Elicitation and Evaluation of Statistical Forecasts

Nicolas S. Lambert[*][†]

June 13, 2011

Abstract

This paper studies mechanisms for eliciting and evaluating statistical forecasts. Nature draws a state at random from a given state space, according to some distribution p. Prior to Nature's move, an expert, who knows p, provides a forecast for a given statistic of p. The mechanism defines the expert's payoff as a function of the forecast and the subsequently realized state. When the statistic is continuous with a continuum of values, I show that the payoffs that provide strict incentives to the expert exist if and only if the statistic partitions the set of distributions into convex subsets. When the underlying state space is finite, and the statistic is discrete with a finite number of values, these payoffs exist if and only if the partition forms a linear cross-section of a Voronoi diagram—a stronger condition than convexity. In both cases, the payoffs can be fully characterized essentially as weighted averages of base functions. I then analyze mechanisms with stronger properties. Finally I describe several applications.

[*] Stanford University Graduate School of Business; email address: nlambert@stanford.edu.

[†] I am deeply indebted to David Pennock and Yoav Shoham for their guidance and invaluable advice. I am also very grateful to Yahoo! Research, where I developed the early version of this work. I thank Google Research and the National Science Foundation (grant CCF-1101209) for their generous support. I am very grateful to Jeremy Bulow, Peter Cramton, Yossi Feinberg, Drew Fudenberg, David Kreps, John Langford, Mohammad Mahdian, Roger Myerson, Michael Ostrovsky, Philip Reny, Alvin Roth, Hugo Sonnenschein, Andrzej Skrzypacz, Balázs Szentes and Robert Wilson for their insightful comments. I thank the participants of the economics, computer science and operations research seminars at Chicago, Columbia GSB, Ecole Polytechnique, Harvard, MIT Sloan, NYU Stern, Stanford GSB, UC Berkeley and Wharton.

1 Introduction

A decision maker often has less information relevant for her decision than does some other agent (an expert). In this paper, I examine mechanisms for eliciting and evaluating information when such information consists of statistical forecasts. Given a state space Ω, Nature draws a state ω^* at random according to some probability distribution p. Before Nature chooses a state, an expert, who knows p, announces a forecast for a "statistic" of interest to the decision maker. In this paper a "statistic" has a fairly general meaning. It encompasses most uses of the term and captures all features a distribution can possess. In particular, a "statistic" can be a real-valued statistic such as the mean, median or variance of a random variable defined on Ω. It can also be multidimensional: for instance, it could be a confidence interval, or a variance-covariance matrix. Finally, it can be a discrete statistic that takes value in a more general set; for example, a ranking of events by likelihood, or a pair of most correlated components of a random vector.

The mechanisms I consider in this paper are scoring rules. A scoring rule assigns to the expert a payoff as a function of his forecast and the one realization of the state of nature, ω^*, that is observed ex-post. I focus on *strictly proper scoring rules*, scoring rules that give the expert (who seeks to maximize his expected payoff) a strict incentive to report truthfully: the expert maximizes his expected payoff if, and only if, he makes a truthful forecast.

I address two central questions. First, for which statistics does a strictly proper scoring rule exist? Second, for a given statistic, how can you construct such a strictly proper scoring rule?

I show that for some statistics, a strictly proper scoring rule does not exist. There are many cases for which the declared forecasts do not procure enough information to grant existence of strict incentives.[1] But there are also many cases for which it is enough.

I distinguish between two classes of statistics: the statistics that are continuous and take a continuum of values (for example, the mean or variance of a random variable), and the statistics that are discrete and take a finite number of values (for example, a ranking of events in order of likelihood, or the most correlated pair of

[1] Note that, in this setting, the expert reports only a value for the statistic of interest, he does not report the full probability distribution.

components in a random vector). For continuous statistics, strict properness can be achieved if and only if the statistic partitions the space of probability distributions into convex sets. That is, each level set of the statistic—those sets of distributions assigned to the same statistic value—must be convex. (So, for example, a strictly proper scoring rule exists for the mean, but not for the variance.) For discrete statistics, strictly proper scoring rules exist if and only if they partition the distributions into pieces of a Voronoi diagram, a geometric object more constraining than convex partitions. (I will show that this implies, for example, that a strictly proper scoring rule can be constructed for the ranking of events, but not for the most correlated components.) In both cases, the proper and strictly proper scoring rules are generated by the (continuous or finite) mixtures of some given functions that are entirely—and often uniquely—determined by the statistic itself.

For statistics that take value in a set endowed with a natural ordering of its elements (for example, the mode and the median of a random variable, that take real values), I introduce an alternative class of scoring rules, the *(strictly) order-sensitive scoring rules*. Order sensitivity means that the closer the forecast is to the true value of the statistic, the larger the expected payoff. For continuous statistics, strict order sensitivity is a property that all strictly proper scoring rules possess. The result carries over to discrete statistics, albeit in a limited sense: when they exist, the strictly order-sensitive scoring rules are exactly those that are strictly proper. However, the discrete statistics that admit a strictly order-sensitive scoring rule form only a small subset of all those that admit a strictly proper scoring rule. These statistics partition the distributions into "slices" separated by hyperplanes, a much stronger constraint. (For example, a strictly order-sensitive scoring rule can be constructed for the median, but not for the mode.)

The literature on forecast evaluation and elicitation goes back to Brier (1950). Brier envisioned a scheme to measure the accuracy of probability assessments for a set of events, in the context of weather forecasting. This scheme, the Brier score, was later recognized as part of a much larger family of functions that possess similar properties, the proper and strictly proper probability scoring rules. These scoring rules and their properties have been exhaustively studied (McCarthy, 1956; De Finetti, 1962; Winkler and Murphy, 1968; Winkler, 1969; Hendrickson and Buehler, 1971; Savage, 1971; Schervish, 1989; Good, 1997; Selten, 1998; Dawid, 2007; Gneiting and Raftery, 2007). A vast majority of the literature has concentrated on eliciting the full information about p. [2] In contrast, this paper deals with predictions regarding partial information in a general sense.

As probability scoring rules elicit the full distribution, any statistic can be theoretically elicited in some indirect fashion. As a practical matter though, eliciting the entire distribution becomes rapidly infeasible as the underlying state space grows large. This can be observed on two grounds. First, describing densities in full entails an amount of communication that, if benign for spaces that comprise only few states, is prohibitive in terms of effort or technological cost for spaces whose states vary along several dimensions. Second, if, to gain sufficient knowledge, the expert must spend a substantial number of hours in study and investigation, it is conceivable that the capability to estimate accurately the entire distribution comes at a much greater cost than learning some of its specific features. This paper does not attempt to model these phenomena, but they provide justification for studying ways to elicit a particular statistic of interest rather than the whole distribution.

This paper embraces the setting of Savage's seminal work (Savage, 1971). Like Savage, in the main part of this paper I concentrate on the basic setting and abstract away from common problems, such as unknown risk-attitudes (Karni, 2009; Offerman et al., 2009), costly access to information (Osband, 1989), or experts endowed with different levels of information or skills (Olszewski and Sandroni, 2007, 2010). Most often, the techniques to solve these problems extend directly to the broader context of statistical forecasts, provided we can build a strictly proper scoring rule. (The final section of this paper briefly considers some of these questions.)

Another branch of the literature is concerned with the testing of forecasts. The typical setting assumes an elicitor who repeatedly interacts with a self-proclaimed expert. At each time period, the expert provides a probability assessment over outcomes that materialize at the following period. For instance, a weather expert is asked, every day, to predict the probability of rain for the following day. The main question of interest is whether, by looking at the sequence of forecasts and the sequence of realizations, the elicitor can make the distinction between a truly knowledgeable expert and an impostor. Naturally, if the elicitor were able to assess the exactitude of

[2] The exceptions I am aware of concern Fan (1975), Bonin (1976), and Thomson (1979) who design compensation schemes to elicit, from local branch managers, the production output that can be attained with some given probability. Also, in the context of government contracting, Reichelstein and Osband (1984) and Osband and Reichelstein (1985) establish incentive contracts that induce a contracting firm to reveal truthfully moments of its prior on the costs associated to the project.

the forecasts, she would be able to enforce appropriate incentives to tell the truth, at least, in a repeated setting. However, such tests do not exist: a main result of the literature essentially asserts that for every test that the true, knowledgeable experts passes, there exists a scheme that, when employed by an expert who ignores all of the state probabilities, makes him successfully pass with arbitrarily high probability (papers of this sort include Foster and Vohra (1998); Fudenberg and Levine (1999); Lehrer (2001); Sandroni et al. (2003); Olszewski and Sandroni (2008); Shmaya (2008); Olszewski and Sandroni (2009c)). More positive results exist under variants of the basic setting, notably with multiple competing experts (see Feinberg and Stewart, 2006; Al-Najjar and Weinstein, 2008)[3], when the expert's computational abilities are sufficiently limited (Fortnow and Vohra, 2009), by imposing strong enough restrictions on the class of distributions Nature can choose amongst (Al-Najjar et al., 2010), or when the expert supplies all of his probability assessments upfront and the tester is allowed to make use of counterfactual predictions (Dekel and Feinberg, 2006; Olszewski and Sandroni, 2009a).

