Elicitation and Evaluation of Statistical Forecasts *

Nicolas S. Lambert†

June 2019

Abstract

An expert has full or partial probabilistic information about a random state. The expert is asked to make a prediction regarding a property of the state distribution, that is, to answer a question about the distribution. The expert receives a payoff that may depend on his own report and the subsequently realized state. For which properties can a payoff rule be devised so as to induce the expert, as a strict best response, to answer the truth? In a finite world, the payoff rules that provide strict incentives to the expert exist if and only if the property partitions the simplex of distributions into a power diagram. These payoff rules can be fully characterized as weighted averages of elementary payoff functions. They can be used both as an incentive device and to evaluate the performance of forecasters.

KEYWORDS: Elicitability, property elicitation, forecast testing.

*This work is a revision of my job market paper and chapters 2–4 of my doctoral dissertation, based on the conference paper "Eliciting Truthful Answers to Multiple-Choice Questions" (with Yoav Shoham) presented at the 10th ACM Conference on Economics and Computation and which received an ACM EC Best Paper Award. Appendix B is based on "Eliciting Properties of Probability Distributions" (with David Pennock and Yoav Shoham) presented at the 9th ACM Conference on Economics and Computation. I am deeply indebted to David Pennock and Yoav Shoham for their guidance and invaluable advice throughout this project. I am grateful to Christopher Chambers, Federico Echenique, Rafael Frongillo, David Kreps, Mohammad Mahdian, Michael Ostrovsky, Andrzej Skrzypacz, William Thomson, and Johanna Ziegel for helpful discussions. I thank many seminar and conference participants for useful comments. A significant part of this work was developed while I was visiting Yahoo! Research and Microsoft Research. Finally, I gratefully acknowledge financial support from Google Research and the National Science Foundation under grant No. CCF-1101209.

†Stanford Graduate School of Business, Stanford University; nlambert@stanford.edu.

1 Introduction

A decision maker often has less information relevant for her decision than does some other agent. In this paper, I examine protocols for eliciting and evaluating information provided by an expert when such information consists of statistical forecasts, or, as defined formally in the paper, properties of probability distributions. Given a finite set of possible states Ω, a random state is drawn according to some probability distribution P. The expert privately observes P or observes some information about P, and is asked to announce the value of a property of interest regarding P. The term *property* has a fairly general meaning, it captures arbitrary features that a distribution can possess. Classical real-valued properties include the mean and median of a variable, measures of dispersion such as the variance, risk measures such as value at risk and expected shortfall. Properties can be multidimensional, for example to represent a confidence interval or a variance-covariance matrix. They need not be numerical, for example they can record the principal components of a random vector, or capture an ordering of events from the most to the least likely.

The protocols or mechanisms considered in this paper are (general) scoring rules, defined by analogy to the classical probability scoring rules developed by Brier (1950), Good (1952), McCarthy (1956), De Finetti (1962) and Savage (1971), among others. A scoring rule assigns a payoff to the expert as a function of his prediction and the state realization. In a large part of the paper, the focus is on strictly proper scoring rules, that is, scoring rules such that the expert, whose payoff is equal to the obtained score, reports truthfully as a strict best response when sufficiently informed. This benchmark is classical in the literature and is motivated by the idea that, in applications, when the expert is indifferent between two responses, we lose the ability to distinguish between a correct and an incorrect expert or to motivate an individual whose access to information is costly. I expand on these points in Section 5 and Appendix A.

I address two central questions. First, for which properties does a strictly proper scoring rule exist? And second, fixing the property, how can we construct such scoring rules, and how can we characterize them? Interestingly, many properties are not "elicitable," i.e., a strictly proper scoring rule does not exist: the declared predictions do not supply enough information to enable the enforcement of strict incentives. However, there are also a number of relevant cases for which they do.

The main body of the paper focuses on properties that take finitely many possible values, which is both tractable and allows for a range of applications (noting that if a property takes a continuum of values, the values can always be partitioned into bins which makes it finite); the case of properties that take a continuum of values is relegated to Appendix B. Properties

are elicitable precisely when they partition the simplex of distributions into a power diagram, a geometric object based on the notion of nearest neighbors. The characterization implies, for example, that the mean, mode, median, some risk measures such as the value at risk, the ranking of events from the most to the least likely are elicitable, whereas measures of dispersion and symmetry such as the variance, skewness, kurtosis, other risk measures such as the expected shortfall, confidence intervals, the pair of most correlated elements of a random vector are not elicitable. If a property is not elicitable then it means that, to motivate the expert, or to evaluate his performance, we must ask for more information, and so the expert must know more. The proper and strictly proper scoring rules are generated by the mixtures of some given baseline functions that are entirely determined by the property itself. I provide several examples of such constructions.

For properties that take values in a set endowed with a natural ordering of its elements, such as the median of a random variable, an alternative class of scoring rules is introduced, the order-sensitive scoring rules. Order sensitivity means that the closer the estimate is to the true value of the property of interest (in terms of its rank in the ordering) the larger the expected payoff. For a strictly proper scoring rule to be strictly order sensitive, the property must partition the distributions into "slices." For example, a strictly order-sensitive scoring rule can be designed for the median, but not for the mode, even though both properties are elicitable.

The objective of this paper is to propose a general framework. The paper informs us as to what can be asked to an expert from the angle of the provision of incentives, and how to do so. Alternatively, since scoring rules are the negative of loss functions, it informs us as to when we can evaluate the performance of an expert over time, if the expert outputs information on uncertainty. The reason for eliciting one property versus another is outside the scope of the paper: I focus on the elicitation problem and abstract away from what the elicitor does with what is elicited. The elicitor may be a decision maker who collects and aggregates forecasts from different experts to help her choose between alternatives. She may be one of potentially many customers who each confronts different choices, and to whom the expert sells information in his area of expertise. She may be an experimenter in the lab, who wants to ask questions to the subjects of an experiment under uncertainty. In this paper, I simply assume that an elicitor wants to learn some distribution property, and the expert need not know how such information is to be used. For example, in risk management we are typically soliciting estimations of risk measures, and may want to know if more information is needed to supply appropriate incentives or to properly evaluate risk managers. In this case, no more is needed when using the value at risk, while more is needed when using the expected shortfall.

Of course, one may attempt to extract full information by eliciting the entire distribution, from which we can derive the value of any property, but it may not be desirable and, sometimes, it may not be feasible in practice. The expert may be partially informed, and may unable to form a precise estimate of the entire true distribution (or it may be overly costly to do so). For example, in the assessment of the risk of a financial portfolio, or of market and credit risk, one often constraints the analysis to the relevant risk measures. As it turns out, asking for too much information to an expert who does not have this information may lead to erroneous reports that are not even consistent with the expert's information, and from which one cannot back out the expert's actual beliefs. In addition, the state space may be too large and the distribution too complicated. Such environments are common in weather forecasting. For example, Accuweather, one of the largest media company selling commercial weather forecasting services, sells highly refined forecasts, spanning up to 90 days and taking various forms. In such environments, one cannot reasonably consider the full state distribution, but only a small part of it. Similarly, in lab experiments with rich uncertainty, it becomes practically difficult for the experimenter to ask for all the state probabilities, even if theoretically possible; in such situations, the experimenter may want to know what sort of questions she can ask for which she can reward truthful answers.

The paper proceeds as follows. The remainder of this section reviews the literature. Section 2 presents the model. Section 3 begins by presenting the main results under the assumption that the expert is fully informed about the state distribution, and Section 4 considers the general case of partially informed experts. Section 5 demonstrates a simple application of the framework to the problem of testing forecasters. Section 6 concludes. Appendix A motivates the concept of elicitability from the viewpoint of motivating an expert to learn information at a cost. Appendix B considers the case of continuous properties. The remaining appendices include the proofs omitted from the main text.

1.1 Related Literature

The literature on forecast evaluation and elicitation goes back to Brier (1950) and Good (1952). Brier and Good envisioned schemes to measure the accuracy of probability assessments for a set of events, in the context of weather forecasting. These schemes, the quadratic and logarithmic scoring rules, were later recognized as part of a much larger family of functions, the strictly proper probability scoring rules, first axiomatized by McCarthy (1956), De Finetti (1962), and Savage (1971). Over the years, proper scoring rules have been extensively studied (Gneiting and Raftery (2007) provide a survey of the literature). This stream of the literature concerns mostly *probability* scoring rules, whose purpose is to motivate or evaluate the expert

regarding estimates of the entire probability distribution. In contrast, this paper is concerned with the elicitation of more specific information regarding state uncertainty.

Closer to this paper are the works of Fan (1975), Bonin (1976), and especially Thomson (1979), who design compensation schemes to elicit, from local branch managers, the production output that can be attained with some given probability. Relatedly, Savage (1971) designs scoring rules to elicit expectations of random variables, and, in the context of government contracting, Reichelstein and Osband (1984) and Osband and Reichelstein (1985) propose incentive contracts that induce a contracting firm to reveal truthfully moments of its prior about project costs. The most general structure is provided in second chapter of the dissertation of Osband (1985), which allows to elicit more general information embedded within linear partitions of distributions of random vectors. In the recent years, several works on property elicitation have appeared, notably in the statistics and computer science literature. It is difficult to do justice to this stream of literature given its interdisciplinary nature and its ramification to other fields. Notably, in an influential article, Gneiting (2011) discusses elicitable properties and studies a number of cases applied to statistical problems, Abernethy and Frongillo (2012) study the elicitation of linear properties, Frongillo and Kash (2015a) deal with the communication complexity of eliciting properties, Frongillo and Kash (2015b) and Fissler and Ziegel (2016) discuss the elicitability of multidimensional properties and their applications to risk management. Finally, while property elicitation explores the ability to elicit information in the space of distributions, Chambers and Lambert (2018) study elicitation in the time dimension and describe proper scoring rules for dynamic beliefs.

This paper also relates to the stream of the literature about forecast testing. In forecast testing, the question is about how to know if a forecaster is well informed about the probability distribution. Foster and Vohra (1998) initially showed the impossibility of testing forecasters for calibration tests, and the most general results, that apply to essentially any test, were obtained independently by Shmaya (2008) and Olszewski and Sandroni (2008). The manipulability results continue to hold when forecasters are asked to communicate partial information on state distributions under the form of an elicitable property, however it is possible to evaluate their relative performance, as do Al-Najjar and Weinstein (2008) and Feinberg and Stewart (2008) in the context of probability forecasts. I expand on this point in Section 5.

2 Model

There is an expert and a finite set Ω of possible relevant states, where $\Delta(\Omega)$ denotes the set of probability distributions over Ω. For state ω and probability distribution P, $P(\omega)$ is

the probability that ω occurs.

Before the state publicly realizes, an individual, referred to as the elicitor, who could be a decision maker, a manager, or an experimenter, wants to elicit, from the expert, some facts about the state distribution. For example, she may want to ask what is the mean of some random variables, which ones of such and such events is more likely to occur, and so on. These "facts" represent partial information about the state distribution and are captured by distribution properties. Formally, I define a *distribution property*, or simply *property*, as a pair (Θ, F). The first element, Θ, is the *value set* of the property, i.e., a set in which the property takes values. The second element, F, is the *level-set function*, it is a multivalued map from Θ into $\Delta(\Omega)$. The level-set function records, for every property value θ, the collection $F(\theta)$ of all the probability distributions which have property value θ. For example, the mean of a random variable X can be written as a pair $(\mathbf{R}, F)$, where $P \in F(m)$ if, and only if, the mean of X under P, $\int X \, \mathrm{d}P$, equals m. For every $\theta \in \Theta$, $F(\theta)$ must be nonempty. In the main body of the paper, Θ is finite; I refer to these properties as finite properties, or simply properties. The case of continuous Θ is considered in Appendix B and I refer to those properties as continuous properties. Equivalently, with finite properties, the expert is asked to answer some question about the uncertainty over states, and there are finitely many possible answers.

A property may assign a unique value to every probability distribution. If so, the level sets $\{F(\theta), \theta \in \Theta\}$, are pairwise disjoint. These properties are said to have *no redundancy*. Other properties may associate several values to the same distribution. When such is the case, some level sets overlap, and the property has some amount of redundancy. For example, the median exhibits some redundancy, because distributions for random variables may have more than one median. A *property function* is defined as a function Γ that associates some property value $\Gamma(P) \in \Theta$ to every distribution $P \in \Delta(\Omega)$ (formally, the requirement is $\Gamma^{-1}(\theta) \subseteq F(\theta)$ for each θ). When the property has no redundancy, there exists exactly one property function, which then suffices to conveniently represent the property. But in general we need two or more property functions to describe a distribution property.

Throughout I restrict attention to the properties (Θ, F) that satisfy two conditions:

(a) $\bigcup_\theta F(\theta) = \Delta(\Omega)$, which means that the property is well defined for every distribution of $\Delta(\Omega)$;

(b) for all $\theta_1 \neq \theta_2$, $F(\theta_1) \not\subseteq F(\theta_2)$, which means that no property value is purely redundant.

The two assumptions are without loss of generality. All properties can be redefined on the entire set $\Delta(\Omega)$ by assigning a dummy value to the distributions for which it is not originally defined. And as the scoring rules we seek to construct offer the same expected score for all

correct predictions, removing property values that are fully redundant does not impact the analysis.

The payoffs to the expert are specified by mean of scoring rules, whose original definition is a straightforward extension that accounts for the sort of general predictions being considered here. Given a property with value set Θ, a *scoring rule* is a function $S : \Theta \times \Omega \mapsto \mathbf{R}$ that assigns to every prediction θ and every state ω a real-valued score $S(\theta, \omega)$. The scores that scoring rules generate may be interpreted and used in various ways, as long as the expert complies with the general principle that higher expected payoffs are always preferred. In most of this paper I assume that the scoring rule specifies the payoff the expert receives in exchange for his prediction.

The elicitation mechanisms are specified by the property (F, Θ) the elicitor wants to learn and the scoring rule S that assigns payoff values. The timing of events is as follows. First, Nature selects a distribution $P \in \Delta(\Omega)$, which the expert observes. Second, the expert reports a prediction $\theta \in \Theta$. Third, public state ω is drawn at random according to P and the expert gets payoff $S(\theta, \omega)$.

Of course, whether there actually exists such a true state distribution is a matter of interpretation. We may equivalently assume that the expert has subjective beliefs on the state uncertainty, and that he reports according to these subjective beliefs. In the next section, I consider the case of a fully informed expert who knows P, or an expert whose beliefs reduce to a single state distribution; that is, the expert is probabilistically sophisticated. In the section that follows, I consider the case of an expert who forms vague beliefs and is averse to ambiguity with maxmin preferences in the sense defined by Gilboa and Schmeidler (1989). Experts are assumed to be risk neutral, however it is not a binding assumption: if the expert is an expected utility maximizer (with or without ambiguity aversion), then, no matter the expert's utility function, the expert will continue to report as if he were risk neutral when paid in "probability currency."[1]

The objective is to construct scoring rules that induce the expert to provide a correct property value as a best response or as a strict best response. In the terminology of De Finetti (1962) and Savage (1971), such scoring rules are called *proper* and *strict proper*, respectively. That is, proper scoring rules ensure that all true predictions yield the maximum expected payoff under the actual state distribution. Strictly proper scoring rules ensure that the maximum expected payoff is attained if, and only if, the prediction is correct.

[1]If the expert gets score s normalized to be within $[0, 1]$ following his report and the observed state, and if he is given a lottery ticket worth x dollars with probability s and y dollars with probability $1 - s$, for $x > y$, then independently of the utility function, the expert will want to report as he were risk neutral. The random payments over two fixed prizes have the effect of "linearizing" the (generally nonlinear) utility function. It is a common method Savage originally refers to as paying in probability currency.

Definition 1 *A scoring rule S for property (Θ, F) is* proper *if, for every prediction θ and every distribution $P \in \Delta(\Omega)$, whenever θ is true under P, i.e., whenever $P \in F(\theta)$, then*

$$\theta \in \arg\max_{\widehat{\theta} \in \Theta} \mathrm{E}_{\omega \sim P}[S(\widehat{\theta}, \omega)].$$

Scoring rule S is strictly proper *if it is proper and if, for every prediction θ and every distribution $P \in \Delta(\Omega)$, whenever θ is false under P, i.e., whenever $P \notin F(\theta)$, then*

$$\theta \notin \arg\max_{\widehat{\theta} \in \Theta} \mathrm{E}_{\omega \sim P}[S(\widehat{\theta}, \omega)].$$

I abuse notation and use the same symbol for a random element and its realization. To avoid all ambiguity, in expectations, the notation $\mathrm{E}_{\omega \sim \mu}[g(\omega)]$ is used to denote the expected value of g when random element ω is distributed according to μ. In the sequel $S(\theta, P)$ denotes the expected score $\mathrm{E}_{\omega \sim P}[S(\theta, \omega)]$.

One primary objective of this paper is to describe properties that can be elicited truthfully and as a strict best response, as long as the expert knows the state distribution. These properties are said to be "elicitable."

Definition 2 *A property is* elicitable *when there exists a strictly proper scoring rule.*

Even though the concept of elicitability works with fully informed experts, as explained in Section 4, it is not necessary for the expert to know the state distribution exactly for the property to be elicited truthfully.

The concept of properness is concerned with how the expected scores of correct predictions compare with those of incorrect predictions, not with how the expected scores of incorrect predictions compare with one another. Suppose, for example, that we elicit the mean of a random variable, and that the true mean is 100. Proper scoring rules tell us that forecasting mean 100 maximize expected payoffs. But they do not tell us, a priori, how the payoffs from forecasting mean 99 compare to those from forecasting mean 10. In such a case, however, it may be desirable that the expert who reports 99 gets more than the expert who reports 10. This idea is captured by the concept of order sensitivity.

Suppose the value set of the property of interest is ordered. Given a true forecast θ and two incorrect forecasts θ_a, θ_b, whenever θ_a is "in-between" θ and θ_b, θ_a can be viewed as "more accurate" prediction than θ_b according to the ordering of the property values. Order sensitive scores reward forecast θ_a at least as much as forecast θ_b. This concept relates to the notion of scoring rule efficiency of Friedman (1983) and Nau (1985). Scoring rule efficiency compares probabilistic predictions according to their distance to the true distribution, with

respect to some metric. Order sensitivity compares statistical predictions according to their rank relative to the true property value.

Definition 3 *A scoring rule S for a property (Θ, F) is* order sensitive *with respect to the ordering $\prec$ on the value set Θ if, for all distributions P, all forecasts θ true under P, i.e., such that $P \in F(\theta)$, and all forecasts θ_a, θ_b such that either $\theta \preceq \theta_a \prec \theta_b$ or $\theta_b \prec \theta_a \preceq \theta$, $S(\theta_a, P) \geq S(\theta_b, P)$. Scoring rule S is* strictly order sensitive *when the inequality is strict whenever $P \notin F(\theta_b)$.*

Through the remainder of the paper, let $\mathbf{R}^\Omega$ be the space of (real valued) functions on Ω, equivalently, $\mathbf{R}^\Omega$ is the space of random variables. Every state distribution P is an element of $\mathbf{R}^\Omega$ by associating a distribution with its density—since there finitely many states, both concepts are equivalent. Since $\mathbf{R}^\Omega$ is a linear space, I refer to members of the space as vectors, and in this context, distributions are also vectors. The set $\mathbf{R}^\Omega$ is naturally endowed with the scalar product

$$\langle X, Y \rangle = \sum_{\omega \in \Omega} X(\omega) Y(\omega).$$

The expected payoff to the expert who announces θ is then $S(\theta, P) = \langle S(\theta, \cdot), P \rangle$, where $S(\theta, \cdot)$ is the state-contingent payoff.

It is useful to view the set $\Delta(\Omega)$ as a simplex in the Euclidean space $\mathbf{R}^\Omega$, as most results of this paper have a geometric interpretation. In particular, every property has a simple graphical representation as a finite covering of the simplex. In most cases of interest, the covering is a partition except at the boundary points of the level sets. Figure 1 illustrates the case of two properties in the context of weather prediction where the states of interest are snow, rain and shine. Figure 1(a) represents the property associated with the most likely weather state, and Figure 1(b) the property associated with the ordering of weather states from most to least likely.

3 Fully Informed Experts

In this section, I assume the expert is fully informed about the state distribution: he knows P, and responds to incentives so as to maximize his expected payoff. Under full information, elicitable properties are exactly the properties for which an incentive device exists that yields a truthful report as a strict best response, and these incentive devices are captured by strictly proper scoring rules. In the next section I argue that full information is not required, but that in general the expert needs to know the property he is being asked about to provide informative reports.

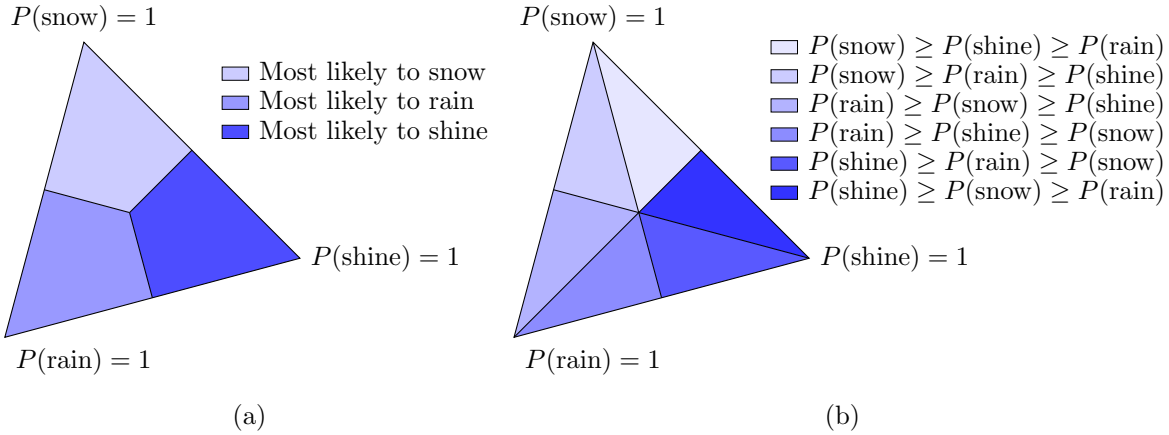

Figure 1: Graphical representations for two finite properties.

3.1 Elicitability and Proper Scoring Rules

A first objective is to understand when a property can be elicited with strict incentives, and when it cannot.

