find further reasons that favor them. We find new reasons that seem natural; and then we realize that they are only natural if judged by our antecedent prejudice for probabilistic credences. Still further reasons are needed, and the regress of reasons proceeds.

11.11. Naturalness Gone Astray

Reinhard Selten (1998) provides a sobering illustration of the precariousness of accepting conditions on the basis of their naturalness. His interest is what he calls "the quadratic scoring rule." It is used in something like an elicitation context in which a predicted probability distribution x is scored against a true probability distribution p by means of the "expected score loss." His quadratic scoring rule is given in one form (p. 48) as

$$L(\mathbf{x} \mid \mathbf{p}) = \sum_{i=1,r} (x_i - p_i)^2,$$

where the two distributions $\mathbf{x} = <x_1, \ldots, x_r>$ and $\mathbf{p} = <p_1, \ldots, p_r>$ adopt the indexed values x_i and p_i over outcomes $i = 1, \ldots, r$. Selten (p. 43) reports: "As far as the author knows, Brier (1950) was the first one who described this rule."

This scoring rule formula differs from the Brier score formula given above as (2a). The difference is easy to see, since the reference probability p_i of formula (2a) enters linearly into the expression, whereas the true probability p_i of Selten's formula enters as a quadratic. A simple relation, however, connects the two formulae and is such that this formality does not make a difference to their common function of eliciting distributions. If we write the Brier formula (2a) as "$B(\mathbf{x} \mid \mathbf{p})$," then we have $L(\mathbf{x} \mid \mathbf{p}) = B(\mathbf{x} \mid \mathbf{p}) - B(\mathbf{p} \mid \mathbf{p})$.[19]

The principal result of Selten's paper is a demonstration that its four axioms are satisfied uniquely by the quadratic scoring rule. This uniqueness is a strong result. Selten goes to some pains to justify the naturalness of what might be the most contentious of the axioms, the fourth axiom, "neutrality." It requires that the loss function L be symmetric in the two distributions:

$$L(\mathbf{p} \mid \mathbf{x}) = L(\mathbf{x} \mid \mathbf{p})$$

Selten's plea for the axiom is strong and plausible:

> The interpretation of axiom 4 becomes clear if one looks at the hypothetical case that one and only one of two theories p and q is right, but it is not known which one. The expected score loss of the wrong theory is a measure of how far it is from the truth. It is only fair to require that this measure is "neutral" in the sense that it treats both theories equally. If p is wrong and q is right, then p should be considered to be as far from the truth as q in the opposite case that q is wrong and p is right.

19 I thank an anonymous reviewer for BSPS*Open* for pointing out this connection.

> A scoring rule should not be prejudiced in favor of one of both theories in the contest between p and q. The severity of the deviation between them should not be judged differently depending on which of them is true or false.
>
> A scoring rule which is not neutral is discriminating on the basis of the location of the theories in the space of all probability distributions over the alternatives. Theories in some parts of this space are treated more favorably than those in some other parts without any justification. Therefore, the neutrality axiom 4 is a natural requirement to be imposed on a reasonable scoring rule. (p. 54)

It is easy to accept this plea and, with it, neutrality as a reasonable demand for any scoring rule. It does seem natural. The comfort will surely be disturbed when one realizes that Selten's naturalness requirement eliminates virtually all the many, strictly proper scoring rules discussed above. All that remains is Selten's quadratic rule. This elimination might be justifiable if Selten's analysis had found some inadequacy in the function of all of these other strictly proper scoring rules. His analysis shows no such inadequacy; and it cannot. For all strictly proper scoring rules are by explicit design adequate to their function of eliciting a probabilistic credence.

The symmetry of Selten's neutrality condition does not derive from the function of the scoring rule. Rather, it calls to mind the idea that distances between two points in ordinary space treat the two points symmetrically. Distance AB in a geometric space is the same as distance BA. It is easy to assent to the corresponding symmetry in the context of Selten's analysis, since his scoring rule is, in a more abstract sense, a measure of the distance separating two distributions in a probability space. However, there is a difference from the geometric case. Distances in geometry must treat the points A and B symmetrically, since the notion of distance itself does not distinguish or privilege one point over another. There is no corresponding symmetry in the two distributions, p and x. One is the true probability distribution; the other is an elicited distribution. In terms of function of the scoring rule, there is no need to treat them symmetrically. Asymmetric strictly proper scoring rules still serve their function of elicitation.