In this paper there is a single expert, who is known to be informed, and known to maximize his expected transfer. On the other hand, I impose no restriction on computational abilities nor on the distributions of Nature, there is a single data point (the one realization ω^* of the state) and I require strict incentives for truth-telling.

This paper proceeds as follows. Section 2 details the model. Sections 3 and 4 focus respectively on continuous and discrete statistics. Section 5 discusses some applications and extensions. Section 6 concludes. Full proofs are relegated to the Online Appendix.

2 Preliminaries

Denote by Ω the space of states of Nature, and by $\mathcal{G}$ a σ-algebra of possible events. Let $\Delta\Omega$ be a set of probability distributions over $(\Omega, \mathcal{G})$, that comprises all the distributions considered possible.

Statistics. This paper is concerned with the elicitation and evaluation of statistical forecasts. I give to statistics a broad meaning: they can describe any feature that a

[3] Although as noted by Olszewski and Sandroni (2009b), it is vitally important that one of the experts be a true expert.

distribution may possess. Formally, a *statistic* is defined as a pair (Θ, F). Θ is the *value set* of the statistic. It is a set in which the statistic takes value. F is the *level-set function.* For each statistic value θ, F records the collection $F(\theta)$ of all distributions of $\Delta\Omega$ under which θ is a correct value. For instance, the mean of a random variable X could be written as a pair $(\mathbb{R}, F)$. Here a distribution p belongs to $F(m)$ if, and only if, the mean of X under p, $\int_\omega X(\omega)\mathrm{d}p(\omega)$, equals m. For expositional convenience, in the sequel Θ contains no more elements than the feasible values for the statistic.

Many statistics assign a unique value to each distribution. If so, the level sets $F(\theta), \theta \in \Theta$, are pairwise disjoint. Such statistics are said to *exhibit no redundancy.* The mean, variance, or entropy share this property. Other statistics may associate different values to the same distribution. If so, the level sets overlap, and the statistic contains some amount of redundancy. For instance, the median exhibits some redundancy: in some cases, several median values may be associated with the same distribution. A *statistic function* is defined as a function Γ that associates a true statistic value $\Gamma(p)$ to each distribution p (formally, $\Gamma^{-1}(\theta) \subseteq F(\theta)$ for each θ). When the statistic has no redundancy, it is associated with only one statistic function. This function then suffices to conveniently represent the statistic. In general however, one needs at least two statistic functions to fully describe a statistic.

I restrict the analysis to the statistics (Θ, F) that satisfy two conditions: *(a)* $\bigcup_\theta F(\theta) = \Delta\Omega$ (the statistic is well defined for all admissible distributions); *(b)* for all $\theta_1 \neq \theta_2$, $F(\theta_1) \nsubseteq F(\theta_2)$ (no statistic value is fully redundant). These assumptions are without loss of generality: all statistics can be redefined on the entire set $\Delta\Omega$ by assigning a dummy value to the distributions for which it is not originally properly defined. Moreover, as the scoring rules I seek to construct offer the same expected score for all correct forecasts, removing statistic values that are fully redundant does not impact the analysis.

Scoring rules. The payoffs to the expert are specified by scoring rules, whose original definition is expanded to account for general statistical forecasts. Given a statistic with value set Θ, a *scoring rule* is a function $S : \Theta \times \Omega \mapsto \mathbb{R}$ that assigns to each forecast θ and each state ω a real-valued score $S(\theta, \omega)$. Payoffs given by scoring rules may be interpreted and used in various ways, as long as the expert complies with the general principle that higher expected payoffs are systematically preferred. To make the discussion concrete, in this paper a scoring rule specifies the payment the expert receives in exchange for his forecast report.

The mechanisms considered throughout this paper are entirely specified by the statistic (F, Θ) we want to learn and the scoring rule S that assigns payoff values. They operate in three stages.

t=1 Nature selects a distribution $p \in \Delta\Omega$.
The expert learns p.

t=2 The expert reports a forecast $\theta \in \Theta$.

t=3 Nature draws a state ω^* at random according to p.
The expert receives an amount $S(\theta, \omega^*)$.

Properness. If the expert is risk-neutral, his optimal response maximizes the expected payment. To induce the expert to answer honestly, we must construct scoring rules that are proper or strictly proper, in Savage's terminology. Proper scoring rules ensure that all true forecasts get the maximum expected payment. With strictly proper scoring rules, the maximum is attained if and only if the forecast is correct. I adapt the original definition to the case of statistical forecasts:

Definition 1. A scoring rule S for statistic (Θ, F) is *proper* if, for all forecasts θ and all distributions p, whenever θ is true under p, i.e., whenever $p \in F(\theta)$,

$$\theta \in \arg\max_{\hat{\theta} \in \Theta} \mathop{\mathsf{E}}_{\omega \sim p} [S(\hat{\theta}, \omega)] \ .$$

S is *strictly proper* if it is proper and if, for all forecasts θ and all distributions p, whenever θ is false under p, i.e., whenever $p \notin F(\theta)$,

$$\theta \notin \arg\max_{\hat{\theta} \in \Theta} \mathop{\mathsf{E}}_{\omega \sim p} [S(\hat{\theta}, \omega)] \ .$$

Because the mechanism controls precisely how much utility the expert gets for each draw of Nature, risk-neutrality is not a binding assumption. Given any expert with strictly increasing utility for money u, any (strictly) proper scoring rule S can be transformed into another one, $u^{-1} \circ S$, that (strictly) induces honest reports from such experts, and conversely. Variations in risk-attitude do not affect our ability to elicit a given statistical information. It merely implies a transformation of the compensation structure. Even if we were to ignore the expert's utilities, enforcing strict incentives remains possible. For example, we can use a two-stage mechanism that starts by estimating the agent's preferences, as in Offerman et al. (2009), or reward experts with lottery tickets to neutralize risk-aversion, as in Karni (2009).

In the same fashion, experts need not be endowed with complete knowledge—but it is convenient to think they are. As long as the expert knows the true value of the statistic, (strictly) proper scoring rules guarantee a (strictly) maximal expected payment, regardless of Nature's distribution. Had the expert the chance to learn the full distribution, he would find himself unable to take advantage of that extra piece of information.

Order sensitivity. Properness looks at how the expected scores of correct forecasts compare with expected scores of incorrect forecasts. However, properness does not look at how the expected scores of incorrect forecasts compare with one another. Yet, a number of situations arise in which such comparisons are desirable. A prominent example is that of eliciting the mean of a random variable. Say, the true mean is 100. Proper scoring rules dictate that forecasting 100 will maximize expected payments. But what about forecasting 99 versus 10? Properness *a priori* does not preclude that the latter forecast yields a higher expected payment than the former.

More generally, suppose the statistic takes value in a set that is naturally ordered. Given a true forecast θ and two incorrect forecasts θ_1, θ_2, it seems pretty uncontroversial that whenever θ_1 is "in-between" θ and θ_2, it constitutes a "more accurate" forecast than θ_2 does. In such situations, it makes sense to offer a contract that, on average, rewards forecast θ_1 with a higher amount than it does for θ_2. This requirement is captured via the notion of order sensitivity.

Definition 2. A scoring rule S for a statistic (Θ, F) is *order sensitive* with respect to total order $\prec$ 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_1, θ_2 such that either $\theta \preceq \theta_1 \prec \theta_2$ or $\theta_2 \prec \theta_1 \preceq \theta$,

$$\underset{\omega\sim p}{\mathsf{E}}[S(\theta_1,\omega)] \geq \underset{\omega\sim p}{\mathsf{E}}[S(\theta_2,\omega)] \ .$$

S is *strictly order sensitive* when the inequality is strict whenever $p \notin F(\theta_2)$.

For convenience, I use the short notation $S(\theta, p)$ to denote the expected score under p, $\mathsf{E}_{\omega\sim p}[S(\theta,\omega)]$.