Indeed, strictly proper scoring rules may exist, but do not always exist. For example, consider the problem of predicting which one of a finite number of events $E_1, \ldots, E_m \subset \Omega$ is most likely. Such a property can be elicited via the scoring rule defined by $S(E_i, \omega) = \mathbb{1}\{\omega \in E_i\}$ which is immediately seen as being strictly proper. Now consider a simple example in a three-state world, with a random variable X taking values 1, 2 or 3. Let us look at the property that indicates whether X has "high" or "low" variance, where the levels of variance are determined with respect to some arbitrary threshold. This property is depicted in Figure 2. In this case a strictly proper scoring rule does not exist. If a scoring rule S strictly motivates the expert to make a truthful report when P has low variance, then $S(\text{"low variance"}, x) > S(\text{"high variance"}, x)$, as a distribution with an almost-sure state $X = x$ has a zero variance. But expected payoffs $S(\theta, P)$ are linear in the true distribution P. So for such a scoring rule, $S(\text{"low variance"}, P) > S(\text{"high variance"}, P)$ for every state distribution P: the expert induced to make truthful predictions when the variance is low is always best off reporting low variance levels even when the true variance is high. Thus, in this case, the amount of information included in the predictions are insufficient to enforce strict incentives.

A necessary condition for existence of strictly proper scoring rules is that the level sets of the property be convex; that is, the distributions that share the same property value must form a convex shape. This insight was already present in the pioneering work of Osband (1985), in the context of distributions on $\mathbf{R}^k$. Here, consider two distributions over states, P and Q. The argument relies on the simple observation that the expected payoff to the expert

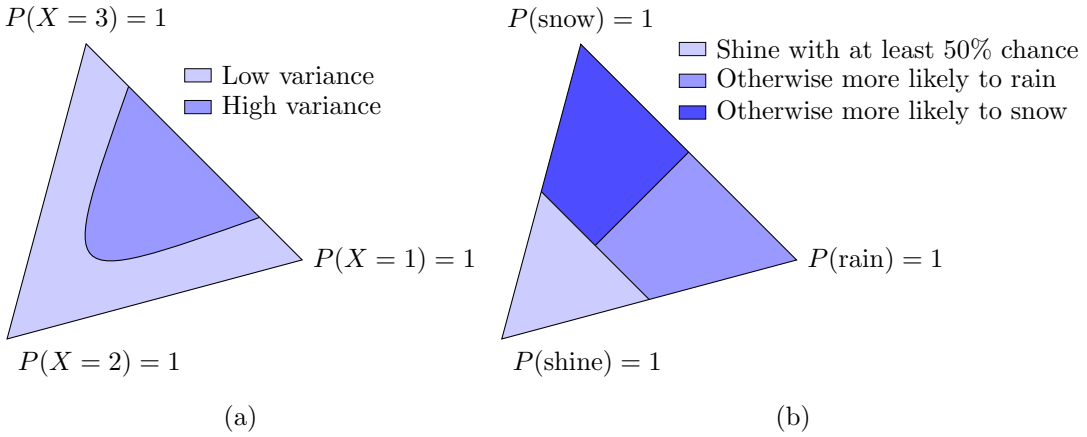

Figure 2: Two nonelicitable properties.

when predicting θ under any mixture of P and Q,

$$\mathrm{E}_{\omega\sim\lambda P+(1-\lambda)Q}[S(\theta,\omega)],$$

equals the mixture of the expected payoffs when predicting θ separately on P and Q,

$$\lambda\ \mathrm{E}_{\omega\sim P}[S(\theta,\omega)] + (1-\lambda)\ \mathrm{E}_{\omega\sim Q}[S(\theta,\omega)].$$

Suppose S is a strictly proper scoring rule and θ is a prediction that is correct for both P and Q. By reporting θ, the expert maximizes the expected payoff under both distributions. Per the above equality, the payoff remains optimal under any mixture of P and Q. Since S is strictly proper, it must be the case that θ is a correct prediction for all mixtures of P and Q. Hence all level sets must have a convex shape. Clearly, in the case of the variance depicted in Figure 2(a), the property does not partition the distributions into convex subsets.

However, convexity is generally not sufficient.[2] The exact characterization makes use of a well-known geometric structure called a *Voronoi diagram*. Voronoi diagrams specify, for a set

[2]This can be seen with the property pictured in Figure 2(b). Consider three possible states of the weather tomorrow: shine, rain, or snow. We want to know if it will shine with at least 50% chance (θ_A), or, if not, whether it is more likely to rain (θ_B) or to snow (θ_C). The property partitions the distributions in convex subsets. Yet there does not exist a strictly proper scoring rule. To see this, let us use the notation $P = (P(\text{shine}), \mathrm{P}(\text{rain}), \mathrm{P}(\text{snow}))$. Let $P_0 = (1,0,0)$, $P_1 = \left(\frac{1}{2}, \frac{1}{2}, 0\right)$, $P_2 = \left(\frac{1}{2}, 0, \frac{1}{2}\right)$, $P_3 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. Consider a proper scoring rule S. Both predictions θ_A and θ_B are true under P_1, so $S(\theta_A, P_1) = S(\theta_B, P_1)$. Similarly, $S(\theta_A, P_2) = S(\theta_C, P_2)$, $S(\theta_A, P_3) = S(\theta_B, P_3) = S(\theta_C, P_3)$, $S(\theta_B, P_0) = S(\theta_B, P_0)$. By linearity of the expected score, $2S(\theta_A, P_3) = S(\theta_A, P_1) + S(\theta_A, P_2)$, so $2S(\theta_C, P_3) = S(\theta_B, P_1) + S(\theta_C, P_2)$ implying $S(\theta_B, P_1) = S(\theta_C, P_1)$. Also, since the vectors P_0, P_1, P_2 are independent, $S(\theta_B, \cdot)$ is entirely specified by $S(\theta_B, P_0), S(\theta_B, P_1), S(\theta_B, P_3)$, and $S(\theta_C, \cdot)$ is entirely specified by $S(\theta_C, P_0), S(\theta_C, P_1), S(\theta_C, P_3)$. However, $S(\theta_B, P_0) = S(\theta_C, P_0)$, $S(\theta_B, P_3) = S(\theta_B, P_3)$, and $S(\theta_B, P_1) = S(\theta_C, P_1)$. Hence $S(\theta_A, \cdot) = S(\theta_B, \cdot)$ and S cannot be strictly proper.

of points called *sites*, the regions of the space that comprise the points closest to each site. Specifically, consider a metric space $\mathcal{E}$ with distance d, together with vectors $x_1, \ldots, x_n \in \mathcal{E}$ that are the sites. The *Voronoi cell* for site x_i includes all the vectors whose distance to x_i is less than or equal to the distance to any other site x_j. The collection of all the Voronoi cells is called the *Voronoi diagram* for the sites $x_1, \ldots, x_n$. Observe that the set of distributions, when viewed as a simplex in $\mathbf{R}^\Omega$ inherits its Euclidean metric. In this context it makes sense to talk about *Voronoi diagrams of distributions*, as well as *Voronoi diagrams of random variables*, since random variables are the elements of $\mathbf{R}^\Omega$. Voronoi diagrams have applications in several fields; see Aurenhammer (1991) and De Berg et al. (2008) for a literature review.

To understand the role that Voronoi diagrams play in the characterization, it is helpful to start off with a simple sufficient condition: if the level sets of a property form a Voronoi diagram of distributions, then the property is elicitable. The argument is as follows. Let (Θ, F) be a property. Let each level set $F(\theta)$ be the Voronoi cell of some distribution $Q_\theta \in \Delta(\Omega)$. Suppose that the expert is allowed to announce a full distribution Q, and is rewarded according to the Brier score $S(Q, \omega) = 2Q(\omega) - \|Q\|^2$. Aside from being strictly proper, the Brier score has the property that the closer the announced distribution is to that of Nature (in the Euclidean distance), the larger the expected payoffs (Friedman, 1983). In consequence if we were to force the expert to choose his report among the set of Voronoi sites $\{Q_\theta, \theta \in \Theta\}$, his best response would be to produce the Q_θ that is the closest to Nature's distribution. By forcing the expert to report one of these distributions, the expert reports the Voronoi cell that contains the distribution of Nature, thereby revealing a true value for the property. Because there is a one-to-one mapping between property values θ and sites Q_θ, the reward scheme corresponds to asking a value θ for the property and paying the expert according to the strictly proper scoring rule $S(\theta, \omega) = 2Q_\theta(\omega) - \|Q_\theta\|^2$.

That the property partition $\Delta(\Omega)$ into a Voronoi diagram of distributions is not necessary, because the logic of the above argument applies to other probability scoring rules and other distances. But it leads the way to the exact characterization, which turns out to be a generalization of this result. Instead of focusing on a Voronoi diagram in the space of distributions, we look at a Voronoi diagram in the entire space $\mathbf{R}^\Omega$. Specifically, the properties that are elicitable are precisely those whose level sets are included in a Voronoi diagram of random variables.

Theorem 1 *A property (Θ, F) is elicitable if, and only if, there exists a Voronoi diagram $\{\mathcal{C}_\theta\}_{\theta \in \Theta}$ in the space of random variables such that for every $\theta \in \Theta$, $F(\theta) = \mathcal{C}_\theta \cap \Delta(\Omega)$.*

Intersections of Voronoi diagrams with linear subsets are otherwise known as *power diagrams* in these subsets (Imai et al., 1985, Aurenhammer, 1987). Power diagrams are often

interpreted as extensions of Voronoi diagrams in which a weight factor on the sites shifts the distances between vectors and sites. With that identification in mind, Theorem 1 can be reformulated as follows: *the property is elicitable if and only if the level sets of the property form a power diagram of distributions.*

Proof of Theorem 1. Let (Θ, F) be a property and let $\{X_\theta\}_{\theta\in\Theta}$ be a family of random variables indexed by property values. Consider the Voronoi diagram of this family in the space $\mathbf{R}^\Omega$. Denote by $\mathcal{C}_\theta$ the Voronoi cell for X_θ, that is, the set of all the functions $X : \Theta \to \mathbf{R}$ that are at least as close to X_θ as to any other site $X_{\theta'}$, with respect to the Euclidean distance. Suppose $F(\theta)$ is the part of $\mathcal{C}_\theta$ that is located on the simplex.

Consider the scoring rule $S(\theta, \omega) = 2X_\theta(\omega) - \|X_\theta\|^2$. The expected payoff for prediction θ, under distribution P, is

$$\mathrm{E}_{\omega\sim P}[S(\theta, \omega)] = 2\langle X_\theta, P\rangle - \|X_\theta\|^2 = \|P\|^2 - \|P - X_\theta\|^2.$$

This means that the expected payoff of prediction θ under P is maximized across all possible predictions if and only if

$$\|P - X_\theta\| \leq \|P - X_{\widehat{\theta}}\| \qquad \forall\widehat{\theta} \in \Theta,$$

which is to say that P belongs to the Voronoi cell $\mathcal{C}_\theta$ of X_θ. Since $F(\theta) = \mathcal{C}_\theta \cap \Delta(\Omega)$, the expected payoff of prediction θ under P is maximized if and only if $P \in F(\theta)$, thereby establishing the strict properness of S.

To get the converse, assume there exists a strictly proper scoring rule S for a property (Θ, F). We need to construct random variables $\{X_\theta\}_{\theta\in\Theta}$ such that the associated Voronoi diagram in $\mathbf{R}^\Omega$ partitions the simplex $\Delta(\Omega)$ into the level sets of the properties. To do so, we will use $X_\theta(\omega) = S(\theta, \omega) + k_\theta$, where k_θ is a constant to be specified later.

Saying that distribution P is in the Voronoi cell of X_θ is saying that

$$\|P - X_\theta\|^2 \leq \|P - X_{\widehat{\theta}}\|^2 \qquad \forall\widehat{\theta} \in \Theta,$$

or equivalently, after expanding the terms,

$$-\|S(\theta, \cdot) + k_\theta\|^2 + 2k_\theta + 2\langle S(\theta, \cdot), P\rangle \geq -\|S(\widehat{\theta}, \cdot) + k_{\widehat{\theta}}\|^2 + 2k_{\widehat{\theta}} + 2\langle S(\widehat{\theta}, \cdot), P\rangle \quad \forall\widehat{\theta} \in \Theta.$$

If the choice in k_θ is such that $\|S(\theta, \cdot) + k_\theta\|^2 - 2k_\theta$ equals a constant c independent of θ, we can cancel these terms and the last inequality becomes

$$\mathrm{E}_{\omega\sim P}[S(\theta, \omega)] \geq \mathrm{E}_{\omega\sim P}[S(\widehat{\theta}, \omega)] \qquad \forall\widehat{\theta} \in \Theta.$$

Note that, for every θ, $\|S(\theta,\cdot)+k_\theta\|^2-2k_\theta$ is a parabola as a function of k_θ. As long as c is chosen to be greater that $\|S(\theta,\cdot)\|^2$ uniformly across property values—so that it intersects all the parabolas—it is always possible to select constants k_θ that satisfy this requirement. For such a choice of k_θ and X_θ, we have that for every distribution of Nature P, announcing prediction θ maximizes the expected payoff if and only if P is located in the Voronoi cell of X_θ. As S is strictly proper, a prediction θ maximizes the expected payoff if and only if it is true, that is, if $P \in F(\theta)$. Combining the two statements, we find that every level set $F(\theta)$ is the part of the Voronoi cell of X_θ located on the the simplex $\Delta(\Omega)$. ■

Theorem 1 essentially asserts that, as scoring rules vary, their associated value functions project onto power diagrams—or equivalently onto linear cross sections of Voronoi diagrams. Indeed, given a scoring rule S, the expert gets as expected payoff $\max_\theta S(\theta,P)$. The expected payoff, as a function of the true distribution of Nature P, is the value function. Saying that S is strictly proper is equivalent to saying that the projection of associated value function on the domain of distributions partitions $\Delta(\Omega)$ *exactly* as the level sets of the property $F(\theta)$ do. The properties we can elicit via strictly proper scoring rules therefore correspond to the projections of all the value functions. In the case of a finite property, the value function describes the upper envelope of a finite number of nonvertical hyperplanes. Moreover, by an appropriate choice of S, any such envelope can be obtained. Therefore the level sets of properties we can elicit via strictly proper scoring rules correspond exactly to the projections of hyperplane envelopes, which turn out to be the power diagrams.

The geometric characterization of the Voronoi test is appealing. As long as the dimension of the simplex of distributions is small, a quick visual check gives a good sense of whether the property satisfies the condition of Theorem 1. Figures 3, 4, 5, and 6 depict on the left side the simplex of distributions partitioned into level sets for, respectively, the rounded mean, the median, the most likely state and the ranking of states according to their probabilities, all of which are classical exemplars of finite properties. The right side of each figure maps a Voronoi diagram (along with the sites, all located on the simplex) that matches exactly the partition of level sets. Hence, all these properties are elicitable. Naturally the number of states must be kept artificially low to enable a 2-dimensional rendering of the simplex. However the Voronoi construction typically extends directly to higher dimensional simplexes. And in most cases, the 2-dimensional visual test is sufficient to get convinced of its existence. The examples in the remainder of this section provide the scoring rules that enable us to elicit those properties.[3]

[3]The ranking of states is not included in the list of examples and I discuss it here briefly. More generally, let $E_1,\dots,E_n$ be arbitrary events of any finite state space Ω, and consider the property that gives a ranking of these events by their likelihood. A simple generalization of Example 1 below yields that the scoring rules of the form $S(\sigma,\omega)=\kappa(\omega)+\sum_{i=1}^n \lambda_i \mathbb{1}\{\omega\in E_{\sigma(i)}\}$ are strictly proper, where σ is an ordering of events and

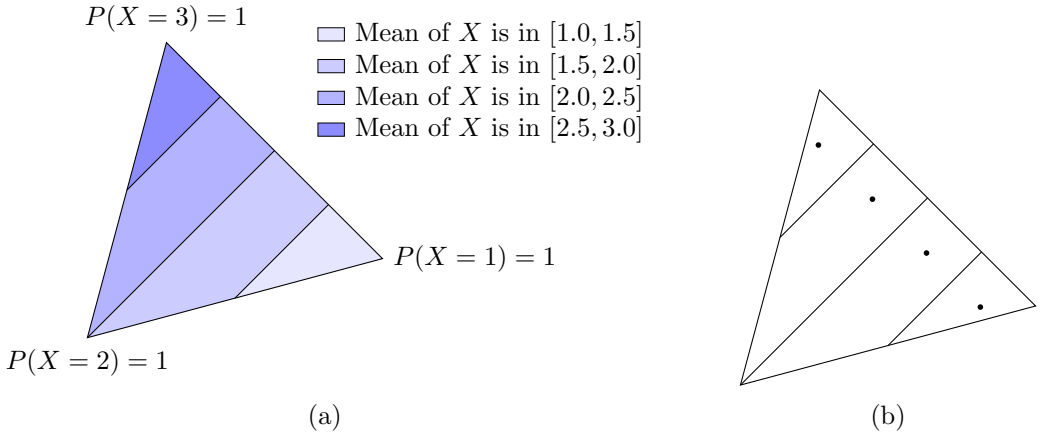

Figure 3: The mean.

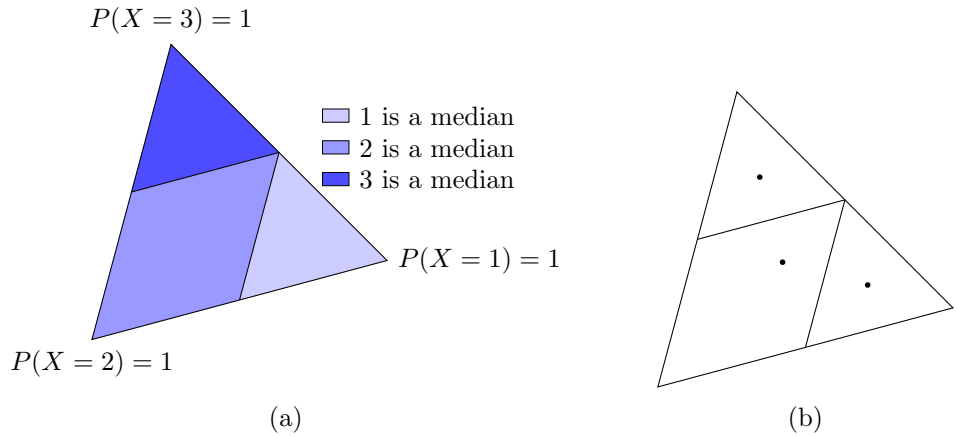

Figure 4: The median.

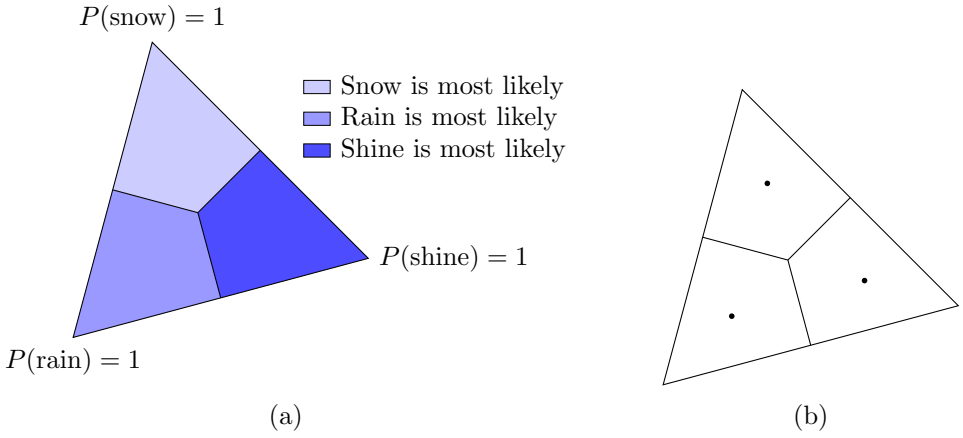

Figure 5: The most likely state of Nature.

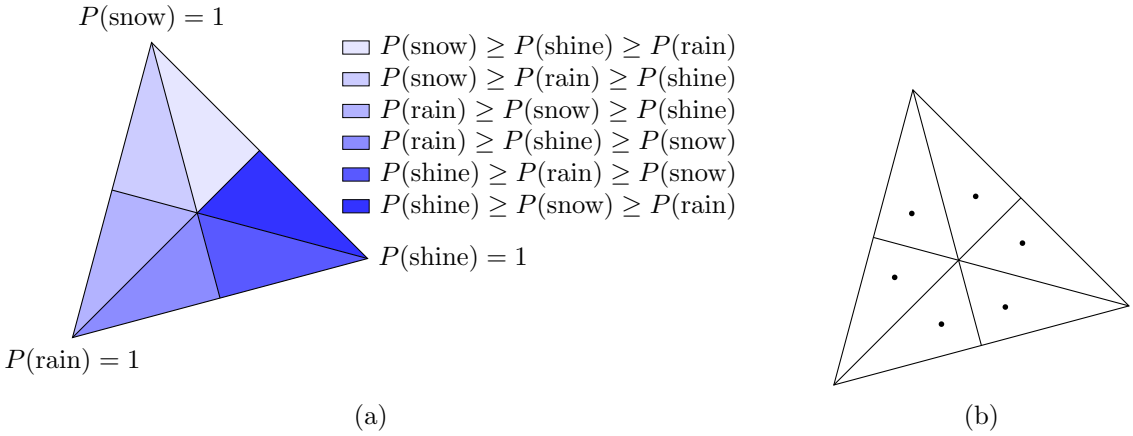

Figure 6: The ranking of states from most to least likely.

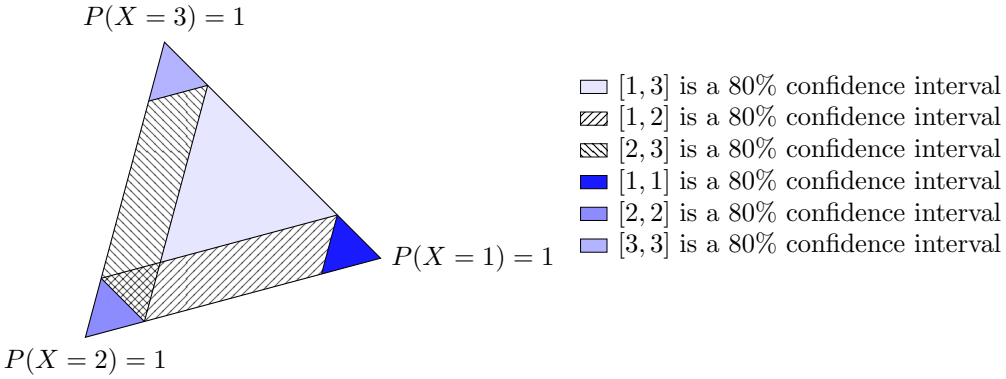

Figure 7: Confidence intervals.