It is of course "natural" to treat them symmetrically. What results is an appealing simplification. We reduce the many scoring rules possible to a unique rule with a simple expression. That simple rule is likely easier to work with computationally than many of the more complicated strictly proper scoring rules. There is an aesthetic comfort in the formula. Its symmetry is visible from inspection, and we can see without calculation that it takes a minimum value just when our elicited distribution x_i coincides with the true distribution p_i.

However, the appeal of these factors should not lead us to think of the naturalness condition as anything more than an aesthetically motivated restriction unrelated to the rules' function. It establishes no necessity for the quadratic scoring rule.

11.12. Conclusion

What makes the circularity of this accuracy-based approach harder to see at the outset is that it draws on a well-established literature on scoring rules in meteorology, economics, and subjective Bayesianism. This literature developed scoring rules for other purposes. They were used to reward meteorologists for their probabilistic predictions when scored against the actual frequencies of weather conditions; or they were used to encourage subjects to match their publicly declared probabilities with their true but hidden probabilities. For these purposes, it was appropriate to work with a narrow subset of scoring rules, adapted antecedently to probability measures. Using different rules, ill-adapted to probabilities would have no point.

Matters change when we try to use scoring rules to demonstrate the necessity of probabilities. Now, the careful selection of the same scoring rules ceases to be the practical adaption of the rules to the intended use. It amounts to the covert assumption of the very thing that is to be proven. For these favored rules—the Brier score and its generalization as strictly proper scoring rules—strongly favor probabilistic credences. As we saw above, if a subject harbors non-probabilistic credences and these scoring rules are used to elicit them, the subject will be rewarded for lying and reporting probabilistic credences.

All would be well with accuracy-based vindications if solid, independent grounds could be found for use of these favored rules. However, no such grounds have emerged and, I argue, none can emerge. For all such grounds must covertly assume exactly what they seek to demonstrate. Instead, as we have seen repeatedly, the grounds will succumb under scrutiny. We are forever trapped in an endless regress of reasons.

Appendix 11.A. Dominance Relations for n-Power Scoring Rule with $n > 1$

The n-power loss functions

$$L_1 = (1 - x_1)^n + x_2^n + x_3^n + \ldots + x_r^n$$

$$L_2 = x_1^n + (1 - x_2)^n + x_3^n + \ldots + x_r^n$$

$$\ldots$$

$$L_r = x_1^n + x_2^n + x_3^n + \ldots + (1 - x_r)^n \tag{4a}$$

admit dominating points that lie on an $r - 1$ dimensional hypersurface of the r dimensional space of credences $x_1, x_2, \ldots, x_r$. Each point on the surface is a minimum for all r loss functions among a set of points lying on a curve in the space of credences. We write this curve as $x_i(\lambda)$, $I = 1, \ldots, r$, where λ is a path parameter. A dominance point is identified by means of the derivatives of the loss functions with respect to λ. The first derivatives are

$$\frac{dL_1}{d\lambda} = -n(1-x_1)^{n-1}\frac{dx_1(\lambda)}{d\lambda} + nx_2^{\,n-1}\frac{dx_2(\lambda)}{d\lambda} + \ldots + nx_r^{\,n-1}\frac{dx_r(\lambda)}{d\lambda} \tag{13}$$

and similarly for $L_2, \ldots, L_r$. The second derivatives are

$$\begin{aligned}\frac{d^2L_1}{d\lambda^2} = n(n-1)(1-x_1)^{n-2}\left(\frac{dx_1(\lambda)}{d\lambda}\right)^2 - n(1-x_1)^{n-1}\frac{d^2x_1(\lambda)}{d\lambda^2} \\ +n(n-1)x_2^{\,n-2}\left(\frac{dx_2(\lambda)}{d\lambda}\right)^2 + nx_2^{\,n-1}\frac{d^2x_2(\lambda)}{d\lambda^2} + \cdots \\ \ldots + n(n-1)x_r^{\,n-2}\left(\frac{dx_r(\lambda)}{d\lambda}\right)^2 + nx_r^{\,n-1}\frac{d^2x_r(\lambda)}{d\lambda^2}\end{aligned} \tag{14}$$