3 Continuous Statistics

In this section, the state space Ω is a compact metric space, and $\mathcal{G}$ its Borel σ-algebra. Let μ be an arbitrary reference measure on $(\Omega,\mathcal{G})$. The set of distributions of Nature, $\Delta\Omega$, is the set of all probability distributions over $(\Omega,\mathcal{G})$ that are absolutely continuous with respect to μ, and whose density functions are continuous and strictly positive over Ω. (Positiveness is not required in the technical arguments but makes the results applicable to a broader set of statistics.) Denote by $\mathcal{C}(\Omega)$ the set of continuous functions over Ω endowed with the L^1 norm $\|f\| = \int |f| \mathrm{d}\mu$. For notational convenience, probability distributions are identified with their density functions. $\Delta\Omega$ is viewed as a subset of $\mathcal{C}(\Omega)$ that inherits the same topology.

For the remaining of this section I restrict the study to statistics (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) statistic function is continuous and nowhere locally constant[4].

I call such statistics *regular real-valued continuous statistics*. As long as we are interested in statistics that vary along a single dimension, these assumptions are not very restrictive. Common statistics such as the mean and moments, median and quantiles, variance, entropy, skewness, kurtosis (when $\Omega = [a,b]$) and covariance (when $\Omega = [a,b]\times[c,d]$) all satisfy the three conditions. To guarantee existence of expectations, I only consider the scoring rules S such that the function $S(\theta,\cdot)$ is μ-integrable, for all θ.

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

[4] The statistic function is not constant on any open set of distributions.

Proposition 1. *Consider a statistic that satisfies conditions* (Real-Valued) *and* (No Redundancy)*, and whose statistic function is continuous in some topology on* $\Delta\Omega$.[5] *A scoring rule is (strictly) proper if and only if it is (strictly) order sensitive.*

Strictly proper scoring rules do not exist for all continuous statistics. A necessary condition is that the level sets of the statistics be convex; the distributions that share the same statistic value must form a convex shape. Consider two distributions over states, p and q. The argument relies on the simple observation that the expected payment to the expert when forecasting θ under any mixture of p and q,

$$\mathop{\mathsf{E}}_{\omega\sim\lambda p+(1-\lambda)q}[S(\theta,\omega)] \ ,$$

equals the mixture of the expected payments when forecasting θ separately on p and q,

$$\lambda \mathop{\mathsf{E}}_{\omega\sim p}[S(\theta,\omega)] + (1-\lambda) \mathop{\mathsf{E}}_{\omega\sim p}[S(\theta,\omega)] \ .$$

Suppose S is a strictly proper scoring rule and θ is a forecast that is correct for both p and q. By reporting θ, the expert maximizes the expected payment both under both distributions. Per the above observation, the payment remains optimal under any mixture of p and q. Since S is strictly proper, it must be the case that θ is a correct forecast for all mixtures of p and q. Hence all level sets must have a convex shape. As it turns out, this necessary condition is also sufficient for the statistics examined in this section.

Theorem 1. *Let* (Θ, F) *be a regular real-valued continuous statistic. Each of the following statements imply the other two:*

1. *There exists a strictly proper scoring rule for* (Θ, F)*.*
2. *For all* $\theta \in \Theta$*,* $F(\theta)$ *is convex.*
3. *For all* θ *in the interior of* Θ*, there exists a continuous linear functional* L_θ *on* $\mathcal{C}(\Omega)$ *such that, for all distributions* $p \in \Delta\Omega$*,* $L_\theta(p) = 0$ *if and only if* $p \in F(\theta)$*.*

For instance, the mean, moments, median, quantiles of a random variable pass the convexity test (part (2) of the theorem)—and so can be elicited via a strictly proper

[5] By topology on $\Delta\Omega$ I mean topology inherited from the linear topological space $\mathcal{C}(\Omega)$, *i.e.*, one that makes addition and scalar multiplications continuous operations.

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 incentivize experts to report values for these statistics.

The intuition of the proof is as follows. Using standard notation, $\mathcal{L}$ denotes the Lebesgue measure, $L^1(\mu)$ the space of μ-integrable real functions on $(\Omega, \mathcal{G})$, and $L^\infty(\mu)$ the space of μ-measurable essentially bounded real functions on $(\Omega, \mathcal{G})$. The L^∞ norm is defined by $\|f\|_\infty = \sup\{m \mid \mu(f > m) = 0\}$. For members of $L^1(\mu)$ and $L^\infty(\mu)$ I make the usual μ-almost everywhere identification.

We have already seen how part (1) of the theorem implies part (2).

Now assume part (2) is true. I show it implies part (3). By continuity, Θ is an interval. Let θ be any statistic 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 statistic value is less than θ, equal to θ, and greater than θ. If we require that every level set of the statistic be convex, a continuity argument can be applied to prove that all three sets inherits this convexity. The convexity of these three sets, along with compactness of the state space, enables the application of the Hahn-Banach separating hyperplane theorem under the L^∞-norm topology. Therefore there exists a hyperplane $\mathcal{H}_\theta$ that separates $\mathcal{D}^{<\theta}$ from $\mathcal{D}^{>\theta}$ and is closed in the L^∞ norm. As it turns out, the closeness carries over in the L^1-norm topology. The statistic being nowhere locally constant, the hyperplane ends up being the linear span of the set $\mathcal{D}^{=\theta}$, from which part (3) of the theorem follows. Hence part (2) implies part (3).

Finally, assume part (3) is true. I show it implies part (1). Assume without loss of generality that $L_\theta(p)$ is strictly positive when p's statistic value is greater than θ, and strictly negative when p's statistic value is less than θ. Making use of the fact that the (topologic) dual of $L^1(\mu)$ is topologically isomorphic to $L^\infty(\mu)$, a version of the Riesz Representation theorem (for example, Theorem 1.11 of Megginson, 1998) gives a function $g_\theta \in L^\infty(\mu)$ such that, for all $f \in L^1(\mu)$,

$$L_\theta(f) = \int_\Omega f g_\theta \, \mathrm{d}\mu \ ,$$

where L_θ has been extended by continuity to the whole space $L^1(\mu)$, by the Banach extension theorem. It can be shown that the continuity of the statistic implies that $\theta \mapsto g_\theta$ is uniformly continuous on every segment of Θ, with respect to the L^1 norm. This allows existence of a function $H : \Theta \times \Omega \mapsto \mathbb{R}$ that is $(\mathcal{L} \times \mu)$-measurable and is a good approximation of the g_θ's, in the sense that, for almost every θ, $\|H(\theta, \cdot) - g_\theta\| = 0$. Then, defining

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

for an arbitrary θ_0 and in virtue of Fubini's theorem, we get, for all distributions p,

$$\begin{aligned} \mathop{\mathsf{E}}_{\omega \sim p}[S(\theta, \omega)] &= \int_\Omega \left(\int_{\theta_0}^{\theta} H(t, \omega) \mathrm{d}t \right) p(\omega) \mathrm{d}\mu(\omega) \ , \\ &= \int_{\theta_0}^{\theta} \left(\int_\Omega H(t, \omega) p(\omega) \mathrm{d}\mu(\omega) \right) \mathrm{d}t \ . \end{aligned}$$

Let θ^* be the unique value of the statistic under p. The condition imposed on H insures that, for almost all $t \in (\theta, \theta^*)$, $\int_\Omega H(t, \cdot)\mathrm{d}\mu = L_t(p) > 0$ (with a symmetric inequality for $t \in (\theta^*, \theta)$), thereby making S strictly proper. Hence part (3) implies part (1).

Once it is established that the statistic of interest passes the convexity test of Theorem 1, it remains to design the incentive contracts. The family of those contracts turns out to be well structured. In the characterization below, I impose a smoothness condition on the scoring rules. Since the statistic varies continuously with the underlying state distribution, it is reasonable to require that payments vary smoothly with the expert's forecast. To formalize the idea, I say of a scoring rule S that is it *regular* when the scoring rule is uniformly Lipschitz continuous in its first variable: there exists $K > 0$ such that for all $\theta_1, \theta_2 \in \Theta$ and all $\omega \in \Omega$,

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

Looking at the regular scoring rules as the only acceptable scoring rules does not limit the range of statistics to which the characterization applies. Regular strictly proper scoring rules are guaranteed to exist whenever the criteria of Theorem 1 are satisfied. On the other hand this restriction is useful since it permits a simple description of the desirable incentive contracts.