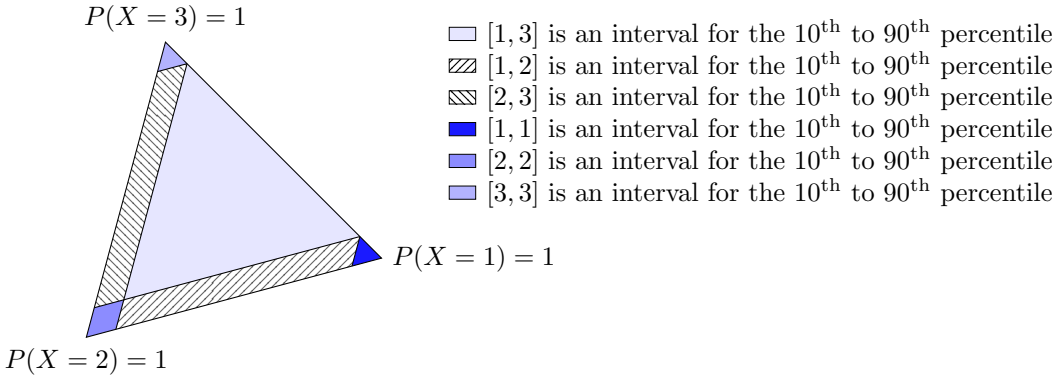

Figure 8: Intervals for the 10^{th} and 90^{th} percentiles.

It is not difficult to exhibit properties that fail the Voronoi test. We already saw the variance fails the convexity test in Figure 2(a), a condition weaker than the Voronoi test. In Figure 7, we are interested in a 80% confidence interval for a random variable.[4] The property fails the Voronoi test because of the large overlap for two of its level sets, corresponding to intervals $[1, 2]$ and $[2, 3]$. In this region, the densities cannot be equidistant to two distinct random variables. Whatever the scoring rule being used, there will be cases where a expert who reports one of the two intervals will not maximize his expected payoff, even when both intervals are correct. In general, to be able to elicit the predictions of the finite properties considered here, there must exist situations for which two or more predictions are simultaneously correct. But those situations should almost never happen, in the sense that the level sets should be a proper partition of the space of distributions except for a measure zero set of points which belong to two or more level sets. To properly elicit confidence intervals, we must reduce the overlap. For example we can require that predictions take the form of symmetric intervals as in Figure 8, that are the ranges between the 10^{th} and the 90^{th} percentiles. It is easily seen that the Voronoi test is then satisfied, and so the property is elicitable.

In general, to learn a property that is not elicitable, one must ask the expert a more precise question whose answer conveys more than just the information of the property. For example, the variance of a random variable is not elicitable at any precision level. However, it is possible to elicit the variable's mean and its second moment to an arbitrary precision, from which the variance is derived to an arbitrary precision. The expected shortfall is a commonly used risk measure that is not elicitable but can be obtained from conditional moments and quantiles, both of which are elicitable to an arbitrary precision.

Although Voronoi diagrams and convex partitions look alike, a Voronoi test can be much stronger than a convexity test. This is especially true in high dimensions. Nonetheless, most properties that exhibit a high level of symmetry are naturally shaped as Voronoi diagrams. Nonsymmetric cases can arise as well. For example, the property that gives the most likely of a list of (possibly overlapping) events. In such cases, the Voronoi sites are typically off the density simplex, precisely to shape the asymmetric structures. These cases are somewhat harder to visualize.

Now let us focus on a property that passes the Voronoi test of Theorem 1. How can we construct strictly proper scoring rules? The next result asserts that the proper and strictly proper scoring rules are essentially the mixtures of a finite number of carefully chosen proper

$\sigma(i)$ denotes the i-the most likely event, $\mathbb{1}\{\omega \in E\}$ has value 1 if E occurs and 0 otherwise, κ is arbitrary and $\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$.

[4]For a discrete random variable X, $[a, b]$ is a 80% confidence interval if the probability that $X \in [a, b]$ is at least 80% and if there is no interval $[c, d] \subsetneq [a, b]$ for which $X \in [c, d]$ with at least 80% probability.

scoring rules. These proper scoring rules form a base. Fixed once and for all, the base is entirely determined by the property being elicited.

Theorem 2 *Let (Θ, F) be a property that satisfies the Voronoi test in Theorem 1. There exist $\ell \geq 1$ proper scoring rules $S_1, \ldots, S_\ell$, called a base, such that a scoring rule S is proper (resp. strictly proper) if, and only if,*

$$S(\theta, \omega) = \kappa(\omega) + \sum_{i=1}^{\ell} \lambda_i S_i(\theta, \omega), \qquad \forall \theta \in \Theta, \omega \in \Omega,$$

for some function $\kappa : \Omega \mapsto \mathbf{R}$, and nonnegative (resp. strictly positive) reals λ_i, $i = 1, \ldots, \ell$.

Theorem 2 has a fairly strong interpretation. Given any elicitable property, there exists a fixed, finite number of baseline payoff functions, here represented by S_i, such that all the possible ways to elicit the property are payoff-equivalent to randomizing over those fixed payoff functions S_i. The probabilities of drawing each payoff function are fixed arbitrarily. One can also scale the payoffs by a constant factor and add an additive state-dependent payoff $\kappa(\omega)$, it does not affect incentives. However, the probabilities of draws, the constant scale and the additive random payoffs are the only degrees of freedom. This linear representation is particularly convenient when one wants to satisfy an optimality criterion, as illustrated in Appendix A.

Sketch of proof for Theorem 2. The full proof of Theorem 2 is in Appendix C. It is based on the following idea. Let S be a proper scoring rule. Properness is captured by the following constraints:

$$S(\theta, P) \geq S(\widehat{\theta}, P) \qquad \forall \theta, \widehat{\theta} \in \Theta, \forall P \in F(\theta).$$

There are uncountably many inequalities. However, whenever the property satisfies the criterion of Theorem 1 the sets $F(\theta)$ are polyhedra. Observing that the inequalities are linear in P, they need only be satisfied at the extreme vertices of these polyhedra. Thus the properness condition boils down to a finite system of homogeneous inequalities. By standard arguments (see, for example, Eremin (2002)), the solutions form a polytope that consists of a cone in the space of scoring rules, which is being copied and translated infinitely many times along some linear subspace. The directrices of the cone generate the "base" scoring rules. The kernel of the system, which gives rise to the translations, produces the complementary state-contingent payoffs. Adding strict properness substitutes some weak inequalities for strict ones in the above system, which complicates matters. Nonetheless the outcome remains intuitive: the strict inequalities only slightly perturb the solution space by excluding the

boundary of the translated cone. In effect, this exclusion is responsible for the strictly positive weights to all the scoring rules of the base. ■

Below are several examples that illustrate the representation of Theorem 2. I show how to obtain the base scoring rules for Examples 2–5 in Section 3.2. The proof that the scoring rules of Example 1 are proper and strictly proper is immediate. The converse, that there exists no other proper scoring rule, is tedious and omitted.[5]

Example 1 *Consider the property that gives the most likely of n events $E_1, \ldots, E_n$ of state space Ω. The events are arbitrary, and not necessarily pairwise incompatible. This property is elicitable, and it turns out to have only one base scoring rule. All the proper (resp. strictly proper) scoring rules S are written*

$$S(E_j, \omega) = \begin{cases} \kappa(\omega) + \lambda & \text{if } \omega \in E_j, \text{ i.e., } E_j \text{ is true,} \\ \kappa(\omega) & \text{if } \omega \notin E_j, \text{ i.e., } E_j \text{ is false,} \end{cases}$$

for arbitrary functions κ and nonnegative (resp. strictly positive) scalar λ. Note that these scoring rules also elicit the mode of a random variable as a special case.

Example 2 *Suppose the state is the realization x of some random variable X that can take n possible values $x_1, \ldots, x_n$. The median of X is an elicitable property, and the proper (resp. strictly proper) scoring rules take the form*

$$S(m, x) = \kappa(x) + \sum_{i=1}^{n-1} \lambda_i \cdot \left\{ \begin{array}{ll} -1 & \text{if } m > x_i, x \leq x_i \\ 0 & \text{if } m \leq x_i \\ +1 & \text{if } m > x_i, x > x_i \end{array} \right\},$$

where the scalars $\lambda_1, \ldots, \lambda_{n-1}$ are nonnegative (resp. strictly positive).

In this case, too, the family of all proper and strictly proper scoring rules takes a very simple form. After algebraic manipulation, it can be seen that all the scoring rules S that are proper (resp. strictly proper) for the median are written even more simply as

$$S(m, x) = \kappa(x) - |g(m) - g(x)|,$$

for arbitrary functions κ and g, where g is nondecreasing (resp. strictly increasing).

Note that, although the setting of this section is discrete, the scoring rules just displayed remain proper (resp. strictly proper) when X takes a continuum of values. Although, in this

[5] A proof is available upon request.

case, the scoring rules (and below, for the variance at risk) differ from the schemes defined by Thomson (1979) who elicits quantiles, it can be shown that they are equivalent.[6]

Example 3 *Let us divide the range* $[0,1]$ *into* n *intervals of equal size,* $\left[\frac{j-1}{n}, \frac{j}{n}\right]$, $j = 1, \ldots, n$. *Suppose the state is binary,* $\omega \in \{0,1\}$, *and consider the property corresponding to the interval of probability for state 1. This property is elicitable, and the proper (resp. strictly proper) scoring rules are*

$$S\left(\left[\tfrac{j-1}{n}, \tfrac{j}{n}\right], \omega\right) = \kappa(\omega) + \sum_{i=1}^{n-1} \lambda_i \cdot \left\{ \begin{array}{ll} n-i & \text{if } j > i, \omega = 1 \\ 0 & \text{if } j \leq i \\ -i & \text{if } j < i, \omega = 0 \end{array} \right\},$$

where the scalars $\lambda_1, \ldots, \lambda_{n-1}$ *are nonnegative (resp. strictly positive). After simplification, we find that the proper (resp. strictly proper) scoring rules for probability intervals take the form*

$$S\left(\left[\tfrac{j-1}{n}, \tfrac{j}{n}\right], \omega\right) = \kappa(\omega) + (g(j) - g(1))\omega + \frac{1}{n}\sum_{i=1}^{j-1}(g(j) - g(i)),$$

for arbitrary functions κ *and* g, *where* g *is nondecreasing (resp. strictly increasing).*

An interesting special case is $\kappa(\omega) = -\omega/2$ *and* $g(k) = k/n$. *We then have*

$$S\left(\left[\tfrac{j-1}{n}, \tfrac{j}{n}\right], \omega\right) = \kappa(\omega) + \frac{1}{n}(j-1)\omega + \frac{j(j-1)}{n^2}.$$

As n *grows large, probability intervals become increasingly finer and eventually converge to singleton probabilities. Informally, at the limit, reporting an interval* $\left[\frac{j-1}{n}, \frac{j}{n}\right]$ *becomes the same as reporting a probability* p, *where* $j/n \to p$. *Then,* $1/n \to 0$ *and* $j(j-1)/n^2 \to p^2/2$, *and thus, in the limit, we obtain scoring rule* $S(p,\omega) = -\omega/2 + p\omega - p^2/2 = -\frac{1}{2}(p-\omega)^2$, *and we rediscover the Brier score or quadratic loss.*

Example 4 *Suppose random variable* X, *with possible realizations* $x_1, \ldots, x_n$, *is associated with the value change of a portfolio investment over some fixed period of time. The value at risk (VaR) is a common risk measure of the loss of investment. Formally, the value at risk of the investment at confidence level* α *defined by a value* v *such that the probability of a loss greater than* v *is at least* $1-\alpha$, *and at the same time, the probability of a loss less than* v *is at least* α*—that is, the value at risk is the same as the* α*-quantile of the loss. Suppose the state is the realization* x *of the (random) investment gain* X, *so that* $-x$ *is the loss over the period considered. The value at risk is an elicitable property, and the proper (resp. strictly*

[6] A proof is available upon request.

proper) scoring rules are

$$S(v,x) = \kappa(x) + \sum_{i=1}^{n-1} \lambda_i \cdot \left\{ \begin{array}{ll} -(1-\alpha) & \text{if } v > -x_i, x \geq x_i \\ 0 & \text{if } v \leq -x_i \\ +\alpha & \text{if } v > -x_i, x < x_i \end{array} \right\},$$

where the scalars $\lambda_1, \ldots, \lambda_{n-1}$ are nonnegative (resp. strictly positive).

After simplification, we find that all the scoring rules S that are proper (resp. strictly proper) for the value at risk at confidence level α are written

$$S(v,x) = \kappa(x) + (2\alpha - 1)(g(v) - g(x)) - |g(v) - g(x)|,$$

for arbitrary functions κ and g, where g is nonincreasing (resp. strictly decreasing).

Example 5 *As in Example 2, suppose the state is the realization of some random variable X. For $J \geq 2$, let $m_1 < \cdots < m_J$ and let us consider the property associated with the interval $[m_{j-1}, m_j]$ that includes the mean of X. Assume that the values of the m_j's are chosen so that for all distributions, at least one interval includes the mean and every interval includes the mean for some distribution. The mean interval property is elicitable, and the proper (resp. strictly proper) scoring rules take the form*

$$S\left([m_{j-1}, m_j], x\right) = \kappa(x) + \sum_{i=1}^{J-1} \lambda_i \cdot \left\{ \begin{array}{ll} x - m_i & \text{if } m_j > m_i, \\ 0 & \text{if } m_j \leq m_i \end{array} \right\},$$

where the scalars $\lambda_1, \ldots, \lambda_{J-1}$ are nonnegative (resp. strictly positive).

3.2 Strictly Order-Sensitive Scoring Rules

I now discuss order sensitivity and its interplay with properness. Consider a scoring rule that takes value in a set attached with a natural ordering of its elements. The result below is a test for the existence of strictly order-sensitive scoring rules. As expected, the test is stronger than the Voronoi test of Theorem 1. But it is also easier to carry out. A property passes the test if and only if it partitions the distributions into "slices," as in Figure 9(a), and as opposed to Figure 9(b).

Theorem 3 *Let $(\Theta = \{\theta_1, \ldots, \theta_n\}, F)$ be a property, with $\theta_1 \prec \cdots \prec \theta_n$. There exists a scoring rule that is strictly order sensitive with respect to the order relation $\prec$ if, and only if, for all $i = 1, \ldots, n-1$, $F(\theta_i) \cap F(\theta_{i+1})$ is a hyperplane of $\Delta(\Omega)$.*[7]

[7]Hyperplanes of distributions can be viewed as hyperplanes in the Euclidean space $\mathbf{R}^\Omega$ that intersect the

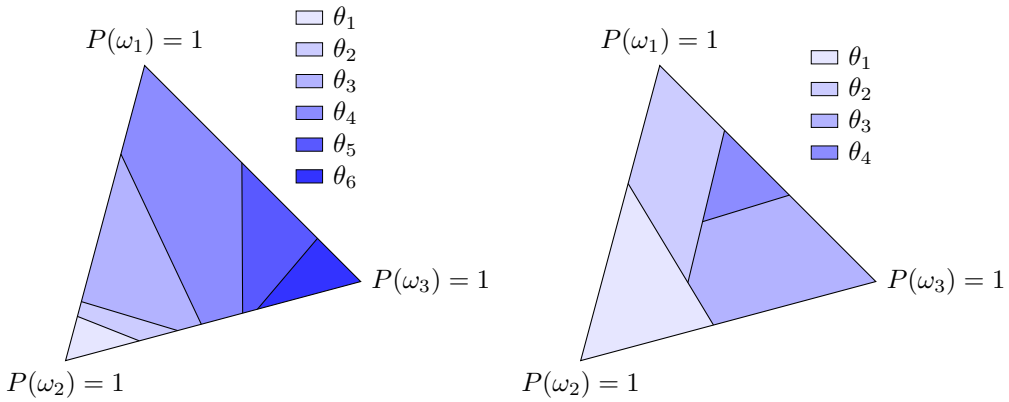

Figure 9: Strict order sensitivity can be enforced on the left property only.

Sketch of proof for Theorem 3. The proof of Theorem 3 is in Appendix C. The proof idea is best conveyed through an example. Consider a strictly order-sensitive scoring rule for a property whose value set Θ contains three elements, θ_1, θ_2 and θ_3. If both θ_1 and θ_3 are correct predictions under some distribution P, but θ_2 is not, then the expected payoff, under P, is maximized only when responding θ_1 or θ_3. Adding a small perturbation to P, we can pull out a distribution $\widetilde{P}$ for which the only true prediction is θ_1, while announcing θ_3 yields an expected payoff that is nearly maximized and larger than that derived from announcing θ_2. This contradicts strict order sensitivity. This means that, whenever we choose some $P \in F(\theta_1)$ and $Q \in F(\theta_3)$, the segment of distributions must go through $F(\theta_2)$. More generally, suppose the property takes more than three values. For any two distributions $P \in F(\theta_i)$ and $Q \in F(\theta_j)$, $i < j$, the segment of distributions starting from P and ending at Q must pass by, in order, through $F(\theta_i), F(\theta_{i+1}), \ldots, F(\theta_j)$, by which the hyperplane separation holds. The converse can be made clear through an explicit construction of the strictly order-sensitive scoring rules, which is the object of Theorem 4. ■

For example, we can apply Theorem 3 to the case of the median and the mode of a random variable X. For the median, Figure 4 suggests that the property passes the slice test of Theorem 3.[8] In contrast, consider the mode of X. This property gives the most likely value of X. Figure 5, for which the mode is a special case, clearly indicates that the property fails the test of Theorem 3.[9]

simplex of the state distributions.

[8]Indeed, choosing two consecutive values for X, x and y, we easily verify that $F(x) \cap F(y)$ is a hyperplane. If both are possible median values under a distribution P, then $P(X \leq x) \geq \frac{1}{2}$, $P(X \geq x) \geq \frac{1}{2}$, and $P(X \leq y) \geq \frac{1}{2}$, $P(X \geq y) \geq \frac{1}{2}$. Hence $P(X > x) = P(X \geq y) \geq \frac{1}{2}$, and, as $P(X \leq x) + P(X > x) = 1$, $P(X \leq x) = \frac{1}{2}$. The converse is immediate. This means that the set $F(x) \cap F(y)$ is the hyperplane defined by $\sum_{z \leq x} P(X = z) = \frac{1}{2}$. Hence the criterion of Theorem 3 is satisfied.

[9]To be convinced of this assertion, choose two consecutive values of X, x and y. The set $F(x) \cap F(y)$

Strictly order-sensitive scoring rules are also strictly proper, and the form of the strictly proper contracts follows the rule given in Theorem 2. One benefit of properties that admit strictly order-sensitive scoring rules is that the base scoring rules are easily derived; they are $0-1$ factors of the normals to the boundaries of consecutive level sets. Together Theorem 2 and Theorem 4 can be used to obtain the strictly proper scoring rules for any property that satisfies the condition of Theorem 3.

Theorem 4 *Let $(\Theta = \{\theta_1, \ldots, \theta_n\}, F)$ be a property with $\theta_1 \prec \cdots \prec \theta_n$. Assume there exists a strictly order-sensitive scoring rule S with respect to the order relation $\prec$. The scoring rules $S_1, \ldots, S_{n-1}$, defined by*

$$S_i(\theta_j, \omega) = \begin{cases} 0 & \text{if } j \leq i, \\ \mathbf{n}_i(\omega) & \text{if } j > i, \end{cases}$$

form a base, with $\mathbf{n}_i$ being a positively oriented normal (i.e., oriented towards $F(\theta_{i+1})$) to the hyperplane of random variables in $\mathbf{R}^\Omega$ generated by $F(\theta_i) \cap F(\theta_{i+1})$.

The proof of Theorem 4 is in Appendix C.

It remains to characterize the order-sensitive scoring rules. As it turns out, as long as a strictly order-sensitive scoring rule exists, all the proper (resp. strictly proper) scoring rules are also order sensitive (resp. strictly order sensitive), so that the characterization of Theorem 4 still applies. The result implies that when a property admits a strictly order-sensitive scoring rule, it does so for exactly two order relations, one being the reverse of the other. As demonstrated in the sketch proof of Theorem 3, the result breaks down without the existence requirement. It breaks down even when restricted to weak order sensitivity, which, obviously, exists for all properties.

Proposition 1 *Let $(\Theta = \{\theta_1, \ldots, \theta_n\}, F)$ be a property with $\theta_1 \prec \cdots \prec \theta_n$. Assume there exists a strictly order-sensitive scoring rule with respect to the order relation $\prec$. A scoring rule is proper (resp. strictly proper) if and only if it is order sensitive (resp. strictly order sensitive), with respect to $\prec$.*

The proof of Proposition 1 is in Appendix C.

Put together, Theorem 4 and Proposition 1 are particularly useful to obtain the base necessary to the design of proper scoring rules. I illustrate this use with a few cases below.

First, let us return to Example 2 about the median property of random variable X. Let x_i be the i-th smallest value taken by X. The hyperplane that separates two consecutive level

contains all distributions P such that $P(X = x) = P(X = y)$, equality that indeed defines a hyperplane. However it is only part of a hyperplane, because there are distributions that assign the same probability to both x and y, and yet whose most likely values are attained elsewhere. As $F(x) \cap F(y)$ does not cover an entire hyperplane of distributions, it fails the above criterion.

sets of the median, for respective values x_i and x_{i+1}, is, as established previously, specified by equation $\sum_{k\leq i} P(x_k) = \frac{1}{2}$. And so, the functions $\mathbf{n}_i$ defined by

$$\mathbf{n}_i(x) = \begin{cases} -1 & \text{if } x \leq x_i, \\ +1 & \text{if } x > x_i, \end{cases}$$

are positively oriented normals for each $i = 1, \ldots, n-1$. The normals generate the $n-1$ base scoring rules used in Example 2,

$$S_i(m, x) = \begin{cases} -1 & \text{if } m > x_i, x \leq x_i, \\ 0 & \text{if } m \leq x_i, \\ +1 & \text{if } m > x_i, x > x_i. \end{cases}$$

The base scoring rules for the variance at risk of Example 4 are obtained in a similar fashion. The hyperplane that separates two consecutive values of variance at risk, x_i and x_{i+1}, is specified by equation $\sum_{k\leq i} P(x_k) = \alpha$. A set of positively oriented normals is then

$$\mathbf{n}_i(x) = \begin{cases} -(1-\alpha) & \text{if } x \leq x_i, \\ +\alpha & \text{if } x > x_i, \end{cases}$$

for $i = 1, \ldots, n-1$, which are then used to generate the base, and eventually yield the proper and strictly proper scoring rules displayed in Example 4.