and similarly for $L_2, \ldots, L_r$. To identify a dominance point, we set all the first derivatives (13) to zero. The results for $dL_1/d\lambda = 0$ and $dL_i/d\lambda = 0$ are, respectively,

$$-(1-x_1)^{n-1}\frac{dx_1}{d\lambda}+\ldots+x_i^{\,n-1}\frac{dx_i}{d\lambda}+\ldots+x_r^{\,n-1}\frac{dx_r}{d\lambda}=0 \tag{15}$$

$$x_1^{\,n-1}\frac{dx_1}{d\lambda}+\ldots-(1-x_i)^{n-1}\frac{dx_i}{d\lambda}+\ldots+x_r^{\,n-1}\frac{dx_r}{d\lambda}=0$$

Subtracting the second from the first, we recover

$$\frac{dx_i/d\lambda}{dx_1/d\lambda}=\frac{\left[x_1^{\,n-1}+(1-x_1)^{n-1}\right]}{\left[x_i^{\,n-1}+(1-x_i)^{n-1}\right]} \tag{16}$$

This expression (16), with $i = 2, 3, \ldots, r$ can be used to replace expressions for $dx_2/d\lambda, dx_3/d\lambda, \ldots, dx_r/d\lambda$ in (15), rewritten as

$$(1-x_1)^{n-1}=x_2^{\,n-1}\frac{dx_2/d\lambda}{dx_1/d\lambda}+\ldots+x_i^{\,n-1}\frac{dx_i/d\lambda}{dx_1/d\lambda}+\ldots+x_r^{\,n-1}\frac{dx_r/d\lambda}{dx_1/d\lambda}\,.$$

After some manipulation, the reconfigured equation (15) reduces to the expression that identifies the $r - 1$ dimensional hypersurface of dominance points:

$$1=\frac{x_1^{\,n-1}}{\left[x_1^{\,n-1}+(1-x_1)^{n-1}\right]}+\ldots+\frac{x_i^{\,n-1}}{\left[x_i^{\,n-1}+(1-x_i)^{n-1}\right]}+\ldots+\frac{x_r^{\,n-1}}{\left[x_r^{\,n-1}+(1-x_r)^{n-1}\right]} \tag{12}$$

In the special case of $n = 2$, the Brier score, this relation identifies the hypersurface of additive credences that conform with the probability calculus:[20]

$$1 = x_1 + \ldots + x_i + \ldots + x_r.$$

To determine the disposition of the hypersurfaces of the remaining cases, we write the individual terms of (12) as

$$y_i=\frac{x_i^{\,n-1}}{\left[x_i^{\,n-1}+(1-x_i)^{n-1}\right]}\,.$$

20 For this case, $n - 1 = 1$ and $x_i^{n-1} + (1 - x_i)^{n-1} = x_i + (1 - x_i) = 1$.

They can be inverted to yield

$$x_i = \frac{y_i^{1/(n-1)}}{\left[y_i^{1/(n-1)} + (1-y_i)^{1/(n-1)}\right]}, \tag{17}$$

where, following (12), we have

$$1 = y_1 + \ldots + y_i + \ldots + y_r.$$

A special case is $r = 2$, for any $n > 1$. For then $y_2 = (1 - y_1)$ we have

$$x_1 = \frac{y_1^{1/(n-1)}}{\left[y_1^{1/(n-1)} + (1-y_1)^{1/(n-1)}\right]} = \frac{y_1^{1/(n-1)}}{\left[y_1^{1/(n-1)} + y_2^{1/(n-1)}\right]}$$

$$x_2 = \frac{y_2^{1/(n-1)}}{\left[y_2^{1/(n-1)} + (1-y_2)^{1/(n-1)}\right]} = \frac{y_2^{1/(n-1)}}{\left[y_2^{1/(n-1)} + y_1^{1/(n-1)}\right]}$$

so that the dominance points are also additive: $1 = x_1 + x_2$.