Assume the statistic passes the convexity test of Theorem 1. My 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 scores) 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 2. *Let (Θ, F) be a regular real-valued continuous statistic such that, for each θ, the set $F(\theta)$ is convex. There exists a $(\mathcal{L} \otimes \mu)$-measurable bounded proper scoring rule S_0 such that a regular scoring rule for the statistic is proper (resp. strictly proper) if, and only if, for all θ and μ-almost every ω,*

$$S(\theta, \omega) = \kappa(\omega) + \int_{\theta_0}^{\theta} \xi(t) S_0(t, \omega) \mathrm{d}t \ , \tag{1}$$

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

The idea of the proof works as follows. Most of the work in the proof of Theorem 1 is done for the purpose of building a function $H : \Theta \times \Omega \mapsto \mathbb{R}$ that is such that, for almost every θ and all $p \in \Delta\Omega$, the quantity

$$\int_{\Omega} H(\theta, \omega) \mathrm{d}p(\omega)$$

is respectively strictly positive when p's statistic value is greater than θ, strictly negative when it is less than θ, and zero when it equals θ. Now substitute S_0 for H in (1). By Fubini's Theorem, for all $p \in \Delta\Omega$, and all statistic values θ, θ^* where θ^* is a correct statistic value under p,

$$S(\theta^*, p) - S(\theta, p) = \int_{\theta}^{\theta^*} \xi(t) \left(\int_{\Omega} H(t, \omega) \mathrm{d}p(\omega) \right) \mathrm{d}t \ , \tag{2}$$

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

I now show how to 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 $G : \Theta \times \Omega \mapsto \mathbb{R}$, which can be chosen to be

$(\mathcal{L} \times \mu)$-measurable, 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 continuous functional Ψ_θ on $L^1(\mu)$ by

$$\Psi_\theta(f) = \int_\Omega G(\theta,\omega) f(\omega) \mathrm{d}\mu(\omega) \ .$$

The first step consists in showing that, 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 payment is maximized, its derivative, when it exists, must be zero. Then, defining the continuous functional

$$\Phi_\theta(f) = \int_\Omega H(\theta,\omega) f(\omega) \mathrm{d}\mu(\omega) \ ,$$

an equality $\Psi_\theta = \xi(\theta)\Phi_\theta$ must hold for some $\xi(\theta)$. This is a mere consequence of the fact that the kernel of Φ_θ is the linear span of the distributions having statistic 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 1). If, in addition, S is strictly proper, then the integral of ξ must be strictly positive on every segment, a direct consequence of (2).

The key to list all the (strictly) proper scoring rules is thus to find a possible candidate for S_0. Such candidate can typically be obtained in two ways. When we have a particular strictly proper scoring rule handy, we can differentiate it. Or we can apply part (2) of theorem 1 and write the Riesz representation of L_θ. I illustrate both ways.

Let $\Omega = [a, b]$ be a segment of the real line, and μ be the Lebesgue measure. The distribution mean, whose statistic function is written

$$\Gamma(p) = \int_a^b \omega p(\omega) \mathrm{d}\omega \ ,$$

is (trivially) continuous, and satisfies the convexity condition of Theorem 1, therefore admits a strictly proper scoring rule. In fact, we can easily find one: observing that the mean squared error $\mathsf{E}_{\omega\sim p}[(\theta-\omega)^2]$ 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, $S_0(\theta,\omega)=\omega-\theta$. Altogether, Theorem 2 gives all the regular (strictly) proper scoring rules for the mean,

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

The median, whose statistic function Γ is defined implicitly by

$$\int_a^{\Gamma(p)} p(\omega)\mathrm{d}\omega=\frac{1}{2}\ ,$$

is also continuous. This definition provides a continuous linear functional on $\mathcal{C}([a,b])$,

$$L_\theta(f)=\int_a^b\left[\mathbb{1}\{\theta\le\omega\}-\frac{1}{2}\right]f(\omega)\mathrm{d}\omega\ ,$$

leading to a candidate for the base scoring rule S_0,

$$S_0(\theta,\omega)=\mathbb{1}\{\theta\le\omega\}-\frac{1}{2}=\frac{1}{2}\left[\mathbb{1}\{\theta\le\omega\}-\mathbb{1}\{\omega\le\theta\}\right]\ .$$

Applying theorem 2 and letting $g(\theta)=\frac{1}{2}\int_a^\theta\xi(t)\mathrm{d}t$, I derive the regular (strictly) proper scoring rules for the median,

$$S(\theta,\omega)=\kappa(\omega)-|g(\theta)-g(\omega)|\ .$$

By varying ξ under the constraints of Theorem 2, it is easily seen that g can be any weakly increasing (strictly increasing) Lipschitz function on $[a,b]$. By the same logic, the (strictly) proper scoring rules that elicit an α-quantile take the form

$$S(\theta,\omega)=\kappa(\omega)+(2\alpha-1)(g(\theta)-g(\omega))-|g(\theta)-g(\omega)|\ .$$

It should be noted that continuity of the state space is not required. Dealing with finite state spaces is somewhat easier, because when equipped with the discrete metric, each of them is a compact metric space that makes continuous any function.

A prevalent example is the dichotomous state space $\Omega = \{0, 1\}$. The probability of occurrence of $\omega = 1$ is, obviously, a continuous statistic. Following the above example of the mean and according to Theorem 2, the form of its (strictly) proper scoring rules is given by

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

which is precisely the Schervish representation of probability scoring rules (Schervish, 1989). However, discrete state spaces must be used with care: statistics can be continuous with continuous states and lose continuity with discrete states. This true, in particular, of the median and quantiles.

Behind the integral representation lies a simple interpretation. From the perspective of the risk-neutral expert, being remunerated according to a (strictly) proper scoring rule is essentially the same as participating to an auction that sells off some carefully designed securities. To see why, assume statistic 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

$$P(y \leq x)\, \mathsf{E}[R_y(\omega^*) \mid y \leq x] = \int_{y \leq x} R_y(\omega^*) f(y)\mathrm{d}y \ . \tag{3}$$

Take any strictly proper scoring rule S of the form (1). Choosing density $f(y) = \xi(y)/\int_\Theta \xi$ and security $R_y(\omega) = \left(\int_\Theta \xi\right) S_0(y, \omega)$, (3) 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 payment. 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 statistic 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. These 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. To elicit the median of a random variable, we should auction off securities whose net payoff $R_\theta(x) = 1$ when $x \geq \theta$ and $R_\theta(x) = -1$ if $x < \theta$ (or a positive factor of these) where θ is the parameter. Here the effect is opposite. The price of the security is fixed, say, 1, and buyers bid on a threshold parameter that only affect the gross payoff, which is then either 0 or 2, depending on whether the underlying variable realizes below or above the threshold.

As a final remark, the results of this section heavily rely on the continuity of the statistics, which can be severely constraining. While it may seem promising to look for a weaker notion of continuity so as to include a broader range of statistics, a simple argument shows that the choice of L^1-norm topology is not arbitrary: it turns out to be the strongest topology—and so the weakest continuity notion—which can be used. Indeed, whenever the statistic takes values in a bounded set, any statistic that is continuous in some topology, but not in the L^1 norm, never admits a strictly proper scoring rule.[6]

[6] To see why, consider a bounded statistic function Γ that is continuous in some topology. Take a sequence $(p_n)_n$ that converges towards some density p_∞ in the L^1 norm. If S is strictly proper, then $\mathsf{E}_{\omega \sim p_n}[S(\Gamma(p_\infty), \omega)] \leq \mathsf{E}_{\omega \sim p_n}[S(\Gamma(p_n), \omega)]$. But simultaneously, as n grows to infinity, $\mathsf{E}_{\omega \sim p_n}[S(\Gamma(p_n), \omega)]$ gets arbitrarily close to $\mathsf{E}_{\omega \sim p_\infty}[S(\Gamma(p_n), \omega)]$ when S is regular. Since S is also strictly order sensitive (Proposition 1 holds regardless of the topology that makes the statistic continuous), the only cluster point of $(\Gamma(p_n))_n$ is $\Gamma(p_\infty)$ and so $\Gamma(p_n) \to \Gamma(p_\infty)$.

4 Discrete Statistics

The preceding section deals with statistics that are continuous, and take values in a continuum. This section deals with statistics that are discrete, and take values in a finite set. I call such statistics *discrete statistics.*

Allowing for discontinuities comes at the expense of a constraint on the state space Ω: throughout this section Ω is a finite set. The set of distributions of Nature $\Delta\Omega$ is the unrestricted set of all distributions of Ω. Distributions remain identified with their densities and the terms are used interchangeably.

The principal benefit of dealing with finite state spaces lies in that the space of random variables, $\mathbb{R}^\Omega$, can be expressed as a Hilbert space. While the results of the preceding section rely on the properties of Banach spaces, the results of this section rely on the properties of Hilbert spaces. I endow $\mathbb{R}^\Omega$ with the (Euclidian) inner product

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

This enables to write an expected payment of the form

$$\mathop{\mathsf{E}}_{\omega \sim p} [S(\theta, \omega)] \ ,$$

as the inner product between random variable $S(\theta, \cdot)$ and density function p,

$$\langle S(\theta, \cdot), p \rangle \ .$$

Writing expected payments as inner products plays a key role in the technical arguments of this section.