Next let us return to Example 3, that describes the property that gives a probability interval $\mathcal{I}_j = \left[\frac{j-1}{n}, \frac{j}{n}\right]$ for state 1. These intervals are naturally ordered by $\mathcal{I}_1 \prec \cdots \prec \mathcal{I}_n$. The hyperplane that separates two consecutive level sets is specified by the set of distributions P such that both $P(1) \in \mathcal{I}_i$ and $P(1) \in \mathcal{I}_{i+1}$, that is, $P(1) = i/n$. Hence the following

$$\mathbf{n}_i(\omega) = \begin{cases} 1 - i/n & \text{if } \omega = 1, \\ -i/n & \text{if } \omega = 0, \end{cases}$$

defines a positively oriented normal for every $i = 1, \ldots, n-1$, from which we can derive the base scoring rules of Example 3.

Finally, let us return to Example 5, which concerns the mean interval of some random variable X. Using the same notation as this example, the hyperplane that separates to consecutive intervals $[m_{i-1}, m_i]$ and $[m_i, m_{i+1}]$ for the mean is given by the equation $\sum_k x_k P(x_k) = m_i$, and the functions $\mathbf{n}_i$ defined by $\mathbf{n}_i(x) = x - m_i$ are positively oriented normals, which generate

the scoring rules proposed in Example 5.

4 Partially Informed Experts

In this section, the expert is no longer assumed to know perfectly the state distribution. Instead, he is partially informed and believes that several state distributions are possible. This belief, the set of distributions the expert deems possible, is a closed subset of $\Delta(\Omega)$. When a state distribution is included in the belief, that distribution is *compatible* with the belief. The expert is endowed with maxmin preferences as defined in Gilboa and Schmeidler (1989), he evaluates a risky prospect as the worst case expected payoff under all the possible state distributions compatible with his belief. Note that, when asked to supply a property value as report, the expert may decide randomly over several values—unlike the special case of full information, randomized reporting strategies cannot be eliminated without losing generality.

With partial information, the expert may be certain that some property values are correct (or incorrect) but, because the belief generally does not reduce to a singleton, he may also be unsure. Let us say that an expert *knows a property value* when, according to his belief, this property value matches the state distribution for sure. Similarly, let us say that an expert *knows a property* when he is able to identify at least one property value that he knows, even though he may not be able to identify all the property values that match the actual state distribution.

In a context of partial information, it is natural to ask is if the incentive schemes known to work with fully informed experts continue to work with partially informed experts when these experts have the information we request from them.

Proposition 2 *Suppose a property is elicitable and an expert is rewarded according to a proper scoring rule. If the expert knows the property value θ, then reporting θ is a best response.*

The argument is immediate. Let $\mathcal{B}$ denote the expert's belief. If the expert knows θ, then $\mathcal{B} \subseteq F(\theta)$. Let $\widetilde{\theta}$ be any property value other than θ. As S is proper, $S(\theta, P) \geq S(\widetilde{\theta}, P)$ for every $P \in F(\theta)$, and so also for every $P \in \mathcal{B}$. Hence, for all state distributions the expert believe possible, supplying θ makes him as least as well off as reporting any other property value, and so at least as well off as any reporting strategy, including randomized strategies.

Thus, Proposition 2 says that a report the expert knows to be correct continues to be a best response, so that strictly proper scoring rules cause no distortion when applied to a partially informed expert. Of course one may also ask if, as in the case of a fully informed

expert, the incentives are strict. When a property value is possible according to the expert's belief, meaning that at least some state distribution is compatible with the belief, by extension let us say of the property value that it is *compatible* with the belief. Note that a response that is compatible with an expert's belief need not be correct. One may want to ensure that, at least, an expert believes that his own responses are possibly correct. As long as the expert knows the property, this condition holds.

Proposition 3 *Suppose a property is elicitable and an expert is rewarded according to a strictly proper proper scoring rule. If the expert knows the property, then the expert can only best respond by reporting values compatible with his belief. In addition, if the diameter of the expert's belief falls below some positive threshold, then the expert only best responds by reporting correct values.*

Proof. Let the expert have belief $\mathcal{B}$ and know the property value θ. Let $\widetilde{\theta}$ be a property value not compatible with $\mathcal{B}$. For every $P \in \mathcal{B}$, we have $P \in F(\theta)$ but also $P \notin F(\widetilde{\theta})$, and so $S(\theta, P) > S(\widetilde{\theta}, P)$, because S is strictly proper. By Proposition 2 reporting θ is a best response, so that any reporting strategy in which the expert chooses to report an incompatible property value with positive probability cannot be a best response.

Finally, if a property value does not match the state distribution, then the Euclidean distance between the state distribution and the level set of that value, which by Theorem 1 is a Voronoi cell, is nonzero. Hence, if the expert's belief has a small enough diameter, this belief, which includes the actual state distribution, does not intersect with that level set so that the incorrect property value is not compatible with the belief. Since there are finitely many property values, it is always possible to select a positive threshold so that any belief with a diameter less than the threshold is guaranteed not to intersect any level set of a property value that does not match the state distribution; in that case, no incorrect property value is compatible with the expert's belief. ■

Thus, as long as the expert knows the property of interest, a partially informed expert is induced to answer coherently with his own belief in all mechanisms for which telling the truth is a strict best response for a fully informed expert. If the belief is precise enough, then the partially informed expert will always best respond by sending a report guaranteed to be correct, but it need not be otherwise. Except for the case of binary properties, it is generally not possible to design a scoring rule that ensures that all best responses always yield correct reports even when the expert knows the property.[10] In order to achieve this effect, one must

[10] A minimal example is as follows. Consider an environment with binary state $\omega \in \{\text{good}, \text{bad}\}$. The property is the probability of the good state belonging to interval $[0, 1/3]$, $[1/3, 2/3]$, or $[2/3, 1]$, respectively. No matter the scoring rule employed, there always exists a belief and a property value fully compatible with

make use of the ambiguity aversion of the expert and have the elicitor pick one of several possible scoring rule, after the expert makes the report, and without committing to any randomized selection.

If an expert does not know a property, then in general little can be said: the expert's only best response may be to send a report inconsistent with his own belief.

Proposition 4 *Suppose a property is elicitable and an expert is rewarded according to a strictly proper scoring rule. If the expert does not know the property, then it is possible that in all best responses, the expert never supplies a report compatible with his belief.*

Proof. Consider the case of a binary state $\omega \in \{\text{good}, \text{bad}\}$. The property is whether the good state has probability less than (or equal to) 1/3, more than (or equal to) 2/3, or between 1/3 and 2/3 (endpoints included). Let S be defined as follows:

$$\begin{aligned} S\left(\left[0, \tfrac{1}{3}\right], \text{good}\right) &= -2, & S\left(\left[0, \tfrac{1}{3}\right], \text{bad}\right) &= 1, \\ S\left(\left[\tfrac{1}{3}, \tfrac{2}{3}\right], \text{good}\right) &= 0, & S\left(\left[\tfrac{1}{3}, \tfrac{2}{3}\right], \text{bad}\right) &= 0, \\ S\left(\left[\tfrac{2}{3}, 1\right], \text{good}\right) &= 1, & S\left(\left[\tfrac{2}{3}, 1\right], \text{bad}\right) &= -2. \end{aligned}$$

It is immediate to verify that S is strictly proper for the property being considered.

Suppose the expert is believes that the probability of the good state is either 1/4 or 3/4. The expert does not know this property. There are two property values compatible with the belief, $[0, 1/3]$ and $[2/3, 1]$. However, if he reports any of these two values with positive probability, the expert's worst case expected payoff is negative, whereas the payoff is always zero when reporting the only incompatible property value $[1/3, 2/3]$.[11] ■

Related to Proposition 4, if the expert knows the property of interest but is asked to report the full state distribution (so, is asked to answer a question he may not know the answer to) and is rewarded according to a strictly proper probability scoring rule, such as the quadratic scoring rule, then it is generally not possible to back out correct property values, even if the property is elicitable. An analogous result holds when the expert is asked to report the value of a property that is finer than the expert's belief.

Proposition 5 *Consider an elicitable property and suppose an expert who knows the property is asked to report an entire distribution, and is rewarded according to a strictly proper probability scoring rule. Then, there can be a unique best response and the reported distribution can be entirely uninformative about the expert's belief.*

the belief such that reporting another, not fully compatible property value is a best response. The proof is simple but tedious and is omitted.

[11] The expert's belief is not convex in this example. With more than two states, it is possible to construct a similar but tedious example in which the belief is convex.

Proof. The argument is general but is best illustrated by a simple example. As in Proposition 4, consider the case of a binary state $\omega \in \{\text{good}, \text{bad}\}$. The property is now whether the good state has probability less than (or equal to) $1/2$, or more than (or equal to) $1/2$. The expert believes that the good state has probability no greater than $1/2$, so he knows the property. Suppose the expert is asked to send an estimate p of the probability that the state is good, and is rewarded, for example, according to the quadratic scoring rule that delivers the payoff $-(1-p)^2$ if the state is good, and $-p^2$ if the state is bad. It is easily seen that the unique best response is to report probability $1/2$, and by symmetry, the same unique best response holds in the reverse case in which the expert believes that the good state has probability no less than $1/2$. ■

These last two results motivate the elicitation of partial information, when the full distribution is not needed. If the expert's belief is too vague and we ask for fine information, the expert may respond in a way that does not reflect his own belief and is uninformative, so as to protect himself against the worst case. The problem is avoided when asking for information just coarse enough so that the expert can form unambiguous assessments.

One may want to ensure that the expert we solicit knows the property. Can the elicitor design a payment scheme that only attracts the experts who are knowledgeable? Formally, let us define a contract as a finite collection of state-contingent payoffs. An expert who is offered such a contract can either reject the offer, and get zero payoff, or accept the offer and choose one state-contingent payoff from the menu. The menu may be indexed by property values, but in general need not be. The goal is to design a contract that an informed expert, who knows the property, accepts, while an expert who does not know the property declines. The result below demonstrates the impossibility to design such screening contracts, in a strong sense, independently of the property of interest.

Proposition 6 *Consider a contract $\mathcal{C}$. If a fully informed expert is at least weakly better off accepting the contract, then an ignorant expert, i.e., an expert whose belief is the entire simplex of distributions, is also at least weakly better off accepting the contract.*

Proof. The proof is short and directly inspired by the manipulability of tests or contracts by minimax arguments, in particular Sandroni (2003) and Olszewski and Sandroni (2007). If an expert is fully informed and forms belief P regarding the state distribution, his expected payoff from state-contingent payoff π is $\langle \pi, P \rangle$. Suppose it is always in the best interest of an expert who is fully informed to accept the contract, no matter his belief. Then,

$$\min_{P \in \Delta(\Omega)} \max_{Q \in \Delta(\mathcal{C})} \sum_{\pi \in \mathcal{C}} Q(\pi) \langle \pi, P \rangle \geq 0.$$

Since $\mathcal{C}$ and Ω are finite, and expected payoffs are linear in P and Q, we can apply von Neumann's minimax theorem to get

$$\max_{Q\in\Delta(\mathcal{C})}\min_{P\in\Delta(\Omega)}\sum_{\pi\in\mathcal{C}}Q(\pi)\langle\pi,P\rangle=\min_{P\in\Delta(\Omega)}\max_{Q\in\Delta(\mathcal{C})}\sum_{\pi\in\mathcal{C}}Q(\pi)\langle\pi,P\rangle.$$

Hence, there exists Q such that, if an ignorant expert randomizes over the state-contingent payoffs of the contract $\mathcal{C}$ according to Q, the worst-case expected payoff of the expert is nonnegative, so that the ignorant expert is always at least weakly better off accepting the contract. ■

It is worth noting that the negative result hinges on the fact to that we consider an expert in isolation. For example, suppose instead that there is a population of experts and consider any elicitable property of interest to the elicitor. In the simplest case, a positive fraction of the population knows the property of interest, while the rest of the population is ignorant. The elicitor offers to two or more experts to work for her in the following manner. If an expert i accepts the offer, he is randomly matched against another accepting expert j and gets $\epsilon+S(\theta_i,\omega)-S(\theta_j,\omega)$ where θ_i is the report made by expert i, and θ_j the report made by expert j (the offer is canceled in the event that the elicitor is unable to get more than one expert to work for her). In this case, if ϵ is positive but small enough, no ignorant expert is willing to accept the offer, while the informed experts are strictly better off accepting. In the general case of a diverse, heterogeneous population, one can ensure that only the best informed experts accept the offer by selecting ϵ arbitrarily small (if $\epsilon=0$, only the most informed experts accept). This fact results from a simple unraveling argument.

5 An Application to Forecast Testing

While the main model of this paper is geared towards the use of scoring rules as incentive devices, one major area where scoring rules are used in practice is the evaluation of forecasters and the calibration of learning models, where the average score of forecasts over time provide an assessment of relative performance. In this section, I apply the framework of this paper to the problem of comparing the quality of two (or more) experts who provide forecasts over time. It is convenient to separate the two cases of Bayesian and non-Bayesian experts.

Bayesian Experts

There are two experts $i=1,2$. As in the model of Section 2, there continues to be a finite set of states Ω, but there are now infinitely many time periods indexed $t=1,2,\ldots$. At every

date t, a state realizes. Alongside the sequence of states, there is a sequence of pairs of signals, one signal for each expert. So, at every date t, three variables are drawn at random: the state ω_t, the signal y_t^1 of expert 1, and the signal y_t^2 of expert 2. Each signal takes value in a finite signal space $\mathcal{Y}$. The sequence of states and signals is generated according to probability measure μ.

There is an elicitable property of interest, (Θ, F), about the state distribution. The two experts provide, from one period to the next, forecasts about the property for the next period's state distribution. It is important that there be at least two experts, however the results of this part and the next generalize to more than two experts.

In this part I suppose that the probability measure μ is common knowledge and that experts form their beliefs based on the signals that they observe. At date t, they know the history of signal realizations and the history of state and forecast realizations, up to date t. (Alternatively, we can assume that signal realizations remain private but that previous signals are irrelevant to infer the state at a given date, conditionally on the state history.) At date t, the private information of expert i is the signal of the next period, y_{t+1}^i. Therefore, at date t, given history h_t, expert i's assessment of the probability that the next state be ω_{t+1} is $\mu(\omega_{t+1} \mid h_t, y_{t+1}^i)$.

Let us say that expert i is *more informed* that expert j, if, at every date and for every history, expert i's signal distribution given the state is more informative than that of expert j according to the Blackwell ordering of information structures (Blackwell, 1951, 1953).

Let S be any strictly proper scoring rule for (Θ, F). The difference of average scores over the first T periods between expert 1 and expert 2 is defined as

$$\Delta S(T) = \frac{1}{T} \sum_{t=1}^{T} (S(\theta_t^1, \omega_t) - S(\theta_t^2, \omega_t)),$$

where θ_t^i is the forecast of expert i for date t. Because scoring rules are the negative of loss functions, computing the average score is a common way to measure the performance of a forecaster over time. Informally, $\Delta S(T)$ is an indication of "how much" expert 1 outperforms expert 2 over the first T time periods.

Proposition 7 *If expert 1 is more informed than expert 2, then with probability 1,*

$$\liminf_{T \to \infty} \Delta S(T) \geq 0.$$

Proof. For both experts $i = 1, 2$,

$$\begin{aligned} \mathrm{E}_{\theta_t^i,\omega_t}\left[S(\theta_t^i,\omega_t) \mid h_{t-1}\right] &= \mathrm{E}_{y_t^i}\left[\mathrm{E}_{\theta_t^i,\omega_t}\left[S(\theta_t^i,\omega_t) \mid h_{t-1}, y_t^i\right] \mid h_{t-1}\right] \\ &= \mathrm{E}_{y_t^i}\left[\max_\theta \mathrm{E}_{\omega_t}\left[S(\theta,\omega_t) \mid h_{t-1}, y_t^i\right] \,\middle|\, h_{t-1}\right], \end{aligned}$$

because expert i's forecast for date t maximizes his expected score given the information available to him, history h_{t-1} and signal y_t^i. At date t and for history h_{t-1}, the information structure of expert i is captured by the conditional probabilities $\mu(\omega_t \mid y_t^i, h_{t-1})$ for $\omega_t \in \Omega$ and $y_t^i \in \mathcal{Y}$. By assumption, for every date and at every history, the information structure of expert 1 is more informative than the information structure of expert 2 according to the Blackwell ordering. Interpreting the problem of announcing the forecast at date t as a decision problem whose utility is given by the score, a direct implication of the Blackwell ordering is that, at every date t and for every history h_{t-1},

$$\mathrm{E}_{y_t^1}\left[\max_\theta \mathrm{E}_{\omega_t}\left[S(\theta,\omega_t) \mid y_t^1, h_{t-1}\right] \,\middle|\, h_{t-1}\right] \geq \mathrm{E}_{y_t^2}\left[\max_\theta \mathrm{E}_{\omega_t}\left[S(\theta,\omega_t) \mid y_t^2, h_{t-1}\right] \,\middle|\, h_{t-1}\right].$$

Hence,

$$\mathrm{E}_{\theta_t^1,\omega_t}\left[S(\theta_t^1,\omega_t) \mid h_{t-1}\right] \geq \mathrm{E}_{\theta_t^2,\omega_t}\left[S(\theta_t^2,\omega_t) \mid h_{t-1}\right].$$

We apply Dawid's calibration theorem (Dawid, 1982) (strictly speaking, we apply an extension of this theorem, see Shiryaev (1996), chapter 7, section 3, corollary 2) to get that with probability 1,

$$\lim_{T\to\infty}\left|\frac{1}{T}\sum_{t=1}^{T} S(\theta_t^i,\omega_t) - \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}_{\theta_t^i,\omega_t}\left[S(\theta_t^i,\omega_t) \mid h_{t-1}\right]\right| = 0$$

which together with the last inequality implies the limit stated in the proposition. ■

Proposition 7 says that an expert almost never outperforms a better informed expert on the average score in the long run. In general, when using strictly proper scoring rules, the more informed expert will outperform the less informed expert in the long run, however, whether it occurs depends on the probability measure that generates the sequence of states and signals, on how fine or how coarse the property is, and on how informative one expert signal is versus the other.

Non-Bayesian Experts

There are two experts $i = 1, 2$, a finite set of states Ω, and infinitely many time periods $t = 1, 2, \ldots$. At every date t, a state ω_t realizes publicly. Let μ be the probability measure

that generates the sequence of states. In this part, experts are not Bayesian, they do not observe private signals. Instead, some experts have knowledge of μ while others do not. An expert is said to be *informed* when he knows the data generating process μ. Informed experts make truthful reports. As for the case of Bayesian experts, there is an elicitable property of interest, (Θ, F), about the state distribution, and the experts provide forecasts about the property at each date t for date $t+1$.

It is the baseline model used in the literature on forecast testing, adapted to the prediction of properties. A classical question in this literature is how to distinguish between the expert who is informed and the expert who is not. While it may be difficult or even impossible to assess the information quality of an expert in isolation—see, for example, Foster and Vohra (1998) for calibration tests, and Shmaya (2008), Olszewski and Sandroni (2008) for general tests (it is worth noting that the arguments that those papers use to prove the manipulability results continue to hold for the properties considered here)—positive results exist for groups of experts if at least one of the experts in the group is informed, by comparing the forecasts of the different experts, as in Al-Najjar and Weinstein (2008) and Feinberg and Stewart (2008). In this part I illustrate how to use long run average scores to distinguish between the informed and the uninformed expert in the context of property forecasting.

In the one-expert case, it is known that randomization helps uninformed experts to cheat and pretend to be informed. To permit the use of such cheating strategies, I assume that in addition to the history of public state realizations, experts have access to a random number generator, which they may use as they wish to generate forecasts. The history of state and forecast realizations is public. Let h_t be the history from date 1 to date t included.

Let S be any strictly proper scoring rule for (Θ, F). The difference of average scores over the first T periods between expert i and expert j is defined as

$$\Delta S_{ij}(T) = \frac{1}{T}\sum_{t=1}^{T}(S(\theta_t^i, \omega_t) - S(\theta_t^j, \omega_t)),$$

where θ_t^i is the forecast of expert i for date t.

Before stating the formal results, I introduce some definitions. Given a sequence of states, the performance of expert i is said to be *asymptotically as good* as that of expert j when, for all $\epsilon > 0$, $\Delta S_{ij}(T) \geq -\epsilon$ for all T chosen large enough.

For $\delta > 0$, let $\mathcal{P}_\delta$ be the set of distributions "close" to the generating process μ,

$$\mathcal{P}_\delta = \{\nu \in \Delta(\Omega^\infty) \mid \forall t, \omega_1, \ldots, \omega_t, |\nu(\omega_t|\omega_1, \ldots, \omega_{t-1}) - \mu(\omega_t|\omega_1, \ldots, \omega_{t-1})| < \delta\}.$$

The definition resembles the notion of merging of probability measures (Blackwell and Dubins,

1962, Kalai and Lehrer, 1994, Lehrer and Smorodinsky, 1996). Given a sequence of states, a sequence of forecasts $(\theta_t)_{t\geq 1}$ is said to be *approximately correct* if forecasts are arbitrarily accurate except possibly for a proportion of periods that vanishes in the long run: for all $\delta > 0$, there exists a distribution $\nu \in \mathcal{P}_\delta$ such that

$$\lim_{T\to\infty} \frac{1}{T} |\{t \leq T \mid \theta_t \text{ is true under } \nu(\cdot|\omega_1, \ldots, \omega_{t-1})\}| = 1,$$

where $|\mathcal{S}|$ is the cardinality of finite set $\mathcal{S}$.

In the proposition below, expert i denotes either expert 1 or 2, and expert j denotes the other expert.

Proposition 8 *If expert i is informed, then almost surely the following obtains:*

(1) The performance of expert i is asymptotically as good as that of expert j.

(2) If the performance of expert j is asymptotically as good as that of expert i, then the sequence of forecasts of expert j is approximately correct.