Otherwise, for $r > 2$ and $n > 2$, we have from (17) that

$$x_1 > \frac{y_1^{1/(n-1)}}{y_1^{1/(n-1)} + y_2^{1/(n-1)} + \ldots + y_r^{1/(n-1)}},$$

since

$$(1 - y_1)^{1/(n-1)} = (y_2 + \ldots + y_r)^{1/(n-1)} < y_2^{1/(n-1)} + \ldots + y_r^{1/(n-1)} \tag{18}$$

by means of inequality (23) below. Using similar relations for $x_2, x_3, \ldots, x_r$, we recover

$$x_1 + x_2 + \ldots + x_r > \frac{y_1^{1/(n-1)} + y_2^{1/(n-1)} + \ldots + y_r^{1/(n-1)}}{y_1^{1/(n-1)} + y_2^{1/(n-1)} + \ldots + y_r^{1/(n-1)}} = 1.$$

It follows that $r > 2$ and $n > 2$ is the case of subadditive credences. Repeating the above analysis for $r > 2$ and $1 < n < 2$, using inequality (24), we recover:

$$x_1 + x_2 + \ldots + x_r < 1,$$

from which it follows that this is the case of superadditive credences.

The hypersurface (12) is picked out by the vanishing of the first derivatives, $dL_1/d\lambda = dL_2/d\lambda = \ldots = dL_r/d\lambda = 0$ for the curves $x_i(\lambda)$, $i = 1, \ldots, r$. To complete the analysis, we need to show that these points are true

minima for the loss functions along the curves, so that the points on the hypersurface are dominance points. This in turn requires identification of the curves.

It will be sufficient to identify one set of curves as follows.[21] In brief, we find the slope of the curve at each point on the hypersurface. We then take as the curve $x_i(\lambda)$ through that point, the straight line that has this slope as its slope everywhere. Select some point on the hypersurface, whose credences X_i satisfy equation (12). We have from (16) that

$$\frac{dx_i}{d\lambda} = \frac{K}{\left[X_i^{n-1} + (1 - X_i)^{n-1}\right]},$$

where K is some undetermined constant that is the same for all $x_i(\lambda)$. The constant is undetermined since its differing values give us the freedom to rescale the parameter λ arbitrarily. We can, for example, alter the value of K if we introduce a new parameterization $\lambda'(\lambda)$ for which

$$\frac{dx_i}{d\lambda'} = \frac{dx_i}{d\lambda} \cdot \frac{d\lambda}{d\lambda'}.$$

To ensure that the path parameterization introduces no nuisance pathologies, it is convenient to set it, by stipulation, proportional to the natural Euclidean path length through

$$d\lambda^2 = \text{constant} \cdot (dx_1{}^2 + dx_2{}^2 + \ldots + dx_r{}^2).$$

We select the constant in this expression so that the undetermined constant K is set to one. That is, we now have

$$\frac{dx_i}{d\lambda} = \frac{1}{\left[X_i^{n-1} + (1 - X_i)^{n-1}\right]} = m_i(X_1, \ldots, X_r) > 0, \tag{19}$$

where $m_i > 0$ since $0 \leq X_i \leq 1$ for all i. The straight line with this slope m_i that passes through the hypersurface point X_i at $\lambda = 0$ is

21 The properties described above do not, I suspect, uniquely define the curves $x_i(\lambda)$. Identifying one set of curves is sufficient to display the dominance properties of the points of the hypersurface.

$$x_i(\lambda) = m_i\lambda + X_i.$$

For all such curves, we have

$$\frac{dx_i}{d\lambda} = m_i > 0 \text{ and } \frac{d^2x_i}{d\lambda^2} = \frac{dm_i}{d\lambda} = 0 \qquad i = 1, \ldots, r.$$

Substituting these properties into the r expressions for $d^2L_i/d\lambda^2$, $I = 1, \ldots, r$, analogous to (14), and recalling $n > 0$, it is easy to see that all the second derivative terms are greater than zero. Hence the point of intersection of each curve X_i with the hypersurface (12) is a true minimum along each curve for all the loss functions $L_1, \ldots, L_r$.

Appendix 11.B. Credences Elicited by n-Power Scoring with $n > 1$

The n-power scoring rule is generated by the functions (5a). The credences $\mathbf{x} = \langle x_1, x_2, \ldots, x_r \rangle$ it elicits for a subject's true probabilistic credences $\mathbf{p} = \langle p_1, p_2, \ldots, p_r \rangle$ are those that minimize the loss function.

$$\begin{aligned} L(\mathbf{p}, \mathbf{x}) = \; & p_1(1 - x_1)^n + \ldots + p_1 x_i^n + \ldots + p_1 x_r^n \\ & + \ldots \\ & + p_i x_1^n + \ldots + p_i (1 - x_i)^n + \ldots + p_i x_r^n \\ & + \ldots \\ & + p_r x_1^n + \ldots + p_r x_i^n + \ldots + p_r (1 - x_r)^n \end{aligned} \tag{10b}$$