I begin with a description of the statistics that accept strictly proper scoring rules. Convexity of the level sets is a necessary condition, by the same argument as for continuous statistics. However it is no longer sufficient, consequence of the lack of continuity.

The exact characterization makes use of a well-known geometric structure called *Voronoi diagram.* Voronoi diagrams specify, for a set of points or sites, the regions of the space that comprise the points closest to each site (Aurenhammer, 1991; De Berg et al., 2008). Specifically, consider a metric space $\mathcal{E}$ with distance d, together with vectors $x_1, \ldots, x_n \in \mathcal{E}$, referred to as *sites.* The *Voronoi cell* for site x_i gathers 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 $x_1, \ldots, x_n$.

To understand the role of Voronoi diagrams, it is helpful to start off with a sufficient condition. The set of distributions, identified with that of density functions, is a subset of the space of random variables. It inherits its Euclidian metric. In this context it makes sense to talk about *Voronoi diagrams of distributions.* This gives a sufficient condition: *if the level sets of a discrete statistic form a Voronoi diagram of distributions, then there exists a strictly proper scoring rule for the statistic.*

The argument is as follows. Let (Θ, F) be a discrete statistic. 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 to that of Nature (in the Euclidian distance), the larger the expected payments (Friedman, 1983). In consequence if we were to force the expert to choose his report among the distributions q_θ, $\theta \in \Theta$, his best response would be to produce the q_θ the closest to Nature's distribution. By forcing the expert to report one of the q_θ's, we effectively ask him to report the Voronoi cell that contains the distribution of Nature. To the extent that the Voronoi cells are the level sets of the underlying statistic, this essentially means reporting a value θ for the statistic and getting rewarded according to the scoring rule $S(\theta, \omega) = 2q_\theta(\omega) - \|q_\theta\|^2$.

The condition is not necessary, because the logic can be replicated to other probability scoring rules and other distances. But this 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 space of random variables. Specifically, *the statistics that admit strictly proper scoring rules are precisely those whose level sets form pieces of a larger Voronoi diagram of random variables.*

The argument is as follows. Let (Θ, F) be a discrete statistic and let $\{X_\theta\}_{\theta\in\Theta}$ be a family of random variables. Consider the Voronoi diagram of this family, and let $\mathcal{C}_\theta$ be the Voronoi cell for X_θ. Define the scoring rule $S(\theta, \omega) = 2X_\theta(\omega) - \|X_\theta\|^2$. The expected payment of forecast θ, under distribution p, is

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

This means that the expected payment of forecast θ under p is maximized across all forecasts if and only if

$$\|p - X_\theta\| \leq \|p - X_{\hat{\theta}}\| \qquad \forall \hat{\theta} \in \Theta \ ,$$

which is to that p belongs to the Voronoi cell $\mathcal{C}_\theta$ of X_θ. Suppose $F(\theta)$ is the piece of $\mathcal{C}_\theta$ located on the simplex of distributions. The expected payment of forecast 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. The following equivalence holds for all $\theta \in \Theta$ and all $p \in \Delta\Omega$:

$$\mathop{\mathsf{E}}_{\omega \sim p}[S(\theta, \omega)] \geq \mathop{\mathsf{E}}_{\omega \sim p}[S(\hat{\theta}, \omega)] \ \ \forall \hat{\theta} \in \Theta \qquad \text{if and only if} \qquad p \in F(\theta) \ .$$

I shall use the family of random variables $\{X_\theta\}_{\theta \in \Theta}$ as the set of sites, defining $X_\theta(\omega) = S(\theta, \omega) + k_\theta$, 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

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

or equivalently, after expanding the terms,

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

If the choice in k_θ is such that $\|S(\theta, \cdot) + k_\theta\|^2 - 2k_\theta$ equals a constant C independent of θ, then after simplification the last inequality becomes

$$\mathop{\mathsf{E}}_{\omega \sim p}[S(\theta, \omega)] \geq \mathop{\mathsf{E}}_{\omega \sim p}[S(\hat{\theta}, \omega)] \qquad \forall \hat{\theta} \in \Theta \ .$$

As long as C is chosen to be greater that $\|S(\theta, \cdot)\|^2$ uniformly across statistic values, getting constants k_θ that satisfy this requirement is always possible. Hence, moving back up on the chain of inequalities, we find that for any given distribution p, forecast θ maximizes the expected payment—and so p belongs to $F(\theta)$—if and only if p is located in the Voronoi cell of X_θ. In summary:

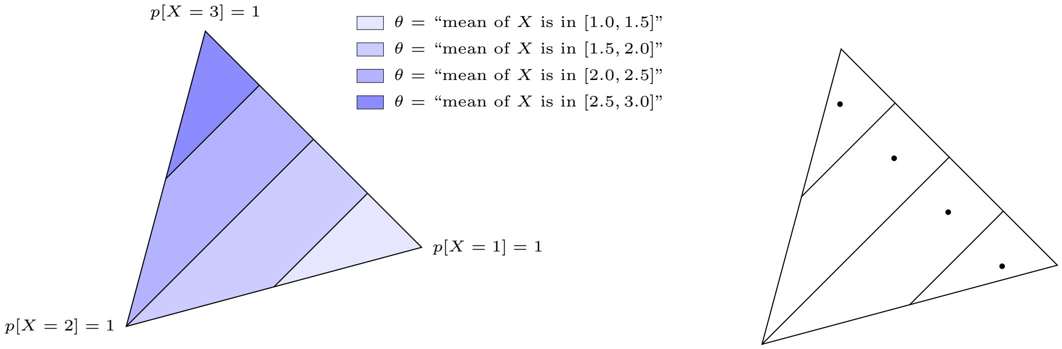
Theorem 3. *A discrete statistic* (Θ, F) *admits a strictly proper scoring rule if and* *only if there exists a Voronoi diagram of random variables, $\{\mathcal{C}_\theta\}_{\theta\in\Theta}$, whose cells are indexed by statistic values, such that for all distributions $p \in \Delta\Omega$ and all statistic values $\theta \in \Theta$,*

$$p \in \mathcal{C}_\theta \qquad \textit{if and only if} \qquad p \in F(\theta) \ .$$

These linear cross-sections of Voronoi diagrams are otherwise known as *power diagrams* (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, the theorem can be reformulated in the following way: *a statistic admits a strictly proper scoring rule if and only if its level sets form a power diagram of distributions.*

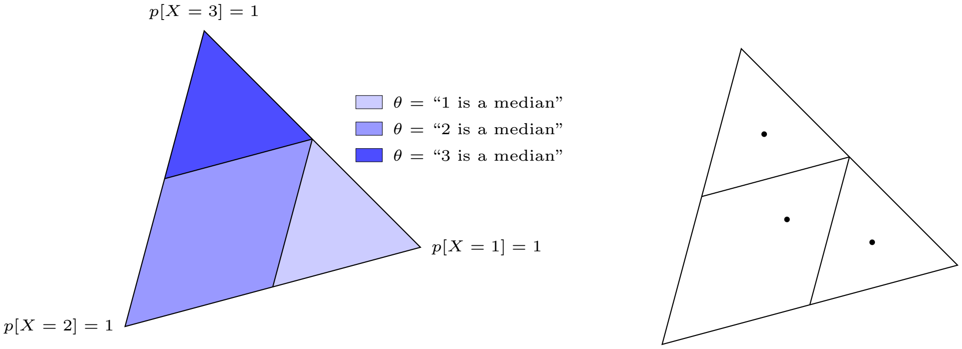
Conducting such Voronoi tests can be more difficult than the simple convexity tests of the continuous statistics. Nevertheless the geometric characterization is appealing, for as long as the dimensionality of the simplex of distributions remains low, a quick visual check gives a good sense of whether the statistic satisfies the condition of Theorem 3. Figures 1 to 4 depict 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 classic exemplars of discrete statistics. 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. We can conclude that all these statistics admit a strictly proper scoring rule. 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 simplices. And in most cases, the 2-dimensional visual test is sufficient to get convinced of its existence.