Proposition 8 formalizes two facts. First, any informed expert is almost surely guaranteed to maximize his performance in the long run, no matter the strictly proper scoring rule employed for the evaluation. Second, if some expert has a performance that becomes as good, asymptotically, as that of an informed expert, then almost surely that expert makes predictions that are essentially as accurate as those of any informed expert except for a fraction of dates that vanishes to zero in the long run. In that sense, evaluating experts with the average score gives a test that separates the informed experts from charlatans, as long as one of the experts is known to be informed. For the case of reports of the full distribution, Feinberg and Stewart (2008) and Al-Najjar and Weinstein (2008) propose tests respectively based on cross-calibration and a comparison of likelihood ratios that also separates the two types of experts. These tests are more advanced and have stronger properties, my objective here is simply to illustrate some basic properties of the average score, commonly used in practice.

Naturally, the assumption that at least one expert is informed is key. Without that assumption, the difference of average scores is not conclusive: it does not enable us tell if one or both experts are informed. Indeed the two-expert case reduces to the one-expert case when both experts make the same forecasts, and in the latter it is already known that no test can distinguish the informed from the uninformed.

The proof makes use of two simple lemmas proved in Appendix D.

Lemma 1 *For any sequence $(u_t)_{t\geq 1}$ of bounded, nonnegative reals, $\lim_{T\to\infty} \frac{1}{T} \sum_{t\leq T} u_t = 0$ if and only if, for all $\epsilon > 0$, $\lim_{T\to\infty} \frac{1}{T} |\{t \leq T \mid u_t < \epsilon\}| = 1$.*

Lemma 2 *For all $\delta > 0$, there exists $\epsilon > 0$ such that for all property values $\theta, \widetilde{\theta}$, and all state distributions $P \in \Delta(\Omega)$ such that θ is true for P, if $S(\widetilde{\theta}, P) \geq S(\theta, P) - \epsilon$, then there exists $Q \in \Delta(\Omega)$ with $\max_\omega |P(\omega) - Q(\omega)| < \delta$ such that $\widetilde{\theta}$ is true for Q.*

Proof of Proposition 8. If expert i is informed, then

$$\mathrm{E}_{\theta_t^i,\omega_t}\left[S(\theta_t^i,\omega_t) \;\middle|\; h_{t-1}\right] = \max_\theta \mathrm{E}_{\omega_t}\left[S(\theta,\omega_t) \mid h_{t-1}\right] \geq \mathrm{E}_{\theta_t^j,\omega_t}\left[S(\theta_t^j,\omega_t) \;\middle|\; h_{t-1}\right].$$

In addition, by Dawid's calibration theorem (Dawid, 1982), almost surely, for both experts $k = 1, 2$, we have

$$\lim_{T\to\infty}\left|\frac{1}{T}\sum_{t=1}^{T} S(\theta_t^k,\omega_t) - \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}_{\theta_t^k,\omega_t}\left[S(\theta_t^k,\omega_t) \;\middle|\; h_{t-1}\right]\right| = 0.$$

Putting these facts together, we get $\liminf_{T\to\infty} S_{ij}(T) \geq 0$. This proves part (1) of the proposition.

Let us prove part (2). By Part (1) we know that the performance of expert i is asymptotically as good as that of expert j almost surely. When the performance of expert j is also asymptotically as good as that of expert i, $\Delta S_{ij}(T) \to 0$. So applying again Dawid's calibration theorem, we get that almost surely,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left(\mathrm{E}_{\theta_t^i,\omega_t}\left[S(\theta_t^i,\omega_t) \;\middle|\; h_{t-1}\right] - \mathrm{E}_{\theta_t^j,\omega_t}\left[S(\theta_t^j,\omega_t) \;\middle|\; h_{t-1}\right]\right) = 0$$

when the performance of expert j is asymptotically as good as that of expert i.

Since $\mathrm{E}_{\theta_t^i,\omega_t}\left[S(\theta_t^i,\omega_t) \mid h_{t-1}\right] \geq \mathrm{E}_{\theta_t^j,\omega_t}\left[S(\theta_t^j,\omega_t) \;\middle|\; h_{t-1}\right]$, Lemma 1 applies and for all $\epsilon > 0$,

$$\lim_{T\to\infty}\frac{1}{T}\left|\left\{\mathrm{E}_{\theta_t^j,\omega_t}\left[S(\theta_t^j,\omega_t)|h_{t-1}\right] \geq \mathrm{E}_{\theta_t^i,\omega_t}\left[S(\theta_t^i,\omega_t)|h_{t-1}\right] - \epsilon\right\}\right| = 1.$$

Now take any $\delta > 0$. By Lemma 2, there exists $\widetilde{\epsilon} > 0$, and some probability measure $\nu \in \mathcal{P}_\delta$ over sequences of states such that, if

$$\mathrm{E}_{\omega_t}[S(\theta_t^i,\omega_t)|h_{t-1}] \geq \mathrm{E}_{\omega_t}[S(\theta,\omega_t)|h_{t-1}] - \widetilde{\epsilon},$$

then θ is true for state distribution P with $P(\omega_t) = \nu(\omega_t \mid h_{t-1})$. So, if we choose $\epsilon = \widetilde{\epsilon}$ in the limit above, we get that for a fraction of periods that converge to 1 as T grows infinite, the forecasts of expert j are true for the state distributions associated with measure of state sequences ν. And hence the sequence of forecasts of expert j is approximately correct. ■

6 Concluding Remarks

In this paper, I discuss the problem of eliciting or evaluating forecasts of properties of probability distributions. The expert's payoff is controlled via a scoring rule that generalizes the classical probability scoring rules, and whose inputs are the expert forecast and the realization of the random state. There exists a simple geometric characterization of the elicitable properties, i.e., the properties that can be elicited truthfully as a strict best response from any expert whose information on the state distribution is precise enough. Moreover, if an expert does not know the property, then asking the expert for this property or for more information may result in erroneous and inconsistent reports. The proper and strictly proper scoring rules are structured as the nonnegatively weighted average of some functions that are fixed and attached to the property under consideration. For properties that take values in an ordered set, additional characterizations are obtained.

In most of the paper, the emphasis is on the fundamental aspect of the problem. I embrace the canonical setting of Savage (1971). I choose to do so because the existence and characterization of proper and strictly proper scoring rules is the critical element in most applications of this literature, without which results are no longer possible. A focus on a more specific environment would prevent the deployment of the theory to other applications. For example, a number of works study environments in which the main players are experts who provide forecasts in probabilities. In many cases, the results of these works continue to hold, possibly with some adaptation, with forecasts of elicitable properties, because they use features of probability elicitation that are shared with the elicitation of general elicitable properties discussed in the general theory of this paper. For example, the methods of Shmaya (2008) and Olszewski and Sandroni (2008) to obtain impossibility results for the testing of informed versus uninformed experts continues to apply when experts forecast general elicitable properties. For positive results, the cross-calibration method of Feinberg and Stewart (2008), which tests multiple competing experts simultaneously, also applies. The fact that the property is elicitable is important. In the latter case, for example, the result owes to a convexity argument which would not be true of general nonelicitable properties.

This paper focuses on individual forecasts, but a stream of the literature also uses scoring rules in a market context. For example, Ostrovsky (2012) studies the information aggregation properties of financial markets operated either by a batch auction or by a dealer. The dealer setting uses a fixed demand/supply schedule modeled via a strictly proper scoring rule, as in Hanson (2003). With minor modifications, one can derive analogous results with the scoring rules presented in this paper and extend the results of Ostrovsky (2012) to a broader class of securities to aggregate information on some particular properties of a potentially complicated

distribution. It is also worth noting that several classical solutions to betting market designs, for example as proposed by Johnstone (2007) and by Lambert et al. (2015), also transpose directly to the predictions of distribution properties, as the only input required is a proper scoring rule. The problem of designing incentive contracts, as in the work of Clemen (2002), or the problem of designing screening contracts, as in the works of Olszewski and Sandroni (2007) and Babaioff et al. (2011), have solutions that can be adapted to the case of more general property forecasts.

Finally, while many properties are elicitable, there are also a number of properties that are not. Yet, eventually, all properties can be elicited in some indirect fashion under the assumption that the expert is full informed—at worst, we can use the standard methods of elicit the full probability distribution, and then subsequently compute the value of any property we wish to have. But asking for the full distribution may be difficult in practice, or unnecessarily cumbersome, and it is not always required. For example, predictions of the variance on its own cannot be elicited, but we have seen that predictions of the mean and variance together can be elicited, which follows from the observation that the mean and the variance are isomorphic to the first and second moments which are both elicitable. More generally, when a property does not convey enough information to induce truthful reports, we can rely on a finer property. A natural question to ask is what is the smallest amount of information we must ask to obtain truthful answers to what we really want to know, and the characterization obtained in this paper may be a modest first step towards an answer.

A Costly Information

Strict properness ensures that truthful responses are the only best responses. The literature usually focuses the search on this criterion, one major reason being that, if there is a cost of acquiring information, then one can simply scale up a strictly proper scoring rule to motivate the expert to learn that information. In this appendix, I explain how this idea works in the context of property elicitation, and argue that the concept of elicitability captures the distribution properties for which there exists an incentive device that motivates learning of the property.

The model is a small modification of the model of Section 2. There continues to be a finite set of states Ω and a distribution P that governs the draw of the state, but the expert no longer knows P. Instead, both the expert and the elicitor share a common prior on P. Let this common prior be μ. It is a probability measure over the elements of $\Delta(\Omega)$ which represent the possible state distributions. Thus, in this model, both the expert and the elicitor start equally uninformed. In particular, they both believe that the probability of state ω is $\mathrm{E}_{P\sim\mu}[P(\omega)]$. Let $\overline{P}$ denote the corresponding prior state distribution.

Consider a property of interest (Θ, F). The expert is asked to provide a forecast for this property. Before his announcement, the expert faces a choice. He can decide to remain uninformed. Or, he can decide to become informed and learn the true state distribution P, but doing so he incurs cost $c > 0$. After the announcement and once the state realizes, the elicitor pays the expert an amount $\pi(\theta, \omega)$ that is a function of the expert's forecast and the realized state. In this more specific context, I refer to π as payment scheme (as opposed to scoring rule). The expert is risk neutral and seeks to maximize his expected payoff. A payment scheme is *incentive compatible* when it induces the expert to make a truthful prediction. The focus is on priors that are *nondegenerate* in the following sense: every property value is false with nonzero probability.

Proposition 9 *If property (Θ, F) is elicitable, then for every nondegenerate prior μ, there exists at least one incentive-compatible payment scheme such that the expert chooses to become informed.*

Proof. Let S be any strictly proper scoring rule for the property of interest. Because S is strictly proper, for every P, $S(\theta(P), P) \geq S(\theta(\overline{P}), P)$. In addition, because μ is nondegenerate, with positive probability on P, $\theta(\overline{P})$ is false under P, and so $S(\theta(P), P) > S(\theta(\overline{P}), P)$. Hence,

$$\mathrm{E}_{P\sim\mu}[S(\theta(P), P)] > \mathrm{E}_{P\sim\mu}[S(\theta(\overline{P}), P)] = S(\theta(\overline{P}), \overline{P}).$$

Thus, there exists $\lambda > 0$ such that if we use payment scheme $\pi = \lambda S$,

$$\mathrm{E}_{P\sim\mu}[\pi(\theta(P), P)] - \pi(\theta(\overline{P}), \overline{P}) > C.$$

The payment scheme is incentive compatible because S is proper. The left-hand side of the inequality is the difference of expected payoff of an informed expert and the expected payoff of an uninformed expert. The difference being greater than the cost of becoming informed, when facing such payment scheme, the expert is strictly better off choosing to become informed. ■

To simplify I take the extreme situation in which the expert goes from fully uninformed to fully informed. Naturally, by the same logic, an analogous result to Proposition 9 holds when the expert starts already partially informed (whereas the elicitor is fully uninformed), and faces the choice of becoming better informed at some cost. One may also enrich the model with different levels of information at different costs.

What is important for the result of Proposition 9 to hold is that the property be elicitable. Here is an instance of what can happen with a nonelicitable property. Consider an environment with two states, a good and a bad state, and a property that distinguishes between two levels of informativeness of the state distribution. If either the good or bad state occurs with probability greater than or equal to $2/3$, then the distribution is said to be "fairly informative." If the good and bad state each occurs with probability less than or equal to $2/3$, then the distribution is said to be "poorly informative." Clearly, this property is not elicitable, because it trivially fails the test of Theorem 1. Further, it is easily verified that the only proper scoring rules (and thus the only incentive-compatible payment schemes) assign a score that may depend on the state but never depend on the forecast. This impossibility to depend on the forecast implies that the only property that an incentive-compatible payment scheme can elicit is the trivial property that carries no information at all on the state distribution. In this case, it is always impossible to induce the expert to become informed, no matter the prior μ or the cost c.

The example is extreme because the impossibility holds for all priors. In the current setting, depending on the prior, it is not always necessary that the property be elicitable to induce the expert to become informed. The reason is that in this simple environment, when the expert chooses to become informed, he becomes informed fully all at once. So, even if a payment scheme pays the same when the expert reports two particular property values, it may still motivate the expert to become informed if it induces him to disentangle between some other property values, as long as the prior μ makes these other values sufficiently likely to occur. Still, elicitability remains a necessary condition for Proposition 9 to hold.

Proposition 10 *If property* (Θ, F) *is not elicitable, then there exists a nondegenerate prior*

μ such that for all incentive-compatible payment schemes, the expert chooses not to become informed, no matter the cost $c > 0$.

Proof. The key ingredient of the proof is the observation that, if the property is not elicitable, then there exists two different property values, θ_a and θ_b, such that for every proper scoring rule S about (Θ, F), we have $S(\theta_a, P) = S(\theta_b, P)$ for all P that makes θ_a or θ_b true. Note the ordering of quantifiers: this observation is not directly implied by the definition of strict properness. It owes instead to the lattice structure of properties.

Specifically, the same argument used in the proof of Theorem 1 can be used to see that any scoring rule that is proper for (Θ, F) elicits a strictly coarser property (at worst, it elicits the trivial property that includes no information at all). And, if scoring rule S_A elicits property A and scoring rule S_B elicits property B, then scoring rule $S_A + S_B$ elicits the information combined in property A and B together, the join of the two properties. So, since there are finitely many elicitable properties that are coarser than (Θ, F), there exists a finest coarser elicitable property. No scoring rule that is proper for (Θ, F) can elicit more than this finest coarser elicitable property, which remains strictly coarser than (Θ, F). Hence the observation above.

Next, take any prior μ which gives positive weight to θ_a and θ_b, and gives zero weight to other property values. In that case, every incentive-compatible payment scheme yields the same expected payoff to the informed and the uninformed expert, who therefore is strictly better off remaining uninformed to avoid the cost. ■

Note that whenever incentive-compatible schemes exist, payoffs can always be chosen to be nonnegative, i.e., the elicitor pays the expert but the expert never pays the elicitor. This feature is often desired in practice. Below I refer to them as *nonnegative* payment schemes.

There are infinitely many incentive-compatible, nonnegative payment schemes. If the elicitor maximizes or minimizes some objective function, then the set of relevant schemes is refined further. Recall from Theorem 2 that incentive-compatible payment schemes are described by a fixed number of parameters. This number depends only on the property being elicited. And importantly, incentive compatibility, nonnegative payoffs, and the incitation to effort all translate into linear constraints for those parameters. Optimal payment schemes are therefore solutions of optimization problems under linear constraints, an appealing property that simplifies greatly the finding of optimal schemes. In particular, if the elicitor's objective function is linear in the different scores—as in the common case of expected payment minimization—then the optimal schemes are the solution of a linear program.

To illustrate this last point, consider the following example. The elicitor wants to elicit the most likely state of Nature. The prior μ is arbitrary but nondegenerate, in the sense defined above. The elicitor wants to obtain truthful forecasts, wants the expert to acquire

the information, and wants to offer nonnegative payoffs while minimizing the expected payoff at the same time.

From Example 1 we know that this property is elicitable and that the incentive-compatible payment schemes take the form

$$\pi(\theta, \omega) = \begin{cases} \lambda + \kappa(\omega) & \text{if } \theta = \omega, \\ \kappa(\omega) & \text{if } \theta \neq \omega, \end{cases}$$

for κ an arbitrary function of the state and λ a nonnegative real number. The constraint that the payoffs be nonnegative is $\kappa(\omega) \geq 0$ for all ω. The constraint that the expert chooses to become informed is written

$$\mathrm{E}[\pi(\theta(P), P)] \geq c + \pi(\theta(\overline{P}), \overline{P}). \tag{1}$$

We note that the incentive constraint to become informed, Equation (1), does not depend on κ which enters additively in both sides of the inequality. To minimize payments to the expert, the elicitor should set $\kappa = 0$. Then, expected payments are minimized when the constraint (1) is binding, which requires to set

$$\lambda = \frac{c}{\mathrm{E}_{P\sim\mu}[\max_\omega P(\omega)] - \max_\omega \mathrm{E}_{P\sim\mu}[P(\omega)]} > 0,$$

since of course $\mathrm{E}_{P\sim\mu}[\max_\omega P(\omega)] > \max_\omega \mathrm{E}_{P\sim\mu}[P(\omega)]$ for nondegenerate prior μ. This characterizes the unique optimal payment scheme,

$$\pi(\theta, \omega) = \frac{c}{\mathrm{E}_{P\sim\mu}[\max_\omega P(\omega)] - \max_\omega \mathrm{E}_{P\sim\mu}[P(\omega)]} \mathbb{1}\{\omega = \theta\}.$$

B Continuous Properties

While the main body of the paper concerns discrete properties, this appendix concerns properties that take values in a one-dimensional continuum. I focus on the distributions that assign positive probability to every state, and slightly abusing notation I continue to denote by $\Delta(\Omega)$ the set of these distributions.

For technical tractability, I restrict attention to properties (Θ, F) that satisfy three conditions:

Real Valued Θ is a subset of the real line.

No Redundancy The sets $F(\theta)$ are pairwise disjoint.

Continuity Their (unique) property function is continuous and nowhere locally constant.[12]

I refer to these properties as *regular real-valued continuous properties*. As long as we are interested in properties that vary along a single dimension, these assumptions are not very restrictive. Common properties such as the mean and moments, variance and covariance, entropy, skewness, kurtosis, all satisfy the three conditions.

Finally, I also focus the discussion on (strict) properness. Indeed, for the properties that satisfy the three conditions, any scoring rule that is proper (resp. strictly proper) is also order sensitive (resp. strictly order sensitive) for the usual ordering on the real line.

Proposition 11 *Consider a property that satisfies conditions* (Real Valued) *and* (No Redundancy)*, and whose property function is continuous. A scoring rule is proper (resp. strictly proper) if and only if it is order sensitive (resp. strictly order sensitive).*

The proof of Proposition 11 is in Appendix E.

Convexity of the level sets remains a necessary condition as for finite properties. However, unlike the case of finite properties, the continuity condition imposed on the properties examined in this appendix makes this condition sufficient.

Theorem 5 *If (Θ, F) is a regular real-valued continuous property, then it is elicitable if and only if, for all $\theta \in \Theta$, $F(\theta)$ is convex.*

For instance, the mean and moments of a random variable pass the convexity test (part (2) of the theorem)—and so can be elicited via a strictly proper scoring rule. However the variance, skewness and kurtosis fail the convexity test. The covariance of two random variables also fails the convexity test. Finally the entropy, which measures the level of uncertainty contained in a probability distribution, fails the convexity test as well. We therefore cannot properly motivate experts to report values for these properties.

Sketch of Proof for Theorem 5. The proof is in Appendix E, and the idea is as follows. We have already seen how elicitability implies convexity, as already argued by Osband (1985). Now let us get the converse. By continuity, Θ is an interval. Let θ be any property value in the interior of Θ. The distributions over states can be partitioned into three subsets, $\mathcal{D}^{<\theta}$, $\mathcal{D}^{=\theta}$, and $\mathcal{D}^{>\theta}$, that are respectively the sets of distributions whose property value is less than θ, equal to θ, and greater than θ. If we require that every level set of the property be convex, a continuity argument shows that all three sets are convex. The separating hyperplane theorem gives existence of a hyperplane $\mathcal{H}_\theta$ that separates $\mathcal{D}^{<\theta}$ from $\mathcal{D}^{>\theta}$. The property being nowhere locally constant, the hyperplane ends up being the linear span of the set $\mathcal{D}^{=\theta}$. This yields existence of a linear functional on $\mathbf{R}^\Omega$ defined as $L_\theta(f) = \langle g_\theta, f \rangle$ where $g_\theta \in \mathbf{R}^\Omega$.

[12] The property function is not constant on any open set of distributions.

Assume without loss of generality that $L_\theta(P)$ is strictly positive when P's property value is greater than θ, and strictly negative when P's property value is less than θ.

It can be shown that the continuity of the property implies that $\theta \mapsto g_\theta$ is continuous on Θ. Thus we can define

$$S(\theta, \omega) = \int_{\theta_0}^{\theta} g_t(\omega)\, \mathrm{d}t,$$

for an arbitrary θ_0. We get, for all distributions P,

$$\begin{aligned} \mathrm{E}_{\omega \sim P}[S(\theta, \omega)] &= \left\langle \int_{\theta_0}^{\theta} g_t \, \mathrm{d}t, P \right\rangle, \\ &= \int_{\theta_0}^{\theta} \langle g_t, P \rangle \, \mathrm{d}t. \end{aligned}$$

Let θ^* be the unique value of the property under P. The condition imposed on g_θ ensures that, for all $t \in (\theta, \theta^*)$, $\langle g_t, P \rangle = L_t(P) > 0$ (with a symmetric inequality for $t \in (\theta^*, \theta)$), thereby making S strictly proper. ■

Once it is established that the property of interest passes the convexity test of Theorem 5, it remains to design the strictly proper scoring rules. In the characterization below, I impose a smoothness condition on the scoring rules. Since the property varies continuously with the underlying distribution, it is reasonable to require that payments vary smoothly with the expert's prediction. To formalize the idea, I say that a scoring rule S is *regular* if it is uniformly Lipschitz continuous in its first variable: it means there must exist $c > 0$ such that for all $\theta_1, \theta_2 \in \Theta$ and all $\omega \in \Omega$,

$$|S(\theta_1, \omega) - S(\theta_2, \omega)| \leq c|\theta_1 - \theta_2|.$$

Looking at the regular scoring rules as the only acceptable scoring rules does not limit the range of properties to which the characterization applies. Regular strictly proper scoring rules are guaranteed to exist whenever the criteria of Theorem 5 are satisfied. On the other hand this restriction is useful in that it permits a simpler description of the scoring rules.