To keep the analysis simple, consider only the generic case in which $p_i > 0$, all i. The first and second derivatives of $L(\mathbf{p}, \mathbf{x})$ with respect to x_1 are

$$\frac{\partial L}{\partial x_1} = -p_1 n(1 - x_1)^{n-1} + (p_2 + \ldots + p_r) n x_1^{n-1} = -p_1 n(1 - x_1)^{n-1} + (1 - p_1) n x_1^{n-1}$$

$$\frac{\partial^2 L}{\partial x_1^2} = p_1 n(n-1)(1 - x_1)^{n-2} + (1 - p_1) n(n-1) x_1^{n-2}$$

and similarly for $x_1, \ldots, x_r$. We seek the minimum loss with respect to x by setting all first derivatives to zero. We find for $i = 1, \ldots, r$, that $\partial L/\partial x_i = 0$ leads to

$$\left(\frac{x_i}{1-x_i}\right)^{n-1} = \left(\frac{p_i}{1-p_i}\right).$$

The values selected by this condition represent a true minimum since $\partial^2 L/\partial x_i^2 > 0$ for $0 \le x_i \le 1$, for all i. Solving for x_i, the credences elicited are

$$x_i = \frac{(p_i)^{1/(n-1)}}{(p_i)^{1/(n-1)} + (1-p_i)^{1/(n-1)}} \quad . \tag{20}$$

The credences elicited will correspond to probabilities p_i only in the case of the Brier rule, $n = 2$. For then we have

$$x_i = \frac{(p_i)^{1/(2-1)}}{(p_i)^{1/(2-1)} + (1-p_i)^{1/(2-1)}} = \frac{(p_i)}{(p_i) + (1-p_i)} = p_i \, .$$

When n is not 2, but $r = 2$, the rule will return additive credence x_1 and x_2:

$$x_1 = \frac{(p_1)^{1/(n-1)}}{(p_1)^{1/(n-1)} + (1-p_1)^{1/(n-1)}} \text{ and } x_2 = \frac{(p_2)^{1/(n-1)}}{(p_1)^{1/(n-1)} + (1-p_2)^{1/(n-1)}} \, .$$

These elicited credences x_1 and x_2 will not correspond to the probabilities p_1 and p_2 unless we have the exceptional cases of $p_1 = 0$ or $p_1 = 0.5$ or $p_1 = 1$.

In all other cases for $n > 1$, we recover subadditive credences (for $n > 2$) or superadditive credences (for $1 < n < 2$).

To begin, consider the case of $n > 2$. For $r > 2$, we have from inequality (23) below that:

$$(p_2 + p_3 + \ldots + p_r)^{1/(n-1)} < (p_2)^{1/(n-1)} + (p_3)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)} \, . \tag{21}$$

Using $1 - p_1 = p_2 + \ldots + p_r$, it becomes

$$(1-p_1)^{1/(n-1)} < (p_2)^{1/(n-1)} + (p_3)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)} \, .$$

Substituting into (20) for the case of $i = 1$, we have

$$x_1 = \frac{(p_1)^{1/(n-1)}}{(p_1)^{1/(n-1)} + (1-p_1)^{1/(n-1)}} > \frac{(p_1)^{1/(n-1)}}{(p_1)^{1/(n-1)} + (p_2)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)}} \, .$$

with similar formulae for $x_2, \ldots, x_r$. We see that these credences are subadditive if we sum them:

$$x_1 + x_2 + \ldots + x_r > \frac{(p_1)^{1/(n-1)} + (p_2)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)}}{(p_1)^{1/(n-1)} + (p_2)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)}} = 1\,,$$

where the credence in the set of all outcomes is 1. For the case of $1 < n < 2$, using (24) below, we have, instead of (21), the inequality:

$$(p_2 + p_3 + \ldots + p_r)^{1/(n-1)} > (p_2)^{1/(n-1)} + (p_3)^{1/(n-1)} + \ldots + (p_r)^{1/(n-1)}\ . \tag{22}$$