It is not difficult to exhibit statistics that fail the Voronoi test. Figure 5 maps two typical situations. In (a), the statistic indicates whether a random variable has "high" or "low" variance. The statistic fails the Voronoi test because one of the level set is not convex. In (b), the statistic gives a 90% confidence interval for a random variable. In spite of its allure of symmetric convex partition, the statistic 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. In general, to be able to elicit the forecasts of a discrete statistic, 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


$p[X = 3] = 1$
$p[X = 1] = 1$
$p[X = 2] = 1$
θ = "mean of X is in $[1.0, 1.5]$"
θ = "mean of X is in $[1.5, 2.0]$"
θ = "mean of X is in $[2.0, 2.5]$"
θ = "mean of X is in $[2.5, 3.0]$"

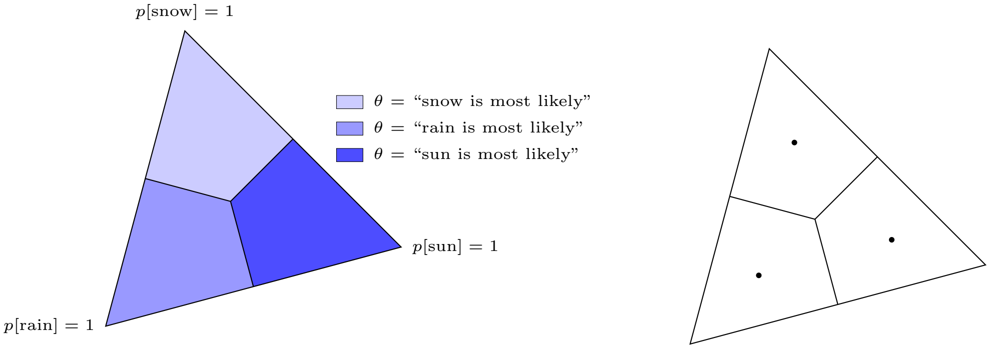
(a) (b)

Figure 1: Level sets for the mean.


$p[X = 3] = 1$
$p[X = 1] = 1$
$p[X = 2] = 1$
θ = "1 is a median"
θ = "2 is a median"
θ = "3 is a median"

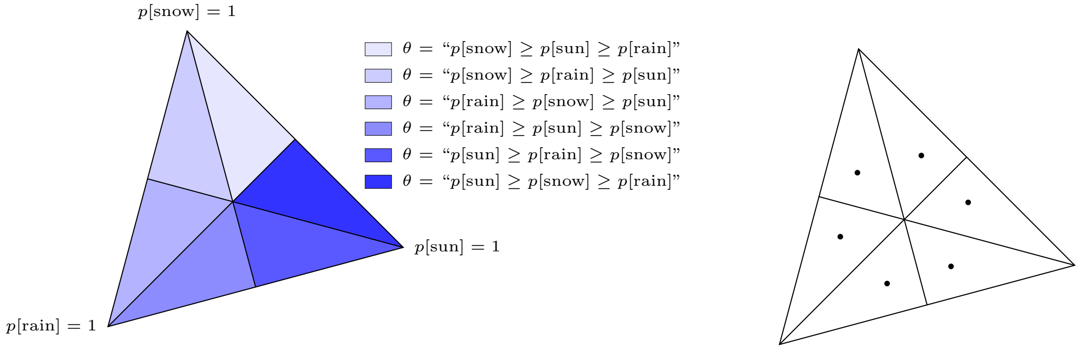
(a) (b)

Figure 2: Level sets for the median.


p[snow] = 1
p[sun] = 1
p[rain] = 1
θ = "snow is most likely"
θ = "rain is most likely"
θ = "sun is most likely"

(a) (b)

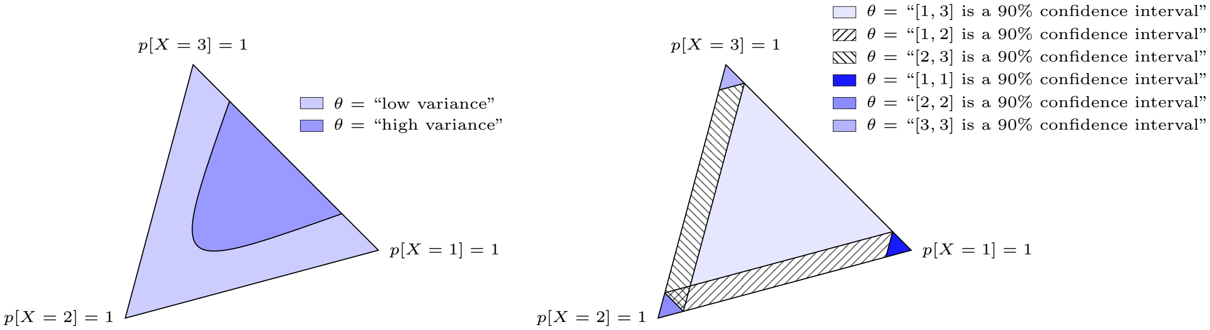
Figure 3: Level sets for the most likely state.



$p[\text{snow}] = 1$
θ = "$p[\text{snow}] \geq p[\text{sun}] \geq p[\text{rain}]$"
θ = "$p[\text{snow}] \geq p[\text{rain}] \geq p[\text{sun}]$"
θ = "$p[\text{rain}] \geq p[\text{snow}] \geq p[\text{sun}]$"
θ = "$p[\text{rain}] \geq p[\text{sun}] \geq p[\text{snow}]$"
θ = "$p[\text{sun}] \geq p[\text{rain}] \geq p[\text{snow}]$"
θ = "$p[\text{sun}] \geq p[\text{snow}] \geq p[\text{rain}]$"
$p[\text{sun}] = 1$
$p[\text{rain}] = 1$

(a) (b)

Figure 4: Level sets for the ranking of state probabilities.



$p[X = 3] = 1$
θ = "low variance"
θ = "high variance"
$p[X = 1] = 1$
$p[X = 2] = 1$
θ = "[1, 3] is a 90% confidence interval"
θ = "[1, 2] is a 90% confidence interval"
θ = "[2, 3] is a 90% confidence interval"
θ = "[1, 1] is a 90% confidence interval"
θ = "[2, 2] is a 90% confidence interval"
θ = "[3, 3] is a 90% confidence interval"
$p[X = 3] = 1$
$p[X = 1] = 1$
$p[X = 2] = 1$

(a) (b)

Figure 5: Level sets for the variance and confidence intervals.

should partition the space of distributions except for a measure zero set. Returning to confidence intervals, it is possible to elicit such forecasts when forcing the interval to be symmetric, which in effect takes away the "thick" overlapping regions.

Although Voronoi diagrams look much like convex partitions, a Voronoi test is a lot stronger than a convexity test, especially in high dimensions. Nonetheless most statistics that exhibit a high level of symmetry are naturally shaped as Voronoi diagrams. Non-symmetric cases can arise as well. For example, the statistic 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 statistic that passes the Voronoi test of Theorem 3. How should we construct strictly proper scoring rules? As for continuous statistics, the payment schemes must obey simple rules that, to some extent, generalize the form given in Theorem 2. The (strictly) proper scoring rules are essentially the mixtures of a finite number of carefully chosen proper scoring rules, that form a base. The scoring rules of the base are entirely determined by the level sets of the statistic being elicited.

Theorem 4. *Assume there exists a strictly proper scoring rule for discrete statistic* (Θ, F)*. 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 $\kappa : \Omega \mapsto \mathbb{R}$*, and nonnegative (resp. strictly positive) reals* λ_i*,* $i = 1, \ldots, \ell$*.*

The proof is based on the following idea. Let S be a proper scoring rule. The properness property is captured by the following inequalities:

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

There are uncountably many inequalities. However, whenever the statistic satisfies the criterion of Theorem 3 the sets $F(\theta)$ are convex 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, responsible for the translations, produces the complementary state-contingent payments. Adding strict properness substitutes some weak inequalities for strict ones in the above system, which significantly 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.

For instance:

- For the statistic that gives the most likely of n arbitrary events $E_1, \ldots, E_n \subset \Omega$, the (strictly) proper scoring rules are

$$S(E_k, \omega) = \kappa(\omega) + \lambda\{\omega \in E_k\} \ .$$

- For the median of a random variable X, the (strictly) proper scoring rules take the form

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

  in which $x_1, \ldots, x_n$ are the values taken by X.

- Divide the range $[0, 1]$ into n intervals of equal size, $\left[\frac{k-1}{n}, \frac{k}{n}\right]$, $k = 1, \ldots, n$. For the statistic that gives an interval that contains the probability for some given event $E \subset \Omega$, the (strictly) proper scoring rules are

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

Some of these examples are detailed below.

The remaining of this section discusses 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. Owing to the lack of continuity, the argument of Proposition 1 cannot be replicated. In fact, when looking at discrete statistics, it immediately appears that many admit a strictly proper scoring rules, but generally much fewer admit strictly order-sensitive ones. (To be convinced of such a fact, it suffices to consider multiple orderings of the same values.)

The result below is a test for the existence of strictly order-sensitive scoring rules. As expected, the test is much stronger than the Voronoi test of Theorem 3. It is also much simpler to carry out. A statistic passes the test if and only if its level sets form "slices" of the space of distributions, separated by hyperplanes.