Assume the property passes the convexity test of Theorem 5. The next result asserts that there exists a particular *base scoring rule*, such that the family of strictly proper scoring rules is fully characterized (up to arbitrary state-contingent payoffs) by integrating the base scoring rule scaled by any nonnegative, nowhere locally zero weight. For proper scoring rules, the weight need only be nonnegative. In addition, the base scoring rule is unique up to a weight factor.

Theorem 6 *Let (Θ, F) be a regular real-valued continuous property such that, for every θ,*

the level set $F(\theta)$ is convex. There exists a bounded scoring rule S_0 such that a regular scoring rule for the property is proper (resp. strictly proper) if, and only if, for all θ and ω,

$$S(\theta, \omega) = \kappa(\omega) + \int_{\theta_0}^{\theta} \xi(t) S_0(t, \omega)\, \mathrm{d}t, \tag{2}$$

for some $\theta_0 \in \Theta$, $\kappa : \Omega \mapsto \mathbf{R}$, and $\xi : \Theta \mapsto \mathbf{R}_+$ a bounded Lebesgue measurable function (resp. and such that, for all $\theta_2 > \theta_1$, $\int_{\theta_1}^{\theta_2} \xi > 0$).

Sketch of proof for Theorem 6. The full proof is in Appendix E, and the idea is as follows. In the proof of Theorem 5, we have built functions $g_\theta \in \mathbf{R}^\Omega$ that are such that, for all θ and all $P \in \Delta(\Omega)$, the value $\langle g_\theta, P \rangle$ is respectively strictly positive when P's property value is greater than θ, strictly negative when it is less than θ, and zero when it equals θ. Now substitute $S_0(t, \omega)$ for $g_t(\omega)$ in (2). For all $P \in \Delta(\Omega)$, and all property values θ, θ^* where θ^* is a correct property value under P,

$$S(\theta^*, P) - S(\theta, P) = \int_{\theta}^{\theta^*} \xi(t) \langle g_t, P \rangle\, \mathrm{d}t, \tag{3}$$

which makes S proper, and even strictly proper with the additional condition of positive integral on ξ.

We now get the converse. First, as $S(\cdot, \omega)$ is assumed to be Lipschitz continuous, it is in particular absolutely continuous. Hence it has an integral representation: there exists a function $G : \Theta \times \Omega \mapsto \mathbf{R}$, where $\theta \to G(\theta, \omega)$ is Lebesgue measurable for every ω such that, for all θ, ω,

$$S(\theta, \omega) = \int_{\theta_0}^{\theta} G(t, \omega)\, \mathrm{d}t,$$

where it is assumed without loss of generality that $S(\theta_0, \cdot) = 0$. Moreover, for all ω, $\theta \mapsto S(\theta, \omega)$ is differentiable for almost every θ and its derivative is given by G. Define the linear functional Ψ_θ on $\mathbf{R}^\Omega$ by

$$\Psi_\theta(f) = \langle G(\theta, \cdot), f \rangle.$$

Hence, outside a set of measure zero, $\theta \mapsto S(\theta, P)$ is differentiable for all distributions P, its derivative at θ being given by $\Psi_\theta(P)$. The second step makes use of the fact that when the expected payoff is maximized, its derivative, when it exists, must be zero. Then, defining the functional

$$\Phi_\theta(f) = \langle g_\theta, f \rangle,$$

an equality $\Psi_\theta = \xi(\theta)\Phi_\theta$ must hold for some $\xi(\theta)$. This is a direct consequence of the fact that the kernel of Φ_θ is the linear span of the distributions having property value θ. And so, in

particular, the kernel of Φ_θ must be included in the kernel of Ψ_θ. ξ must remain nonnegative not to violate the order sensitivity property, which is implied by properness (Proposition 11). If, in addition, S is strictly proper, then the integral of ξ must be strictly positive on every segment, a direct consequence of Equation (3). ■

For example, take $\Omega \subset \mathbf{R}$. The distribution mean is (trivially) continuous, and satisfies the convexity condition of Theorem 5. It therefore admits a strictly proper scoring rule. In fact, we can easily find one: observing that the mean squared error $\mathrm{E}_{\omega\sim P}[(\theta-\omega)^2]$ is minimized precisely when θ equals the mean, paying the expert some positive amount minus the squared error yields a strictly proper scoring rule. Differentiating the quadratic term leads to a possible definition of S_0, $S_0(\theta,\omega)=\omega-\theta$, which is at the heart of the scoring rule design of Osband (1985). Altogether, Theorem 6 gives all the regular proper and strictly proper scoring rules for the mean,

$$S(\theta,\omega)=\kappa(\omega)+\int_a^\theta(\omega-t)\xi(t)\,\mathrm{d}t,$$

that are given in Reichelstein and Osband (1984) and Savage (1971). Another simple example is the dichotomous state space $\Omega=\{0,1\}$. The probability of occurrence of $\omega=1$ is, obviously, a continuous property. Following the above example of the mean and according to Theorem 6, the form of its proper and strictly proper scoring rules is given by

$$S(\theta,\omega)=\kappa(\omega)+\int_0^\theta(\omega-t)\xi(t)\,\mathrm{d}t,$$

which is the Schervish representation of probability scoring rules (Schervish, 1989).

Economic interpretation. Behind the integral representation of scoring rules lies a simple economic interpretation. From the perspective of the risk-neutral expert, being remunerated according to a proper scoring rule is essentially the same as participating to an auction that sells off some carefully designed securities. Assume property values are bounded. Consider the auction that sells securities from a parametric family $\{R_\theta\}_{\theta\in\Theta}$; here $R_\theta(\omega)$ specifies the net payoff of the security R_θ when the realized state of Nature is ω. Net payoff is gross payoff minus initial price, which can be normalized to zero. In this auction, buyers bid on the security parameter. They are asked to bid the maximum value of θ for which they are willing to receive security R_θ. The winner is the bidder with the highest bid (ties are broken arbitrarily). Let us look at the special case in which the expert competes against a dummy bidder, whose bid y is distributed according to some density f. When the state ω^*

materializes, the expert who bids x makes expected profit

$$P(y \leq x)\ \mathrm{E}[R_y(\omega^*) \mid y \leq x] = \int_{y \leq x} R_y(\omega^*) f(y)\,\mathrm{d}y. \tag{4}$$

Take any strictly proper scoring rule S of the form (2). Choosing density $f(y) = \xi(y)/\int_\Theta \xi$ and security $R_y(\omega) = \left(\int_\Theta \xi\right) S_0(y, \omega)$, (4) can be re-written

$$\int_{t \leq x} \xi(t) S_0(t, \omega^*)\,\mathrm{d}t,$$

which is precisely the amount the expert would get through scoring rule S, up to a state-contingent payoff. Conversely, take any bounded density function f nowhere locally zero. The expert's expected profit derived from participation in the auction equals the remuneration he would get with some strictly proper scoring rule. The auction interpretation is especially relevant when used on multiple experts, in which case dummy bidders are not needed. However, with multiple experts, these auctions no longer constitute the only valid incentive devices.

The family of securities that the must be auctioned off depends on the property of interest. When eliciting an event's probability, the goods for sale are securities that pay off the same positive amount if the event occurs and zero otherwise, minus the parameter. When eliciting a distribution's mean, they are securities that pay off a positive factor of the realized value of the underlying variable, minus the parameter. Note that in this special case, the auctions are essentially second-price auctions: for these families of securities, the gross payoff is fixed and buyers, in effect, end up bidding on the price.

When we want experts to choose among a few alternative predictions, we can employ the properties examined in Sections 3 and 4 as approximations of continuous properties. We can combine the results of this appendix and those sections to derive the proper and order-sensitive scoring rules. Let (Θ, F) be a regular real-valued continuous property. Let $\alpha_0, \alpha_n \in \mathbf{R} \cup \{+\infty, -\infty\}$ be respective lower and upper bounds of the possible values for the property, and $\alpha_1 < \cdots < \alpha_{n-1}$ be arbitrarily chosen in the interior of Θ (which is an interval). Consider a finite, approximate version $(\widetilde{\Theta}, \widetilde{F})$ of the continuous property. Instead of the exact value, it rounds up nearby property estimates: it gives an interval of the form $[\alpha_j, \alpha_{j+1}]$ that includes the exact value of the continuous property (Θ, F). $\widetilde{F}([\alpha_j, \alpha_{j+1}])$ is therefore the set of distributions whose values, for the property (Θ, F), lie within $[\alpha_j, \alpha_{j+1}]$.

The collection of intervals, $\widetilde{\Theta}$, is naturally equipped with the ordering $[\alpha_0, \alpha_1] \prec \cdots \prec [\alpha_{n-1}, \alpha_n]$. Suppose there exists a strictly proper scoring rule for the continuous property. Theorem 5 says that each $F(\theta)$ is a hyperplane of $\Delta(\Omega)$. As $\widetilde{F}([\alpha_{i-1}, \alpha_i]) \cap \widetilde{F}([\alpha_i, \alpha_{i+1}]) =$

$F(\alpha_i)$, a direct application of Theorems 3 and 4 yields the following corollary:

Corollary 1 *If property (Θ, F) is elicitable, then there exist strictly proper (and strictly order-sensitive) scoring rules for the approximate property $(\widetilde{\Theta}, \widetilde{F})$. A scoring rule S is strictly proper (or strictly order sensitive) for the approximate property if, and only if,*

$$S([\alpha_j, \alpha_{j+1}], \omega) = \kappa(\omega) + \sum_{i<j} \lambda_i \mathbf{n}_i(\omega),$$

for any function $\kappa : \Omega \mapsto \mathbf{R}$ and strictly positive scalars $\lambda_1, \ldots, \lambda_{n-1}$, $\mathbf{n}_i$ being a positively oriented normal to the hyperplane generated by $F(\alpha_i)$.

Consider, similarly to Example 3, the property that gives, for an event E, some interval $\left[\frac{j-1}{n}, \frac{j}{n}\right]$ that includes the probability of E. Event E's probability is a regular continuous property. The hyperplane of distributions that give probability θ to the event has equation $\sum_{\omega \in E} p(\omega) = \theta$, and $\omega \mapsto \mathbb{1}\{\omega \in E\} - \theta$ is a positively oriented normal. This gives the strictly proper scoring rules

$$S\left(\left[\tfrac{j-1}{n}, \tfrac{j}{n}\right], \omega\right) = \kappa(\omega) + \sum_{i=1}^{n-1} \lambda_i \cdot \left\{ \begin{array}{ll} n-i & \text{if } j > i, \omega \in E \\ 0 & \text{if } j \leq i \\ -i & \text{if } j > i, \omega \notin E \end{array} \right\}.$$

C Proofs of Section 3

In the proofs of this appendix $S(\theta)$ denotes the function $S(\theta, \cdot)$ for a scoring rule $S : \Theta \times \Omega \mapsto \mathbf{R}$. For a subset of $\mathcal{V}$ of a given linear space, $\dim \mathcal{V}$ denotes the dimension of its linear span. A convex polyhedron in a given convex subset $\mathcal{C}$ of a linear space is said to be *nondegenerate* when it has the same dimension as $\mathcal{C}$. Finally, as in the main text, $|\mathcal{S}|$ denotes the cardinality of finite set $\mathcal{S}$.

The proofs make use of the following lemma.

Lemma 3 *If property (Θ, F) is elicitable, then, for all θ, $F(\theta)$ is a nondegenerate closed convex polyhedron of $\Delta(\Omega)$, and, for all $\widetilde{\theta} \neq \theta$, either the intersection of $F(\theta)$ and $F(\widetilde{\theta})$ is a degenerate closed convex polyhedron, or this intersection is empty.*

Proof. The result is implied by Theorem 1, according to which the elements $F(\theta)$, $\theta \in \Theta$, form a power diagram of distributions. ■

C.1 Proof of Theorem 2

The proof uses the following lemma.

Lemma 4 *Let $\mathcal{E}$ be an n-dimensional Hilbert space with an inner product $\langle \cdot, \cdot \rangle$. Let $y_1, \ldots, y_m$ be m vectors that generate $\mathcal{E}$. Consider the two systems of inequalities*

$$\langle y_i, x \rangle \geq 0, \qquad i \in \{1, \ldots, m\} \tag{5}$$

and

$$\langle y_i, x \rangle > 0, \qquad i \in \{1, \ldots, m\}. \tag{6}$$

If both systems admit a nonempty set of solutions, then there exist vectors $s_1, \ldots, s_\ell$ of $\mathcal{E}$ such that the set of solutions of (5) is $\{\lambda_1 s_1 + \cdots + \lambda_\ell s_\ell, \lambda_1, \ldots, \lambda_\ell \geq 0\}$ while the set of solutions of (6) is $\{\lambda_1 s_1 + \cdots + \lambda_\ell s_\ell, \lambda_1, \ldots, \lambda_\ell > 0\}$.

Proof. As (5) is a homogeneous system of weak inequalities, its set of solutions is a cone. Let $\{s_1, \ldots, s_\ell\}$ be a set of directrices of the edges of this cone. As by assumption there exists a nonzero solution, this set is not empty. The parametric form of the solutions of (5) is given by the set $\{\sum_i \lambda_i s_i,\ \lambda_1, \ldots, \lambda_\ell \geq 0\}$ (Eremin, 2002). The remainder of the proof shows that the cone $\mathcal{C} = \{\sum_i \lambda_i s_i,\ \lambda_1, \ldots, \lambda_\ell > 0\}$ is the set of solutions of (6).

Part 1. This part shows that any element of $\mathcal{C}$ is solution of (6).

Each vector s_k of $\{s_1, \ldots, s_\ell\}$ is solution of a $(n-1)$-boundary system of the form

$$\begin{cases} \langle y_i, s_k \rangle = 0, & i \notin I_k, \\ \langle y_i, s_k \rangle > 0, & i \in I_k, \end{cases}$$

for I_k a subset of $\{1, \ldots, m\}$. Let x_0 be a solution of (6). Then x_0 is also solution of (5) and so $x_0 = \sum_i \lambda_i s_i$, with $\lambda_i \geq 0$ for all i. There cannot exist j with $\langle y_j, s_k \rangle = 0$ for all k, otherwise $\langle y_j, x_0 \rangle = 0$ and x_0 would not be solution of (6). Therefore $\cup_k I_k = \{1, \ldots, m\}$.

Let $\widehat{x} \in \mathcal{C}$, with $\widehat{x} = \sum_i \mu_i s_i$, with $\mu_i > 0$ for all i. Since $\cup_k I_k = \{1, \ldots, m\}$, for all j there exists k such that $\mu_k \langle y_j, s_k \rangle > 0$ and $\mu_k \langle y_i, s_k \rangle \geq 0$ for all $i \neq j$. By summation, for all i, $\langle y_i, \widehat{x} \rangle > 0$, and so $\widehat{x}$ is solution of (6).

Part 2. This part shows the converse, that any solution of (6) is in $\mathcal{C}$.

Let $\widehat{x}$ be a solution of (6). Let $\mathcal{B}_0$ be the open ball of diameter δ centered on $\widehat{x}$, and $\mathcal{B}_1$ the open ball of diameter $\frac{3}{4}\delta$ with the same center. If δ is chosen small enough, any vector of $\mathcal{B}_0$ is solution of (6) since its inequalities define an open set of $\mathcal{E}$.

For $\epsilon > 0$, let $t = \epsilon \sum_i s_i$, and let $\mathcal{B}_1' = \mathcal{B}_1 + t$ be ball translated by t. If ϵ is chosen small enough, the open ball $\mathcal{B}_1'$ remains contained in $\mathcal{B}_0$. In this case, $\widehat{x}$, which also belongs to $\mathcal{B}_1'$, is the image of some $x_0 \in \mathcal{B}_1$. As x_0 is solution of (5), we can write $x_0 = \sum_i \lambda_i s_i$, with $\lambda_i \geq 0$ for all i, and hence $\widehat{x} = \sum \mu_i s_i$, with $\mu_i = \lambda_i + \epsilon > 0$ for all i. Therefore $\widehat{x} \in \mathcal{C}$. ■

I return to the proof of Theorem 2. Denote by $\mathcal{S}$ the space of scoring rules, i.e., the linear space of functions $S : \Theta \times \Omega \mapsto \mathbf{R}$, considered as a Hilbert space whose inner product is defined as $\langle S_1, S_2 \rangle = \sum_{\theta,\omega} S_1(\theta,\omega) S_2(\theta,\omega)$.

Part 1. Suppose that property (Θ, F) is elicitable. By definition $S \in \mathcal{S}$ is proper if, and only if, for all $\theta, \widehat{\theta} \in \Theta$,

$$\langle S(\theta), P \rangle = \langle S(\widehat{\theta}), P \rangle \qquad \forall P \in F(\theta) \cap F(\widehat{\theta}), \tag{7}$$

$$\langle S(\theta), P \rangle \geq \langle S(\widehat{\theta}), P \rangle \qquad \forall P \in F(\theta) \backslash F(\widehat{\theta}), \tag{8}$$

with the last inequality being strict if and only if S is strictly proper.

By Lemma 3, for all $\theta \in \Theta$, the level set $F(\theta)$ is a bounded convex polyhedron, and so is the convex hull of a set of vertices $\mathcal{V}_\theta$. We can supplement the set of vertices $\mathcal{V}_\theta$ of each polyhedron $F(\theta)$ by vertices of the other polyhedra that belong to its boundary, in such a way that, for all $\theta, \widehat{\theta} \in \Theta$, and all P belonging to both $F(\theta)$ and $\mathcal{V}_{\widehat{\theta}}$, P also belong to $\mathcal{V}_\theta$. Let us write $\mathcal{V}_\theta$ as $\{P_1^\theta, \ldots, P_{\ell_\theta}^\theta\}$.

Let $S \in \mathcal{S}$ be proper (resp. strictly proper). Let $\theta, \widehat{\theta} \in \Theta$. If $P \in \mathcal{V}_\theta \cap \mathcal{V}_{\widehat{\theta}}$, then $P \in F(\theta) \cap F(\widehat{\theta})$ and so by (7), $\langle S(\theta), P \rangle = \langle S(\widehat{\theta}), P \rangle$. If $P \in \mathcal{V}_\theta \backslash \mathcal{V}_{\widehat{\theta}}$, then $P \in F(\theta)$ and $P \notin F(\widehat{\theta})$, since by construction of $\mathcal{V}_\theta$, $P \in F(\widehat{\theta})$ and $P \in \mathcal{V}_\theta$ implies $P \in \mathcal{V}_{\widehat{\theta}}$. So by (8), $\langle S(\theta), P \rangle \geq \langle S(\widehat{\theta}), P \rangle$ (resp. $\langle S(\theta), P \rangle > \langle S(\widehat{\theta}), P \rangle$).

I now show the sufficiency of these two conditions. Assume that if $P \in \mathcal{V}_\theta \cap \mathcal{V}_{\widehat{\theta}}$, then $\langle S(\theta), P \rangle = \langle S(\widehat{\theta}), P \rangle$, and if $P \in \mathcal{V}_\theta \backslash \mathcal{V}_{\widehat{\theta}}$, then $\langle S(\theta), P \rangle \geq \langle S(\widehat{\theta}), P \rangle$ (resp. $\langle S(\theta), P \rangle > \langle S(\widehat{\theta}), P \rangle$). Let $P \in F(\theta) \cap F(\widehat{\theta})$. Then P is a linear combination of vectors in $\mathcal{V}_\theta$ and $\mathcal{V}_{\widehat{\theta}}$, and since the equality $\langle S(\theta), Q \rangle = \langle S(\widehat{\theta}), Q \rangle$ holds for all vectors Q that belong to these two sets, by linearity $\langle S(\theta), P \rangle = \langle S(\widehat{\theta}), P \rangle$. Now let $P \in F(\theta) \backslash F(\widehat{\theta})$. Then $P = \sum_i \lambda_i P_i^\theta$ for some nonnegative scalars λ_i that sum to one. Since $P \notin F(\widehat{\theta})$, there exists k such that $\lambda_k > 0$ and $P_k^\theta \notin F(\widehat{\theta})$. Hence $P_k^\theta \in \mathcal{V}_\theta \backslash \mathcal{V}_{\widehat{\theta}}$, and $\langle S(\theta), P_k^\theta \rangle \geq \langle S(\widehat{\theta}), P_k^\theta \rangle$ (resp. $\langle S(\theta), P_k^\theta \rangle > \langle S(\widehat{\theta}), P_k^\theta \rangle$). For $i \neq k$, we either have $P_i^\theta \in \mathcal{V}_\theta \cap \mathcal{V}_{\widehat{\theta}}$ or $P_i^\theta \in \mathcal{V}_\theta \backslash \mathcal{V}_{\widehat{\theta}}$, and so in both cases $\langle S(\theta), P \rangle \geq \langle S(\widehat{\theta}), P \rangle$. Hence

$$\langle S(\theta), P \rangle = \sum_i \lambda_i \langle S(\theta), P_i^\theta \rangle \geq \sum_i \lambda_i \langle S(\widehat{\theta}), P_i^\theta \rangle = \langle S(\widehat{\theta}), P \rangle$$

with a strict inequality when S is strictly proper. Therefore, a scoring rule S is proper if, and only if, S is solution of the following finite linear system in the space $\mathcal{S}$,

$$\begin{cases} \langle S(\theta) - S(\widehat{\theta}), P\rangle = 0, & \theta, \widehat{\theta} \in \Theta, P \in \mathcal{V}_\theta \cap \mathcal{V}_{\widehat{\theta}}, \\ \langle S(\theta) - S(\widehat{\theta}), P\rangle \geq 0, & \theta, \widehat{\theta} \in \Theta, P \in \mathcal{V}_\theta \backslash \mathcal{V}_{\widehat{\theta}}, \end{cases} \tag{9}$$

and S is strictly proper if, and only if, S is solution of the system

$$\begin{cases} \langle S(\theta) - S(\widehat{\theta}), P\rangle = 0, & \theta, \widehat{\theta} \in \Theta, P \in \mathcal{V}_\theta \cap \mathcal{V}_{\widehat{\theta}}, \\ \langle S(\theta) - S(\widehat{\theta}), P\rangle > 0, & \theta, \widehat{\theta} \in \Theta, P \in \mathcal{V}_\theta \backslash \mathcal{V}_{\widehat{\theta}}. \end{cases} \tag{10}$$

Part 2. Let $\mathcal{S}_0$ be the space of solutions of the finite system of equalities (in $\mathcal{S}$)

$$\langle S(\theta) - S(\widehat{\theta}), P\rangle = 0, \quad \theta, \widehat{\theta} \in \Theta, P \in \mathcal{V}_\theta \cap \mathcal{V}_{\widehat{\theta}}$$

corresponding to the first part of (9) and (10).