Following analogous reasoning, we arrive at superadditive credences

$$x_1 + x_2 + \ldots + x_r < 1.$$

Appendix 11.C. Useful Inequalities

The equalities used above are derived by considering the function

$$f(x) = (x + y)^{1/(n-1)} - x^{1/(n-1)} - y^{1/(n-1)}$$

for some fixed value of $y > 0$. Its first derivative is

$$\frac{df(x)}{dx} = \tfrac{1}{n-1}\left((x+y)^{(2-n)/(n-1)} - x^{(2-n)/(n-1)}\right).$$

For $n > 2$, the exponent satisfies $-1 < (2 - n)/(n - 1) < 0$. It follows that $df(x)/dx < 0$ for all $x > 0$. Since $f(0) = 0$, we have after integration of $df(x)/dx$ that $f(x) < 0$. That is, for all $x > 0$ and $y > 0$, $n > 2$,

$$(x + y)^{1/(n-1)} < x^{1/(n-1)} + y^{1/(n-1)}$$

Applying this inequality to $(z_2 + z_3 + \ldots + z_r)^{1/(n-1)}$ for all $z_i > 0$, we recover

$$(z_2 + z_3 + \ldots + z_r)^{1/(n-1)} < (z_2 + z_3 + \ldots + z_{r-1})^{1/(n-1)} + (z_r)^{1/(n-1)}$$

and then

$$(z_2 + z_3 + \ldots + z_{r-1})^{1/(n-1)} + (z_r)^{1/(n-1)} < (z_2 + z_3 + \ldots + z_{r-2})^{1/(n-1)} + (z_{r-1})^{1/(n-1)} + (z_r)^{1/(n-1)}$$

Further iteration eventually leads to:

$$(z_2 + z_3 + \ldots + z_r)^{1/(n-1)} < (z_2)^{1/(n-1)} + (z_3)^{1/(n-1)} + \ldots + (z_r)^{1/(n-1)}. \tag{23}$$

For $1 < n < 2$, we have that the exponent in $f(x)$ satisfies $(2 - n)/(n - 1) > 0$. Proceeding as before we now have

$$(x + y)^{1/(n-1)} > x^{1/(n-1)} + y^{1/(n-1)},$$

which eventually leads to:

$$(z_2 + z_3 + \ldots + z_r)^{1/(n-1)} > (z_2)^{1/(n-1)} + (z_3)^{1/(n-1)} + \ldots + (z_r)^{1/(n-1)}. \tag{24}$$

Appendix 11.D. Equivalent Definitions of Strictly Proper Scoring Rules

To show the equivalence of the two definitions I and II of strictly proper scoring rules, it is sufficient to show that definition II entails definition I; and to show the converse entailment.

Strictly Proper II entails Strictly Proper I

The loss function $L(\mathbf{p}, \mathbf{x})$ of (10a) consists of a sum of r terms:

$$p_1\, g_0(x_i) + \ldots + p_i\, g_1(x_i) + \ldots + p_r\, g_0(x_i),$$

where $i = 1, \ldots, r$. Definition II entails that each of these r terms individually is minimized when $x_i = p_i$. To illustrate for $I = 1$, the term is rewritten as

$$\begin{aligned} &p_1\, g_1(x_1) + p_2\, g_0(x_1) \ldots + p_i\, g_0(x_1) + \ldots + p_r\, g_0(x_1) \\ &\quad = p_1\, g_1(x_1) + (p_2 + \ldots + p_i + \ldots + p_r)\, g_0(x_1) \\ &\quad = p_1\, g_1(x_1) + (1 - p_1)\, g_0(x_1) \end{aligned}$$

Hence, this term is minimized uniquely, according to definition II, when $x_1 = p_1$. The corresponding results for the remaining $x_2, x_3, \ldots$ follow analogously. Since $\mathbf{x} = \mathbf{p}$ minimizes each term uniquely, it follows that $\mathbf{x} = \mathbf{p}$ minimizes their sum, $L(\mathbf{p}, \mathbf{x})$, uniquely, which is definition I.