Theorem 5. *Let* $(\Theta = \{\theta_1, \ldots, \theta_n\}, F)$ *be a discrete statistic, 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]

The proof idea is best conveyed through an example. Consider a strictly order-sensitive scoring rule for a statistic whose value set Θ contains three elements, θ_1, θ_2 and θ_3. If both θ_1 and θ_3 are correct forecasts under some distribution p, but θ_2 is not, then the expected payment, under p, is maximized only when responding θ_1 or θ_3. Adding a small perturbation to p, we can pull out a distribution $\tilde{p}$ for which the only true forecast is θ_1, while announcing θ_3 yields an expected fee 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 statistic 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 6.

I illustrate Theorem 5 with two common cases. First, consider the median of a discrete random variable X. Figure 2 suggests that the statistic passes the slice test of Theorem 5.[8] To contrast, consider the mode of X. This statistic gives the most likely value of X. Figure 3, for which the mode is a special case, clearly indicates that the statistic fails the test of Theorem 5.[9]

Strictly order-sensitive scoring rules are also strictly proper, and the form of the strictly proper contracts follows the rule given in Theorem 4. The technique of Theorem 4 is not constructive. It does not procure the scoring rules that compose the base, but simply asserts their existence. One benefit of statistics that admit strictly

[7] Hyperplanes of distributions can be viewed as hyperplanes in the Hilbert space of random variables that intersect the simplex of densities.

[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 5 is satisfied.

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

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.

Theorem 6. *Let $(\Theta = \{\theta_1, \ldots, \theta_n\}, F)$ be a discrete statistic 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 generated by $F(\theta_i) \cap F(\theta_{i+1})$.

The proof is immediate and relegated to the Online Appendix.

Now I elaborate a proposition in the spirit of Proposition 1, with a slight tweak. The key is to assume existence of strict order sensitivity. As long as existence is granted, strictly proper scoring rules are exactly the strictly order-sensitive scoring rules. This implies that when a statistic accepts a strictly order-sensitive scoring 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 5, the result breaks down without the existence requirement. It breaks down even when restricted to weak order sensitivity, which, obviously, exists for all statistics.

Proposition 2. *Let $(\Theta = \{\theta_1, \ldots, \theta_n\}, F)$ be a discrete statistic 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 (strictly) proper if and only if it is (strictly) order sensitive, with respect to $\prec$.*

The proof is elementary. Take, for example, the median statistic of random variable X. Let x_i be the i-th smallest value taken by X. The hyperplane that separates two consecutive level sets of the median, for respective values x_k and x_{k+1}, is, as established previously, specified by equation $\sum_{i \leq k} p[X = x_i] = \frac{1}{2}$. And so, the following

$$\mathbf{n}_k(\omega) = \begin{cases} -1 & \text{if } X(\omega) \leq x_k \ , \\ +1 & \text{if } X(\omega) > x_k \ , \end{cases}$$

defines a positively oriented normal for each k. The normals generate the $n-1$ base scoring rules

$$S_k(m,\omega) = \begin{cases} -1 & \text{if } m < x_i, X(\omega) \leq x_i \ , \\ 0 & \text{if } m \geq x_i \ , \\ +1 & \text{if } m < x_i, X(\omega) > x_i \ . \end{cases}$$

Finite-range statistics can also be employed as approximations of continuous statistics. When used in this fashion, one can combine the results of the present and preceding sections to derive the (strictly) proper and (strictly) order-sensitive scoring rules. Let (Θ, F) be a regular real-valued continuous statistic. Let $\alpha_0, \alpha_n \in \mathbb{R} \cup \{+\infty, -\infty\}$ be respective lower and upper bounds of the possible values for the statistic, and $\alpha_1 < \cdots < \alpha_{n-1}$ be arbitrarily chosen in the interior of Θ (which is an interval). Consider the discrete, approximate version $(\tilde{\Theta}, \tilde{F})$ of the continuous statistic. Instead of the exact value, it rounds up nearby statistical estimates: it gives an interval of the form $[\alpha_i, \alpha_{i+1}]$ that includes the exact value of the continuous statistic (Θ, F). $\tilde{F}([\alpha_i, \alpha_{i+1}])$ is therefore the set of distributions whose values, for the statistic (Θ, F), lie within $[\alpha_i, \alpha_{i+1}]$.

The collection of intervals, $\tilde{\Theta}$, 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 statistic. Theorem 1 says that each $F(\theta)$ is a hyperplane of $\Delta\Omega$. As $\tilde{F}([\alpha_{i-1}, \alpha_i]) \cap \tilde{F}([\alpha_i, \alpha_{i+1}]) = F(\alpha_i)$, a direct application of theorems 5 and 6 gives the following corollary:

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

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

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

For example, consider again the statistic that gives, for an even E, some interval $\left[\frac{k-1}{n}, \frac{k}{n}\right]$ that includes the probability of E. Event E's probability is a regular continuous statistic. 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

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

Rearranging the terms, the strictly proper scoring rules are given by

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

with g any strictly increasing function.

5 Applications

This section illustrates applications of the general model to more specific economic contexts. For expositional convenience, in this section state spaces are finite and statistics are discrete.

5.1 Screening Contracts

In the previous sections the expert has complete knowledge of Nature's distribution. This section illustrates an extreme case of experts with differentiated knowledge. The expert can be one of two types: either completely informed (type I), or completely uninformed (type U). The expert knows his own type, but we, elicitors, don't, thereby creating a problem of adverse selection. We want to hire the informed type while discouraging the uninformed type to participate.

Olszewski and Sandroni (2007, 2010) discuss in details the problem of screening experts. They consider forecasts of the entire distribution over states. In this section I extend the impossibility results to forecasts of statistics. For ease of exposition I focus on a basic setting.

The mechanism is revised as follows. At $t = 1$ Nature chooses her distribution $p \in \Delta\Omega$. The expert of type I learns p, while the expert of type U does not know p. At $t = 2$, the expert is offered a scoring rule contract S to forecast some given statistic (Θ, F). The expert can either accept of decline the offer. If he declines, the expert gets zero payment and exits the mechanism. If he accepts, he must report some forecast θ. At $t = 3$ Nature draws the true state ω^* according to p, and the accepting expert is paid the amount $S(\theta, \omega^*)$. Contracts are binding and both positive and negative payments can be enforced by a third party.

I work under the assumptions of Olszewski and Sandroni. Both types are risk-neutral. The informed expert, knowing Nature's distribution, is able to compute his expected payment. He accepts the contract if and only if he expects to make a strictly positive amount of money, on average, when providing a correct forecast. The uninformed expert, however, does not know Nature's distribution. He evaluates his prospects based on his expected payment computed at the worst possible distribution of Nature. This captures a case of extreme aversion to uncertainty. The uninformed expert accepts the contract under the only condition that he be able to strategically select his reports so as to earn a strictly positive expected monetary amount, even in the worst-case distribution of Nature. In more realistic settings, the expert would probably exhibit aversion to a lesser degree. The idea behind these assumptions is that, if we are able to screen the informed type under such extreme behaviors, it would still be possible to do so under milder, more realistic ones.

However, as long as the statistic of interest can be elicited via a strictly proper scoring rule, it remains impossible to screen the informed type.

Proposition 3. *Let (Θ, F) be a discrete statistic that can be elicited via a strictly proper scoring rule. If the expert is presented with a scoring rule contract that the informed type always accept, then the uninformed type always accepts the contract.*

Proof. The proof a direct consequence of the minimax argument at the heart of Olszewski and Sandroni (2007, 2010), here replicated for completeness. Given a distribution of Nature p, denote by $g(\xi, p) = \int S(\theta, p) \mathrm{d}\xi(\theta)$ the expected amount earned by the expert who randomizes forecasts according to ξ. By a standard application of the von Neumann Minimax Theorem,

$$\max_{\xi \in \Delta\Theta} \min_{p \in \Delta\Omega} g(\xi, p) = \min_{p \in \Delta\Omega} \max_{\xi \in \Delta\Theta} g(\xi, p) \ . \tag{4}$$

Let $\Gamma(p)$ be one possible true value of the statistic under distribution p, and let ξ_p be a distribution that puts full mass on $\Gamma(p)$. For any given distribution of Nature p, the expected payment of the expert who reports $\Gamma(p)$ is $g(\xi_p, p)$. As the informed type always accept S, $g(\xi_p, p) > 0$. By (4), there exists a distribution over possible forecasts, ξ^*, independent of p, such that $g(\xi^*, p) > 0$ for all p. Hence an expert who randomizes forecasts according to ξ^* is guaranteed a strictly positive expected revenue in for all distributions of Nature, and so in particular for all possible realized states. □

5.2 Costly Information Acquisition

The preceding subsection argues that it is theoretically impossible to screen for the informed experts. Now suppose the expert initially receives little information, but is capable to become better informed at some cost. We want to write a contract that induces the expert to exert himself and provide a forecast to the best of his capability. This essentially converts the problem of adverse selection into a problem of moral hazard. The adverse selection problem is unsolvable. Can we solve the problem of moral hazard?