Part 2(a). Let $\mathcal{S}_0^\perp$ be the orthogonal complement of $\mathcal{S}_0$ in $\mathcal{S}$. Let $S \in \mathcal{S}_0$. Then, for any vector X of $\mathcal{S}$, $\langle X, S\rangle = \langle X^{\perp\perp}, S\rangle$, with $X^{\perp\perp} \in \mathcal{S}_0$ and where $X^{\perp\perp} + X^\perp$ is the decomposition of X according to the direct sum $\mathcal{S} = \mathcal{S}_0 \oplus \mathcal{S}_0^\perp$. Therefore, there exists vectors $Y_1, \ldots, Y_m$ in $\mathcal{S}_0$ such that the solutions of (9) in $\mathcal{S}$ are exactly the solutions of the finite system of weak linear inequalities in $\mathcal{S}_0$

$$\langle Y_i, S\rangle \geq 0, \quad i = 1, \ldots, m \tag{11}$$

and the solutions of (10) are the solutions of the finite system of strict linear inequalities in $\mathcal{S}_0$

$$\langle Y_i, S\rangle > 0, \quad i = 1, \ldots, m. \tag{12}$$

Part 2(b). Let $\mathcal{K}$ be the kernel of (11) in $\mathcal{S}_0$, and $\mathcal{K}^\perp$ be its orthogonal complement in $\mathcal{S}_0$. For each Y_i, write $Y_i^{\perp\perp} + Y_i^\perp$ its decomposition according to the direct sum $\mathcal{S}_0 = \mathcal{K} \oplus \mathcal{K}^\perp$.

We can easily describe $\mathcal{K}$: $S \in \mathcal{K}$ if and only if $S \in \mathcal{S}_0$, and if, for all $\theta, \widehat{\theta} \in \Theta$ and all $P \in \mathcal{V}_\theta \backslash \mathcal{V}_{\widehat{\theta}}$, $\langle S(\theta) - S(\widehat{\theta}), P\rangle = 0$. Since $(\mathcal{V}_\theta \cap \mathcal{V}_{\widehat{\theta}}) \cup (\mathcal{V}_\theta \backslash \mathcal{V}_{\widehat{\theta}}) = \mathcal{V}_\theta$, $\mathcal{K}$ is simply the solution of

$$\langle S(\theta) - S(\widehat{\theta}), P\rangle = 0, \quad \theta, \widehat{\theta} \in \Theta, P \in \mathcal{V}_\theta. \tag{13}$$

Any S such that $S(\theta) = S(\widehat{\theta})$ for all $\theta, \widehat{\theta} \in \mathcal{S}$ is solution. By Lemma 3, $F(\theta)$ has as dimension the cardinality of Ω for all θ, and so the linear span of $\mathcal{V}_\theta$ is $\mathbf{R}^\Omega$. Consequently, if S is solution of (13), then $\langle S(\theta) - S(\widehat{\theta}), P\rangle = 0$ for all $\theta, \widehat{\theta}$ and all $P \in \mathbf{R}^\Omega$, implying $S(\theta) = S(\widehat{\theta})$. Hence $\mathcal{K} = \{S \in \mathcal{S} \mid S(\theta, \omega) = S(\widehat{\theta}, \omega)\, \forall \theta \neq \widehat{\theta}\}$.

Part 2(c). Let's consider the following two systems of inequalities in $\mathcal{K}^\perp$:

$$\langle Y_i^\perp, S\rangle \geq 0, \quad i = 1, \ldots, m \tag{14}$$

and

$$\langle Y_i^\perp, S\rangle > 0, \quad i = 1, \ldots, m. \tag{15}$$

If $S \in \mathcal{K}^\perp$, $\langle Y_i, S\rangle = \langle Y_i^\perp, S\rangle$, and the solutions of (11) (resp. (12)) are the elements of $\mathcal{K}$ added to the solutions of (14) (resp. (15)). The systems (14) and (15) have full rank in $\mathcal{K}^\perp$, and since by assumption there exists a strictly proper scoring rule, both admit at least one solution. By Lemma 4, there exist vectors $S_1, \ldots, S_\ell \in \mathcal{K}^\perp$ such that S is solution of (14) (resp. of (15)) if and only if S is a nonnegative (resp. strictly positive) linear combination of $S_1, \ldots, S_\ell$.

Therefore, S is solution of (9) (resp. of (10)) if, and only if, $S = \kappa + \sum_i \lambda_i S_i$, for $\kappa \in \mathcal{K}$ and $\lambda_1, \ldots, \lambda_\ell \geq 0$ (resp. $\lambda_1, \ldots, \lambda_\ell > 0$).

C.2 Proof of Theorem 3

If part. The construction of strictly order-sensitive scoring rules is done in Theorem 6 and in Proposition 1.

Only if part. Let S be a strictly order-sensitive scoring rule.

Step 1. This first step shows that for all i and $j > i+1$, if $P \in F(\theta_i)$ and $P \in F(\theta_j)$ then $P \in F(\theta_{i+1})$. Suppose by contradiction that there exist i and P with $P \in F(\theta_i)$, $P \notin F(\theta_{i+1})$, and $P \in F(\theta_j)$ for some $j > i+1$. By Lemma 3, $F(\theta_i)$ is a convex polyhedron of nonempty relative interior. Since $P \in F(\theta_i)$, there exists a sequence of vectors $\{P_k\}_{k\geq 1}$ of the relative interior of $F(\theta_i)$ that converges to P. By continuity $\lim_{k\to\infty} S(\theta_i, P_k) \to S(\theta_i, P)$. Let $\delta_k = S(\theta_i, P_k) - S(\theta_{i+1}, P_k)$. Since P_k and P both belong to $F(\theta_i)$, but not to $F(\theta_{i+1})$, $\delta_k > 0$, and δ_k converges to $\delta = S(\theta_i, P) - S(\theta_{i+1}, P) > 0$. Therefore $\inf\{\delta_k\}_{k\geq 1} > 0$. Let $\epsilon = \inf\{\delta_k/2\}_{k\geq 1}$. By continuity, there exists K such that

$$|S(\theta_i, P) - S(\theta_i, P_K)| \leq \epsilon/2,$$

and

$$|S(\theta_j, P) - S(\theta_j, P_K)| \leq \epsilon/2,$$

so that, since θ_i and θ_j both contain P, $S(\theta_i, P) = S(\theta_j, P)$ and

$$|S(\theta_i, P_K) - S(\theta_j, P_K)| \leq \epsilon.$$

Hence, $S(\theta_j, P_K) > S(\theta_i, P_K) - \epsilon = S(\theta_{i+1}, P_K) + \delta_K - \epsilon > S(\theta_{i+1}, P_K)$. However, P_K is in the relative interior of $F(\theta_i)$, which means according to Lemma 3 that θ_i is the only true value of the property for P_K. Since $i < i+1 < j$, and S is strictly order sensitive, we should have $S(\theta_{i+1}, P_K) > S(\theta_j, P_K)$, which creates a contradiction.

Step 2. Now let $1 \leq j \leq n-1$. Let $B_j = F(\theta_1) \cup \cdots \cup F(\theta_j)$, and $C_j = F(\theta_{j+1}) \cup \cdots \cup F(\theta_n)$. By Lemma 3, B_j and C_j are polyhedra of dimension $|\Omega|$ and nonempty relative interior, with $B_j \cup C_j = \Delta(\Omega)$. Let $i \leq j < j+1 \leq k$. If $P \in F(\theta_i)$ and $P \in F(\theta_k)$, an iterative application of the claim of Step 1 above yields $P \in F(\theta_i), F(\theta_{i+1}), \ldots, F(\theta_k)$. In particular, $P \in F(\theta_j) \cap F(\theta_{j+1})$. Therefore $B_j \cap C_j = F(\theta_j) \cap F(\theta_{j+1})$. By Lemma 3, the dimension of $F(\theta_j) \cap F(\theta_{j+1})$ is at most $|\Omega| - 1$, so that there is a hyperplane of distributions $\mathcal{H}$ that contains $B_j \cap C_j$. Suppose that there exists a distribution P of $\mathcal{H}$ that does not belong to $B_j \cap C_j$. Since $B_j \cup C_j = \Delta(\Omega)$, $P \in B_j$ or $P \in C_j$. Suppose for example that $P \in B_j$. Then there exists a distribution Q in the relative interior of C_j with $Q \notin \mathcal{H}$. Note that the segment $(P, Q]$ contains only elements of B_j or C_j. Since both sets are closed, the segment intersects $B_j \cap C_j$, which is impossible since $(P, Q]$ does not intersect $\mathcal{H}$. So $B_j \cap C_j$ must be the full hyperplane of distributions $\mathcal{H}$, $\mathcal{H} = B_j \cap C_j = F(\theta_j) \cap F(\theta_{j+1})$, which concludes the proof.

C.3 Proof of Theorem 4

Part 1. Define

$$S(\theta_k, \omega) = \kappa(\omega) + \sum_{1 \leq i < k} \lambda_i \mathbf{n}_i(\omega),$$

with $\lambda_1, \ldots, \lambda_{n-1} \geq 0$, and $\kappa \in \mathbf{R}^\Omega$.

As $\mathbf{n}_k$ is oriented positively, $\langle \mathbf{n}_k, P \rangle \geq 0$ for all $P \in F(\theta_{k+1}), \ldots, F(\theta_n)$, and $\langle \mathbf{n}_k, P \rangle \leq 0$ for all $P \in F(\theta_1), \ldots, F(\theta_k)$. The inequalities are strict if $P \notin F(\theta_k) \cap F(\theta_{k+1})$.

Let $P \in F(\theta_k)$. If $j < k$,

$$\mathrm{E}_{\omega \sim P}[S(\theta_k, \omega)] - \mathrm{E}_{\omega \sim P}[S(\theta_j, \omega)] = \sum_{j \leq i < k} \lambda_i \langle \mathbf{n}_i, P \rangle \geq 0,$$

and, if $j > k$,

$$\mathrm{E}_{\omega \sim P}[S(\theta_k, \omega)] - \mathrm{E}_{\omega \sim P}[S(\theta_j, \omega)] = -\sum_{k \leq i < j} \lambda_i \langle \mathbf{n}_i, P \rangle \geq 0.$$

Therefore S is a proper scoring rule. If, in addition, $\lambda_1, \ldots, \lambda_{n-1} > 0$, the inequalities become strict when $P \notin F(\theta_j)$, making S strictly proper.

Part 2. Now assume S is a proper scoring rule. Then, for all $P \in F(\theta_k) \cap F(\theta_{k+1})$, $1 \leq k < n$, $\langle S(\theta_k), P\rangle = \langle S(\theta_{k+1}), P\rangle$, and so $\langle S(\theta_{k+1}) - S(\theta_k), P\rangle = 0$. Theorem 3 says that $F(\theta_k) \cap F(\theta_{k+1})$ is a hyperplane of $\Delta(\Omega)$. Its linear span is a hyperplane $\mathcal{H}_k$ of $\mathbf{R}^\Omega$, thus $S(\theta_{k+1}) - S(\theta_k) = \lambda_k \mathbf{n}_k$, where $\mathbf{n}_k$ is a normal to $\mathcal{H}_k$ oriented positively.

Let $P \in F(\theta_{k+1})$, $P \notin F(\theta_k)$. As S is proper, $\langle S(\theta_{k+1}), P\rangle \geq \langle S(\theta_k), P\rangle$, so $\lambda_k \langle \mathbf{n}_k, P\rangle \geq 0$. Since $P \notin \mathcal{H}_k$ and $\mathbf{n}_k$ is positively oriented, $\langle \mathbf{n}_k, P\rangle > 0$, implying $\lambda_k \geq 0$ ($\lambda_k > 0$ if S is strictly proper).

Therefore

$$S(\theta_k) = S(\theta_1) + \sum_{1 \leq i < k} (S(\theta_{i+1}) - S(\theta_i)) = \kappa + \sum_{1 \leq i < k} \lambda_i \mathbf{n}_i,$$

with $\kappa = S(\theta_1)$.

C.4 Proof of Proposition 1

Assume the property has a strictly order-sensitive scoring rule with respect to the order relation $\prec$, and let $\theta_1 \prec \cdots \prec \theta_n$ be the elements of the value set of the property. Let S be a proper scoring rule. Theorem 4 shows that S takes the form

$$S(\theta_k, \omega) = \kappa(\omega) + \sum_{1 \leq i < k} \lambda_i \mathbf{n}_i(\omega),$$

with $\lambda_1, \ldots, \lambda_{n-1} \geq 0$. Let $P \in \Delta(\Omega)$. Since the normals are positively oriented, $\langle \mathbf{n}_k, P\rangle \geq 0$ if $P \in F(\theta_{k+1}), \ldots, F(\theta_n)$ and $\langle \mathbf{n}_k, P\rangle \leq 0$ if $P \in F(\theta_1), \ldots, F(\theta_k)$, the inequalities being strict if $P \notin F(\theta_k) \cap F(\theta_{k+1})$. So, for all $\theta, \theta_k, \theta_j$, if $\theta_j \prec \theta_k \prec \theta$ and $P \in F(\theta)$, then

$$\mathrm{E}_{\omega \sim P}[S(\theta_k, \omega)] - \mathrm{E}_{\omega \sim P}[S(\theta_j, \omega)] = \sum_{j \leq i < k} \lambda_i \langle \mathbf{n}_i, P\rangle \geq 0.$$

Similarly, if $\theta \prec \theta_k \prec \theta_j$,

$$\mathrm{E}_{\omega \sim P}[S(\theta_k, \omega)] - \mathrm{E}_{\omega \sim P}[S(\theta_j, \omega)] = -\sum_{k \leq i < j} \lambda_i \langle \mathbf{n}_i, P\rangle \geq 0.$$

Hence S is order sensitive. If S is strictly proper, the λ_i's are strictly positive, making the above inequalities strict, and S becomes strictly order sensitive.

D Proofs of Section 5

D.1 Proof of Lemma 1

Assume without loss of generality that $0 \leq u_t \leq 1$. If

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} u_t = 0,$$

then, for all $\epsilon > 0$,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} u_t \mathbb{1}\{u_t \geq \epsilon\} = 0.$$

As

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{u_t \geq \epsilon\} \leq \frac{1}{T}\sum_{t=1}^{T} \frac{u_t}{\epsilon}\mathbb{1}\{u_t \geq \epsilon\},$$

we have

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{u_t \geq \epsilon\} = 0.$$

For the converse, suppose that $\frac{1}{T}\sum_{t=1}^{T} u_t$ does not converge to zero. Then, there exists $\epsilon > 0$ and a subsequence $(w_t)_{t\geq 1}$ such that, for all T, $\frac{1}{T}\sum_{t=1}^{T} w_t \geq \epsilon$. Hence,

$$\frac{1}{T}\sum_{t=1}^{T} w_t \mathbb{1}\left\{w_t \geq \frac{\epsilon}{2}\right\} \geq \frac{\epsilon}{2},$$

so $\frac{1}{T}\sum_{t=1}^{T} u_t \mathbb{1}\{u_t \geq \epsilon/2\}$ does not converge to zero. As $u_t\mathbb{1}\{u_t \geq \epsilon/2\} \leq \mathbb{1}\{u_t \geq \epsilon/2\}$, $\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{u_t \geq \epsilon/2\}$ does not converge to zero.

D.2 Proof of Lemma 2

For a set of state distributions $\mathcal{P}$, let $\mathcal{P}^c$ be the complement of $\mathcal{P}$ in $\Delta(\Omega)$, and define $\mathcal{B}_\delta(\mathcal{P}) = \{Q \in \Delta\Omega \mid \exists P \in \mathcal{P}, \max_\omega |Q(\omega) - P(\omega)| < \delta\}$.

Fix $\theta, \widetilde{\theta} \in \Theta$. By contradiction, suppose that for some $\delta > 0$ and all $\epsilon > 0$, there exists some $P \in F(\theta)$ such that $P \in \mathcal{B}_\delta(F(\widetilde{\theta}))^c$ and $S(\widetilde{\theta}, P) \geq S(\theta, P) - \epsilon$. Choosing $\epsilon = 1/n$, we can generate a sequence of distributions $(P_n)_{n\geq 1}$ such that $P_n \in F(\theta) \cap \mathcal{B}_\delta(F(\widetilde{\theta}))^c$ and $S(\widetilde{\theta}, P) \geq S(\theta, P) - 1/n$. Observing that $F(\theta) \cap \mathcal{B}_\delta(F(\widetilde{\theta}))^c$ is compact, we can extract a subsequence that converges to some $P_\infty \in F(\theta) \cap \mathcal{B}_\delta(F(\widetilde{\theta}))^c$ and that satisfies, by continuity, $S(\widetilde{\theta}, P_\infty) \geq S(\theta, P_\infty)$. But as S is strictly proper and $P_\infty \in F(\theta)$, it must be the case that $P_\infty \in F(\widetilde{\theta})$, contradicting $P \in \mathcal{B}_\delta(F(\widetilde{\theta}))^c$.

The constant ϵ may depend on the choice of $\theta, \hat{\theta}$, however since there are finitely many such pairs, the result also holds uniformly.

E Proofs of Appendix B

In the proofs of this appendix, I use the following notation. For a finite set $\mathcal{S}$, $|\mathcal{S}|$ continues to denote the cardinality of the set. For an arbitrary function f, $\{f = \alpha\}$ denotes the set $\{x \mid f(x) = \alpha\}$. Given a subset $\mathcal{S}$ of a linear space, $\mathcal{S}^\circ$ denotes the interior of $\mathcal{S}$, $\langle \mathcal{S} \rangle$ its linear span, and $\langle \mathcal{S} \rangle_a$ its affine span. For a scoring rule S, I let

$$\overline{S}_P(t) = \mathrm{E}_{\omega \sim P}[S(t, \omega)].$$

The proofs make use of the following elementary lemmas.

Lemma 5 *If Φ is a linear functional on $\mathbf{R}^\Omega$ such that $\ker \Phi \cap \Delta(\Omega) \neq \emptyset$, then $\ker \Phi$ is the linear span of its intersection with $\Delta(\Omega)$.*

Proof. Let $f_0 \in \Delta(\Omega)$ with $\Phi(f_0) = 0$. Take any $f \in \ker \Phi$. As $f_0 > 0$, if α is chosen large enough, $f + \alpha f_0 > 0$. So, defining $\beta = \|f + \alpha f_0\|_\infty$ and $f_1 = (f + \alpha f_0)/\beta$, we have that $\Phi(f_1) = 0$ and $f_1 \in \Delta(\Omega)$. Hence $f = \beta f_1 - \alpha f_0 \in \langle \ker \Phi \cap \Delta(\Omega) \rangle$. ■

Lemma 6 *If $h : [a, b] \mapsto \mathbf{R}_+$ is a Lebesgue measurable function with $\int h > 0$, then there exists $\epsilon > 0$ such that $\{h \geq \epsilon\}$ has strictly positive measure.*

Proof. As $\{h > 0\}$ is the limit of the monotone increasing sequence of sets $\{h \geq 1/k\}$ and $\{h > 0\}$ has strictly positive measure, the sets $\{h \geq 1/k\}$ must have strictly positive measure as k grows large enough. ■

Lemma 7 *If $h : [a, b] \mapsto \mathbf{R}_+$ is a Lebesgue measurable function that is strictly positive almost everywhere, and $A \subset [a, b]$ is a measurable set of strictly positive measure, then $\int_A h > 0$.*

Proof. As $A \cap \{h > 0\}$ is the limit of the monotone increasing sequence of sets $(A \cap \{h \geq 1/k\})$, for k large enough, the set $A \cap \{h \geq 1/k\}$ must have strictly positive measure, and $\int_A h \geq \lambda(A \cap \{h \geq 1/k\})/k > 0$. ■

E.1 Proof of Proposition 11

The proof is a simple adaptation of Proposition 3 of Nau (1985). Let Γ be the associated property function. Assume S is proper. Let θ_P, θ_Q be two property values, and let P be a

distribution such that $\Gamma(P) = \theta_P$. Consider the case $\theta_P < \theta_Q$ and let R be a distribution such that $\theta_Q \leq \Gamma(R)$. Consider the function $f : \lambda \mapsto \Gamma(\lambda R + (1-\lambda)P)$. Observe that the function is continuous. Noting that $f(0) = \theta_P < \theta_Q \leq \Gamma(R) = f(1)$, there exists some $\lambda_Q \in (0, 1]$ such that $f(\lambda_Q) = \theta_Q$. Let $Q = \lambda_Q R + (1 - \lambda_Q)P$.

As S is proper $S(\theta_P, Q) \leq S(\theta_Q, Q)$. By linearity of the expectation operator, the inequality can be re-written as

$$\begin{aligned}
\lambda_Q S(\theta_P, R) + (1-\lambda_Q) S(\theta_P, P) &\leq \lambda_Q S(\theta_Q, R) + (1-\lambda_Q) S(\theta_Q, P), \\
\frac{1-\lambda_Q}{\lambda_Q}(S(\theta_P, P) - S(\theta_Q, P)) &\leq S(\theta_Q, R) - S(\theta_P, R).
\end{aligned}$$

The left-hand side of the inequality is nonnegative by properness of S, which makes the right-hand side of the inequality nonnegative as well. Hence S is order sensitive. By an analogous procedure, if S is strictly proper, then S is also strictly order sensitive.

E.2 Proof of Theorem 5

Let (Θ, F) the a regular real-valued continuous property and Γ be the associated property function.

Part (1) $\Rightarrow$ (2). Let S be a strictly proper scoring rule. Take $P, Q \in \Delta(\Omega)$, and $0 < \alpha < 1$. Suppose $P, Q \in F(\theta)$. Then, for all $\widehat{\theta} \neq \theta$,

$$\mathrm{E}_{\omega \sim P}[S(\widehat{\theta}, \omega)] \leq \mathrm{E}_{\omega \sim P}[S(\theta, \omega)],$$

and

$$\mathrm{E}_{\omega \sim Q}[S(\widehat{\theta}, \omega)] \leq \mathrm{E}_{\omega \sim Q}[S(\theta, \omega)],$$

and so, by linearity of the expectation operator,

$$\begin{aligned}
\mathrm{E}_{\omega \sim \alpha P + (1-\alpha)Q}[S(\widehat{\theta}, \omega)] &= \alpha\ \mathrm{E}_{\omega \sim P}[S(\widehat{\theta}, \omega)] + (1-\alpha)\ \mathrm{E}_{\omega \sim Q}[S(\widehat{\theta}, \omega)] \\
&\leq \alpha\ \mathrm{E}_{\omega \sim P}[S(\theta, \omega)] + (1-\alpha)\ \mathrm{E}_{\omega \sim Q}[S(\theta, \omega)] \\
&= \mathrm{E}_{\omega \sim \alpha P + (1-\alpha)Q}[S(\theta, \omega)],
\end{aligned}$$

which, by strict properness, implies $\alpha P + (1-\alpha)Q \in F(\theta)$. Hence the convexity of the sets $F(\theta)$.