Strictly Proper I entails Strictly Proper II

Definition I applies for all p_i in $0 \le p_i \le 1$, $i = 1, \ldots, r$. Thus it applies to the case in which only $p_1 > 0$ and $p_2 > 0$, but $p_3 = p_4 = \ldots = p_r = 0$. In this special case, the loss function reduces to

$$L(\mathbf{p}, \mathbf{x}) = p_1\, g_1(x_1) + p_1\, g_0(x_2) + \ldots + p_1\, g_0(x_i) + \ldots + p_1\, g_0(x_r)$$
$$+ p_2\, g_0(x_1) + p_2\, g_1(x_2) + \ldots + p_2\, g_0(x_i) + \ldots + p_2\, g_0(x_r).$$

There are no terms in $L(\mathbf{p}, \mathbf{x})$ in $g_1(x_3), g_1(x_4), \ldots, g_1(x_4)$, but these variables only appear in $g_0(x_3), g_0(x_4), \ldots, g_0(x_r)$. Since all suitable functions for $g_0(x_i)$ are strictly increasing, the condition for minimization must include $x_i = 0 = p_i$, for $i = 3, 4, \ldots, r$. Hence, the minimization of definition I reduces to the simpler problem of minimizing:

$$L(p_1, p_2, x_1, x_2) = p_1\, g_1(x_1) + p_1\, g_0(x_2)$$
$$+ p_2\, g_0(x_1) + p_2\, g_1(x_2).$$

That is, definition I requires minimization for fixed p_1 and p_2 of:

$$L(p_1, p_2, x_1, x_2) = p_1\, g_1(x_1) + (1 - p_2)\, g_0(x_2)$$
$$+ (1 - p_1)\, g_0(x_1) + p_2\, g_1(x_2).$$

Definition I stipulates that the minimum is achieved uniquely when $x_1 = p_1$ and $x_2 = p_2$. Since x_1 and x_2 can be varied independently in seeking the minimum, the minimum can only arise when the terms in which they appear

$$p_1\, g_1(x_1) + (1 - p_1)\, g_0(x_1) \quad \text{and} \quad p_2\, g_1(x_2) + (1 - p_2)\, g_0(x_2)$$

are individually, uniquely minimized by $x_1 = p_1$, for the first, and $x_2 = p_2$, for the second.

Either of these is equivalent to definition II, with the restriction that $0 < p < 1$. The complete definition II allows $0 \leq p \leq 1$. The two missing cases, $p = 0$ and $p = 1$, always conform with definition II. Hence, definition I entails definition II.

REFERENCES

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review* 78(1): pp. 1–3.

Brier, Glenn W. and Roger A. Allen. 1951. "Verification of Weather Forecasts." In *Compendium of Meteorology*, edited by T. F. Malone, pp. 841–48. Boston: American Meteorological Society.

De Finetti, Bruno. 1965. "Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item." *The British Journal of Mathematical and Statistical Psychology* 18: pp. 87–123.

———. 1974. *Theory of Probability: A Critical Introductory Treatment*. Vol. 1. Chichester: John Wiley & Sons.

Gneiting, Tilmann, Adrian T. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102: pp. 359–78.

Joyce, James. 1998. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science* 65: pp. 575–603.

———. 2009. "Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief." In *Degrees of Belief*, edited by F. Huber and C. Schmidt-Petri, pp. 263-97. Dordrecht: Springer Netherlands.

Leitgeb, Hannes and Richard Pettigrew. 2010. "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy." *Philosophy of Science* 77: pp. 236–72.

McCarthy, John. 1956. "Measures of the Value of Information." *Proceedings of the National Academy of Sciences* 42(9): pp. 654–55.

Pettigrew, Richard. 2016. *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Predd, Joel B. et al. 2009. "Probabilistic Coherence and Proper Scoring Rules." *IEEE Transactions of Information Theory* 55: pp. 4786–92.

Rosenkrantz, Roger D. 1981. *Foundations and Applications of Inductive Probability*. Atascadero, CA: Ridgeview Publishing.

Savage, Leonard J. 1971. "Elicitation of Personal Probabilities and Expectations." *Journal of the American Statistical Association* 66: pp. 783–801.

Schervish, Mark. 1989. "A General Method for Comparing Probability Assessors." *The Annals of Statistics* 17: pp. 1856–79.

Schervish, Mark, Teddy Seidenfeld, and Joseph Kadane. 2009. "Proper Scoring Rules, Dominated Forecasts and Coherence." *Decision Analysis* 6: pp. 202–21.

Selten, Reinhard. 1998. "Axiomatic Characterization of the Quadratic Scoring Rule," *Experimental Economics* 1: pp. 43–62.