Part (2) ⇒ (1). First remark that, as Γ is continuous, the set of values taken by the property, Θ, is an interval of the real line. This can be seen by applying the intermediate value theorem to the continuous function $\alpha \mapsto \Gamma(\alpha P + (1-\alpha)Q)$, defined on $[0,1]$ for any $P, Q \in \Delta(\Omega)$.

Step 1. Let us start by showing that if, for all θ, $\{\Gamma = \theta\}$ is convex, then it is also the case that $\{\Gamma \geq \theta\}$, $\{\Gamma > \theta\}$, $\{\Gamma \leq \theta\}$ and $\{\Gamma < \theta\}$ are convex. I prove the first case, the other three work in a similar fashion.

Let $\theta \in \Theta^\circ$, and $P, Q \in \Delta(\Omega)$, with $\Gamma(P) \geq \Gamma(Q) \geq \theta$. Consider the function $f(\alpha) = \Gamma(\alpha P + (1-\alpha)Q)$ defined on $[0,1]$. Note that f is continuous. To prove that $\{\Gamma \geq \theta\}$ is convex, it suffices to show that the image of f is the interval $[\Gamma(Q), \Gamma(P)]$. We already know that $[\Gamma(Q), \Gamma(P)] \subseteq f([0,1])$ by continuity of f, observing that $f(0) = \Gamma(Q)$ and $f(1) = \Gamma(P)$. So let

$$a = \sup\{\alpha \in [0,1] \mid f(\alpha) = \Gamma(Q)\},$$
$$b = \inf\{\alpha \in [0,1] \mid f(\alpha) = \Gamma(P)\}.$$

By continuity of f, the above two sets are closed and nonempty, so $f(a) = f(0) = \Gamma(Q)$ and $f(b) = f(1) = \Gamma(P)$. Also, by convexity of the level sets of Γ, $f([0,a]) = \{\Gamma(Q)\}$ and $f([b,1]) = \{\Gamma(P)\}$. Besides, if, for some $\alpha^* > a$, $f(\alpha^*) < f(0)$ then by continuity $f(\alpha) = f(0)$ for some $\alpha > \alpha^*$, violating a's definition. Similarly, there does not exist α^* with $f(\alpha^*) > f(1)$, and $f([0,1]) = [\Gamma(Q), \Gamma(P)]$. So $\{\Gamma \geq \theta\}$ is convex.

Step 2. Let $\theta \in \Theta^\circ$. Let's start by showing the existence of a nonzero linear functional Φ on $\mathbf{R}^\Omega$, such that

$$\{\Gamma < \theta\} \subset \{\Phi \leq 0\},$$
$$\{\Gamma \geq \theta\} \subset \{\Phi \geq 0\}.$$

By the previous step both $\{\Gamma < \theta\}$ and $\{\Gamma \geq \theta\}$ are convex, and since they are disjoint with nonempty relative interior, we can apply the separating hyperplane theorem and find a nonconstant affine function Φ on the affine span of $\Delta(\Omega)$, $\Phi(\{\Gamma < \theta\}) \leq 0$ and $\Phi(\{\Gamma \geq \theta\}) \geq 0$. Φ naturally extends to a linear functional on $\mathbf{R}^\Omega$.

Step 3. Using the same θ as in the preceding step, as $\{\Gamma < \theta\}$ and $\{\Gamma > \theta\}$ are open sets of $\Delta(\Omega)$, we have that $\{\Gamma < \theta\} \subset \{\Phi < 0\}$ and $\{\Gamma > \theta\} \subset \{\Phi > 0\}$. In summary, we have

existence of a linear functional Φ on $\mathbf{R}^\Omega$ satisfying

$$\{\Gamma < \theta\} \subset \{\Phi < 0\},$$
$$\{\Gamma \geq \theta\} \subset \{\Phi \geq 0\},$$
$$\{\Gamma > \theta\} \subset \{\Phi > 0\}.$$

By a symmetric argument, there exists a linear functional Ψ that satisfies

$$\{\Gamma < \theta\} \subset \{\Psi < 0\},$$
$$\{\Gamma \leq \theta\} \subset \{\Psi \leq 0\},$$
$$\{\Gamma > \theta\} \subset \{\Psi > 0\}.$$

Next we prove that Φ and Ψ are positively collinear. If they are not collinear, then $\ker \Phi \cap \Delta(\Omega) \neq \ker \Psi \cap \Delta(\Omega)$ by Lemma 5. As $\ker \Phi \cap \Delta(\Omega) \subseteq \{\Gamma = \theta\}$ and $\ker \Psi \cap \Delta(\Omega) \subseteq \{\Gamma = \theta\}$, there exist $P, Q \in \{\Gamma = \theta\}$ such that $\Phi(P) = 0$ with $\Psi(P) < 0$, and $\Phi(Q) > 0$ with $\Psi(Q) = 0$. So,

$$\Phi\left(\frac{P+Q}{2}\right) > 0 \quad \text{and} \quad \Psi\left(\frac{P+Q}{2}\right) < 0.$$

By continuity of Φ and Ψ, there exists an open ball $\mathcal{B}$ centered on $(P+Q)/2$, such that $\Phi(\mathcal{B})$ contains only strictly positive values and $\Psi(\mathcal{B})$ contains only strictly negative values. Since $\mathcal{B} \cap \Delta(\Omega) \neq \emptyset$, these assertions imply that Γ is both greater than or equal to θ and less than or equal to θ on $\mathcal{B} \cap \Delta(\Omega)$, and so equals θ on this open set of $\Delta(\Omega)$. This contradicts the regularity assumption on Γ. So Ψ and Φ are collinear, and, by their sign properties above, are positively collinear, implying $\{\Gamma = \theta\} = \ker \Phi \cap \Delta(\Omega)$.

Thus, for all $\theta \in \Theta^\circ$, there exists a linear functional Φ_θ on $\mathbf{R}^\Omega$ such that $\{\Gamma = \theta\} = \ker \Phi_\theta \cap \Delta(\Omega)$.

Step 4. We can choose Φ_θ such that $\|\Phi_\theta\| = 1$, and orient Φ_θ such that $\Phi_\theta(P) > 0$ for some given $P \in \{\Gamma > \theta\}$. By continuity of Γ and convexity of $\Delta(\Omega)$, Φ_θ has the following properties:

$$\{\Gamma < \theta\} = \{\Phi < 0\} \cap \Delta(\Omega),$$
$$\{\Gamma = \theta\} = \{\Phi = 0\} \cap \Delta(\Omega),$$
$$\{\Gamma > \theta\} = \{\Phi > 0\} \cap \Delta(\Omega).$$

Let us write $\Phi_\theta(P)$ as $\langle g_\theta, P \rangle$, for some $g_\theta \in \mathbf{R}^\Omega$.

Step 5. This steps shows that the function $\theta \mapsto g_\theta$ is continuous on Θ°.

Let us begin by showing that, for all $\theta_0 \in \Theta^\circ$, $\lim_{\theta\to\theta_0} \Phi_\theta(f) = 0$ whenever $f \in \ker \Phi_{\theta_0} \cap \Delta(\Omega)$. To see this, let $f \in \{\Gamma = \theta_0\}$, and, for any $\epsilon > 0$, consider the open ball $\mathcal{B}_\epsilon$ of radius ϵ that is centered on f. Note that Φ_{θ_0} takes both strictly positive and strictly negative values on $\mathcal{B}_\epsilon$, meaning that Γ takes values that are both above and below θ_0. By continuity of Γ, there exists some $\delta > 0$ such that $(\theta_0-\delta, \theta_0+\delta) \subset \Gamma(\mathcal{B}\cap\Delta(\Omega))$. In particular, for all $\theta \in (\theta_0-\delta, \theta_0+\delta)$, there is $g \in \mathcal{B}_\epsilon \cap \Delta(\Omega)$ with $\Gamma(g) = \theta$, hence $|\Phi_\theta(f)| = |\Phi_\theta(f-g) + \Phi_\theta(g)| \leq \|\Phi_\theta\|\|f-g\| \leq \epsilon$. Therefore, we have that $\lim_{\theta\to\theta_0} \Phi_\theta(f) = 0$.

Observing that $\ker \Phi_{\theta_0} = \langle \ker \Phi_{\theta_0} \cap \Delta(\Omega)\rangle$ by Lemma 5, the above limit remains valid whenever $f \in \ker \Phi_{\theta_0}$.

Now we can extend the limit to all members of $\mathbf{R}^\Omega$. Take any sequence $\{\theta_k\}$ that converges to θ_0. Then, because g_θ is finite-dimensional and bounded, Φ_{θ_k} converges, uniformly, on a subsequence of k's. Suppose that Φ^∞ is the limit. We have just shown that $\ker \Phi_{\theta_0} \subseteq \ker \Phi^\infty$, which implies that $\Phi^\infty = \alpha\Phi_{\theta_0}$ for some α. Since $\|\Phi_{\theta_k}\| = 1$ by assumption, $\|\Phi^\infty\| = 1$, so $|\alpha| = 1$, and the orientation that was decided of Φ_θ yields $\alpha = 1$. If for any $f \in \mathbf{R}^\Omega$, it was the case that $\Phi_{\theta_k}(f)$ did not converge to $\Phi_{\theta_0}(f)$, then for a subsequence of k's, we would have that Φ_{θ_k} converges to a functional different from Φ_{θ_0}, which we have just ruled out.

Step 6. At last we can construct a strictly proper scoring rule. We let $H(t,\omega) = g_t(\omega)$ if $t \in \Theta^\circ$ and, using that g_t is bounded and continuous, we extend $H(t,\omega)$ by continuity on the entire interval Θ.

Choose any $\theta_0 \in \Theta$ and let

$$S(\theta,\omega) = \int_{\theta_0}^{\theta} H(t,\omega)\,\mathrm{d}t.$$

We have that

$$\begin{aligned}\mathrm{E}_{\omega\sim P}[S(\theta,\omega)] &= \left\langle \int_{\theta_0}^{\theta} H(t,\cdot)\,\mathrm{d}t, P \right\rangle, \\ &= \int_{\theta_0}^{\theta} \langle H(t,\cdot), P\rangle\,\mathrm{d}t.\end{aligned}$$

Suppose for example that $\Gamma(P) > \theta$, then

$$\begin{aligned}\mathrm{E}_{\omega\sim P}[S(\Gamma(P),\omega)] - \mathrm{E}_{\omega\sim P}[S(\theta,\omega)] &= \int_{\theta}^{\Gamma(P)} \langle (H(t,\cdot), P\rangle\,\mathrm{d}t, \\ &> 0\end{aligned}$$

since, for all $t < \Gamma(P)$, $\langle H(t,\cdot), P\rangle = \Phi_t(P) > 0$. And similarly for $t > \Gamma(P)$. Hence S is strictly proper.

E.3 Proof of Theorem 6

If part. In the proof of Theorem 5, we construct a function $H(\theta, \omega)$, that satisfies $|H| \leq 1$, and such that, for all θ and $P \in \Delta(\Omega)$,

$$\langle H(\theta, \cdot), P \rangle$$

is strictly positive when $\Gamma(P) > \theta$, strictly negative when $\Gamma(P) < \theta$, and zero when $\Gamma(P) = \theta$. Let us set $S_0 = H$. Assume that for all ω and θ, scoring rule S takes the form

$$S(\theta, \omega) = \kappa(\omega) + \int_{\theta_0}^{\theta} \xi(t) S_0(t, \omega)\, \mathrm{d}t,$$

for some $\theta_0 \in \Theta$, $\kappa : \Omega \mapsto \mathbf{R}$, and $\xi : \mathcal{I} \mapsto \mathbf{R}_+$ a Lebesgue measurable bounded function.

For all $P \in \Delta(\Omega)$,

$$\overline{S}_P(\theta) = \langle \kappa, P \rangle + \left\langle \int_{\theta_0}^{\theta} \xi(t) S_0(t, \cdot), P \right\rangle.$$

Take, for example, $\theta < \Gamma(P)$:

$$\overline{S}_P(\Gamma(P)) - \overline{S}_p(\theta) = \int_{\theta}^{\Gamma(P)} \xi(t) \langle S_0(t, \cdot), P \rangle\, \mathrm{d}t.$$

As, for all $t < \Gamma(P)$, $\langle S_0(t, \cdot), P \rangle > 0$, we get $\overline{S}_P(\Gamma(P)) - \overline{S}_P(\theta) \geq 0$, implying that S is proper. If, in addition, $\int_{\theta}^{\Gamma(P)} \xi > 0$, then by Lemma 6, there is $\epsilon > 0$ such that $A = \{\xi \geq \epsilon\}$ is of strictly positive Lebesgue measure. Hence,

$$\overline{S}_P(\Gamma(P)) - \overline{S}_P(\theta) \geq \epsilon \int_A \langle S_0(t, \cdot), P \rangle\, \mathrm{d}t$$

which is strictly positive by Lemma 7, making S strictly proper.

Only if part. Let S be a regular scoring rule for Γ, and $\theta_0 \in \Theta$. If S is proper (resp. strictly proper), $(\theta, \omega) \mapsto S(\theta, \omega) - S(\theta_0, \omega)$ is also proper (resp. strictly proper). Thus we can assume with loss of generality that $S(\theta_0, \cdot) = 0$.

As $S(\cdot, \omega)$ is Lipschitz continuous, it is also absolutely continuous and there is a function $G : \Theta \times \Omega \mapsto \mathbf{R}$ such that, for all θ, ω,

$$S(\theta, \omega) = \int_{\theta_0}^{\theta} G(t, \omega)\, \mathrm{d}t.$$

Moreover, for all ω, $\theta \mapsto S(\theta, \omega)$ is differentiable except possibly on a measure zero set $\mathcal{Z}$, and

$$\frac{S(\theta, \omega)}{\partial \theta} = G(\theta, \omega).$$

(The measure zero set generally depend on ω, but as ω only takes a finite number of values, we can always choose $\mathcal{Z}$ to be independent of ω.) G can be chosen such that, if $S(\cdot, \omega)$ is not differentiable at θ, $G(\theta, \omega) = 0$. Finally, as S is Lipschitz continuous, G is bounded.

For all $\theta \in \Theta$, define Ψ_θ on $\mathbf{R}^\Omega$ as $\Psi_\theta(f) = \langle G(\theta, \cdot), f \rangle$.

Assume S is proper, and let $\theta \notin \mathcal{Z}$. If $P \in \{\Gamma = \theta\}$, $\Gamma(P) \notin \mathcal{Z}$ and so $\overline{S}_P(\Gamma(P))' = 0$, which yields $\{\Gamma = \theta\} \subset \ker \Psi_\theta$. By Theorem 5, there exists a linear functional Φ_θ on $\mathbf{R}^\Omega$ such that $\{\Gamma = \theta\} = \ker \Phi_\theta \cap \Delta(\Omega)$. As $\{\Gamma = \theta\}$ is nonempty, applying Lemma 5 yields $\ker \Phi_\theta = \langle \{\Gamma = \theta\} \rangle$ and, as $\{\Gamma = \theta\} \subset \ker \Psi_\theta$ we have that $\ker \Phi_\theta \subseteq \ker \Psi_\theta$. Consequently there exists a real number $\xi(\theta)$ such that $\Psi_\theta = \xi(\theta)\Phi_\theta$. Choose $\xi(\theta) = 0$ if $\Phi_\theta = 0$ or if $\theta \in \mathcal{Z}$.

We can choose without loss $\|\Phi_\theta\| = 1$. In the proof of Theorem 5, we showed that Φ_θ can be chosen such that $\theta \mapsto \Phi_\theta(P)$ be continuous. Writing $\xi(\theta) = \Psi_\theta / \Phi_\theta$ leads to Lebesgue measurability of ξ. Besides, noting that $\|G(\theta, \cdot)\| = \|\Psi_\theta\| = |\xi(\theta)| \|\Phi_\theta\| = |\xi(\theta)|$, boundedness of ξ follows from boundedness of G.

Therefore, for all $P \in \Delta(\Omega)$, and all θ,

$$\overline{S}_P(\theta) = \int_{\theta_0}^{\theta} \Psi_t(P)\, \mathrm{d}t = \int_{\theta_0}^{\theta} \xi(t)\Phi_t(P)\, \mathrm{d}t.$$

By Proposition 11, S is order sensitive. This implies $\xi \geq 0$. Indeed, suppose $\xi(\theta) < 0$ for some $\theta \notin \mathcal{Z}$. Take, for example, $P \in \{\Gamma > \theta\}$. Then,

$$\overline{S}'_P(\theta) = \xi(\theta)\Phi_\theta(P) < 0,$$

and $\overline{S}_P$ is not weakly increasing on $\{t < \Gamma(P)\}$, contradicting order sensitivity of S. Hence $\xi \geq 0$. Assume that, in addition, S is strictly proper. Take any $\theta_1 < \theta_2$ and $P \in \{\Gamma = \theta_2\}$. Then,

$$\begin{aligned} 0 < |\overline{S}_P(\theta_2) - \overline{S}_P(\theta_1)| &= \left| \int_{\theta_1}^{\theta_2} \xi(t)\Phi_t(P)\, \mathrm{d}t \right|, \\ &\leq \|P\| \int_{\theta_1}^{\theta_2} \xi, \end{aligned}$$

implying $\int_{\theta_1}^{\theta_2} \xi > 0$.

References

Abernethy, J. and R. Frongillo (2012). A characterization of scoring rules for linear properties. In *the 25th Annual Conference on Learning Theory (COLT)*.

Al-Najjar, N. and J. Weinstein (2008). Comparative testing of experts. *Econometrica 76*(3), 541–559.

Aurenhammer, F. (1987). Power diagrams: Properties, algorithms and applications. *SIAM Journal on Computing 16*(1), 78–96.

Aurenhammer, F. (1991). Voronoi diagrams: A survey of a fundamental geometric data structure. *ACM Computing Surveys 23*(3), 345–405.

Babaioff, M., L. Blumrosen, N. Lambert, and O. Reingold (2011). Only valuable experts can be valued. In *the 12th ACM Conference on Economics and Computation (EC)*.

Blackwell, D. (1951). Comparison of experiments. In *the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 93–102.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics 24*(2), 265–272.

Blackwell, D. and L. Dubins (1962). Merging of opinions with increasing information. *The Annals of Mathematical Statistics 33*(3), 882–886.

Bonin, J. (1976). On the design of managerial incentive structures in a decentralized planning environment. *American Economic Review 66*(4), 682–687.

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

Chambers, C. and N. Lambert (2018). Dynamic belief elicitation. Working paper.

Clemen, R. (2002). Incentive contracts and strictly proper scoring rules. *TEST 11*(1), 167–189.

Dawid, A. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association 77*(379), 605–610.

De Berg, M., O. Cheong, and M. van Kreveld (2008). *Computational Geometry: Algorithms and Applications*. Springer.

De Finetti, B. (1962). Does it make sense to speak of "good probability appraisers"? In I. Good (Ed.), *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, pp. 257–364. Basic Books.

Eremin, I. (2002). *Theory of Linear Optimization*. VSP International Science Publishers.

Fan, L.-S. (1975). On the reward system. *American Economic Review 65*(4), 226–229.

Feinberg, Y. and C. Stewart (2008). Testing multiple forecasters. *Econometrica 76*(3), 561–582.

Fissler, T. and J. Ziegel (2016). Higher order elicitability and osband's principle. *Annals of Statistics 44*(4), 1680–1707.

Foster, D. and R. Vohra (1998). Asymptotic calibration. *Biometrika 85*(2), 379–390.

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

Frongillo, R. and I. Kash (2015a). On elicitation complexity. In *Advances in Neural Information Processing Systems 28 (NIPS)*.

Frongillo, R. and I. Kash (2015b). Vector-valued property elicitation. In *the 28th Annual Conference on Learning Theory (COLT)*.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics 18*(2), 141–153.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*(494), 746–762.

Gneiting, T. and A. Raftery (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B 14*(1), 107–114.

Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers 5*(1), 107–119.

Imai, H., M. Iri, and K. Murota (1985). Voronoi diagram in the Laguerre geometry and its applications. *SIAM Journal on Computing 14*(1), 93–105.

Johnstone, D. (2007). The parimutuel Kelly probability scoring rule. *Decision Analysis 4*(2), 66–75.

Kalai, E. and E. Lehrer (1994). Weak and strong merging of opinions. *Journal of Mathematical Economics 23*(1), 73–86.

Lambert, N., J. Langford, J. Wortman Vaughan, Y. Chen, D. Reeves, Y. Shoham, and D. Pennock (2015). An axiomatic characterization of wagering mechanisms. *Journal of Economic Theory 156*, 389–416.

Lehrer, E. and R. Smorodinsky (1996). Compatible measures and merging. *Mathematics of Operations Research 21*(3), 697–706.

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America 42*(9), 654–655.

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

Olszewski, W. and A. Sandroni (2007). Contracts and uncertainty. *Theoretical Economics 2*(1), 1–13.

Olszewski, W. and A. Sandroni (2008). Manipulability of future-independent tests. *Econometrica 76*(6), 1437–1466.

Osband, K. (1985). *Providing Incentives for Better Cost Forecasting*. Ph. D. thesis, University of California, Berkeley.

Osband, K. and S. Reichelstein (1985). Information-eliciting compensation schemes. *Journal of Public Economics 27*(1), 107–15.

Ostrovsky, M. (2012). Information aggregation in dynamic markets with strategic traders. *Econometrica 80*(6), 2595–2647.

Reichelstein, S. and K. Osband (1984). Incentives in government contracts. *Journal of Public Economics 24*(2), 257–270.

Sandroni, A. (2003). The reproducible properties of correct forecasts. *International Journal of Game Theory 32*(1), 151–159.

Savage, L. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66*(336), 783–801.

Schervish, M. (1989). A general method for comparing probability assessors. *Annals of Statistics 17*(4), 1856–1879.

Shiryaev, A. (1996). *Probability, 2nd ed.* Springer-Verlag, New York.

Shmaya, E. (2008). Many inspections are manipulable. *Theoretical Economics 3*(3), 367–382.

Thomson, W. (1979). Eliciting production possibilities from a well informed manager. *Journal of Economic Theory 20*(3), 360–380.