To account for different levels of informativeness, I complement the finite state space Ω with two finite signal spaces, Σ_1 and Σ_2. The signals of Σ_1 are given to the expert at no cost, while the signals of Σ_2 are costly. Let p be a prior over the product space $\Omega \times \Sigma_1 \times \Sigma_2$ with full support. p is common knowledge.

Suppose we want to elicit forecasts of discrete statistic (Θ, F). Acquiring the costly signal is desirable only when learning this signal yields better forecasts. To capture the idea, I assume the signals of Σ_2 always give finer information than those of Σ_1. Formally, this means that for all signals $\sigma_1 \in \Sigma_1$ and all forecasts θ that are true under the posterior $p(\cdot \mid \sigma_1)$, there exists a signal $\sigma_2 \in \Sigma_2$ such that θ becomes incorrect under the revised posterior $p(\cdot \mid \sigma_1, \sigma_2)$. Denote by $\mathcal{P}$ the set of distributions that satisfy this assumption.

The mechanism operates as follows. At $t = 1$, Nature draws a state ω^* and signals σ_1^*, σ_2^* according to the common prior $p \in \mathcal{P}$. The expert learns σ_1^* at no expense, and must decide whether to buy signal σ_2^* at cost C. At $t = 2$, the expert is asked to produce a forecast θ for the statistic. At $t = 3$, the true state ω^* is revealed and the expert gets paid the amount $S(\theta, \omega^*)$.

We want to design a scoring rule contract S satisfying the following two conditions:

Proper It is always in the expert's best interest to report a correct forecast, *i.e.*, the scoring rule is proper.

Effort Inducing The expert's expected revenue net the purchase of the costly signal is never less than his expected revenue without purchase. Formally, for all functions $\theta_1 : \Sigma_1 \mapsto \Theta$ such that $\theta_1(\sigma_1)$ is a correct forecast under $p(\cdot \mid \sigma_1)$, and all functions $\theta_2 : \Sigma_1 \times \Sigma_2 \mapsto \Theta$ such that $\theta_2(\sigma_1, \sigma_2)$ is correct under $p(\cdot|\sigma_1, \sigma_2)$,

$$\mathsf{E}[S(\theta_2(\sigma_1, \sigma_2), \omega) \mid \sigma_1] - C \geq \mathsf{E}[S(\theta_1, \omega) \mid \sigma_1] \ , \qquad \forall \sigma_1 \in \Sigma_1 \ .$$

As expected, the statistics that can be elicited via costly information acquisition are exactly those that admit a strictly proper scoring rules.

Proposition 4. *The following two properties obtain:*

- *If a strictly proper scoring rule exists for* (Θ, F)*, then, for all distributions* $p \in \mathcal{P}$*, there exist scoring rule contracts that satisfy conditions* (Proper) *and* (Effort Inducing)*: any strictly proper scoring rule, scaled by a large enough factor, fulfills these two conditions.*
- *If no strictly proper scoring rule exists for* (Θ, F)*, then there exists a distribution* $p \in \mathcal{P}$ *for which no scoring rule contract satisfies both conditions* (Proper) *and* (Effort Inducing) *simultaneously.*

Proof (sketch). For the first point to be true, it suffices to show that, for any strictly proper scoring rule S, the following strict inequality,

$$\mathsf{E}[S(\theta_2(\sigma_1, \sigma_2), \omega) \mid \sigma_1] > \mathsf{E}[S(\theta_1(\sigma_1), \omega) \mid \sigma_1] \ ,$$

holds for all $\sigma_1 \in \Sigma_1$ and all functions θ_1, θ_2 of the form given above. Fix any $\sigma_1 \in \Sigma_1$. As S is strictly proper, for all signals $\sigma_2 \in \Sigma_2$,

$$\mathsf{E}[S(\theta_2(\sigma_1, \sigma_2), \omega) \mid \sigma_1, \sigma_2] \geq \mathsf{E}[S(\theta_1(\sigma_1), \omega) \mid \sigma_1, \sigma_2] \ . \tag{5}$$

Per our assumption on the family $\mathcal{P}$, even though $\theta_1(\sigma_1)$ is true under the initial posterior $p(\cdot \mid \sigma_1)$, there is always some σ_2 that makes this forecast incorrect under the revised posterior $p(\cdot \mid \sigma_1, \sigma_2)$. Hence inequality (5) becomes strict for some σ_2.

Consequently,

$$\begin{aligned}\mathsf{E}[S(\theta_2(\sigma_1,\sigma_2),\omega) \mid \sigma_1] &= \mathsf{E}[\mathsf{E}[S(\theta_2(\sigma_1,\sigma_2),\omega) \mid \sigma_1,\sigma_2] \mid \sigma_1] \\ &> \mathsf{E}[\mathsf{E}[S(\theta_1(\sigma_1),\omega) \mid \sigma_1,\sigma_2] \mid \sigma_1] \\ &= \mathsf{E}[S(\theta_1(\sigma_1),\omega) \mid \sigma_1] \ .\end{aligned}$$

The second part of the proposition is intuitive. Take any scoring rule contract S that satisfies the first condition but is not strictly proper. Then there exists some $q \in \Delta\Omega$ and forecasts $\theta, \hat{\theta}$ such that θ is correct under q but $\hat{\theta}$ is not, and $S(\theta, q) = S(\hat{\theta}, q)$. Now pick any distribution $p \in \mathcal{P}$ and, given arbitrary signals σ_1, σ_2, re-weight p in such a way that $\hat{\theta}$ is correct under $p(\cdot \mid \sigma_1)$ while θ is not. Set $p(\cdot \mid \sigma_1, \sigma_2)$ to be equal to q. As long as σ_2 occurs sufficiently often given σ_1, S violates the second condition. □

5.3 Forecast Evaluation

This subsection illustrates the use of scoring rules as an ex-post evaluation tool to compare experts' forecasting abilities. I use a simple argument to show that, aside from creating appropriate incentives, strict properness also enables the separation between the "good experts" and the "bad experts" among a pool that contains both types.

Let Ω be a finite state space, Ω^∞ be the set of all sequences of states over Ω, and $\mathcal{F}$ be the σ-algebra generated by the finite histories. Let $\Delta\Omega^\infty$ be the set of probability distributions over $(\Omega^\infty, \mathcal{F})$ that assign strictly positive probability to all finite histories. (Θ, F) denotes a statistic on period distributions that takes a finite number of values.

Nature initially chooses a distribution $P \in \Delta\Omega^\infty$. Then, at each period $t = 1, 2, \ldots$ Nature draws a publicly observed state ω_t^*, according to P, and conditionally on the history of realized states $\omega_1^*, \ldots, \omega_{t-1}^*$. At each period t, prior to Nature's move, each expert $i = 1, \ldots, N$ produces a forecast θ_t^i of the statistic of interest regarding the distribution over states at period t.

We want to evaluate the forecasting abilities of the experts. To this end, experts are given performance measures to assess the quality of their forecasts. Each expert i starts with measure $\pi_0^i = 0$. Then, at each period t, after the forecasts have been collected and the true state ω_t^* revealed, the performance of each expert i is revised to

$$\pi_t^i = \frac{t-1}{t}\pi_{t-1}^i + \frac{1}{t}S(\theta_t^i, \omega_t^*) \ ,$$

in which S is a strictly proper scoring rule for the statistic. I argue that, defined as such, performances reflect the extent to which an expert provides good forecasts in comparison to another.

Some definitions are needed to state the formal results. Given a sequence of states, say that the performance of expert i is *asymptotically as good* as that another expert j when, for all $\epsilon > 0$, $\pi_t^i \geq \pi_t^j - \epsilon$ for all t chosen large enough. For $\delta > 0$, define $\mathcal{P}_\delta$ be a set of distributions "close" to that of Nature,

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

The definition resembles the notion of merging of probability measures (Blackwell and Dubins, 1962; Kalai and Lehrer, 1994; Lehrer and Smorodinsky, 1996). I say of a sequence of forecasts $(\theta_t)_{t\geq 1}$ that it is *essentially correct* if forecasts are arbitrarily accurate except possibly for a proportion zero of periods: for all $\delta > 0$, there exists a distribution $Q \in \mathcal{P}_\delta$ such that

$$\lim_{T\to+\infty} \frac{1}{T}|\{t \leq T \mid \theta_t \text{ is correct under } Q(\cdot|\omega_1^*, \ldots, \omega_{t-1}^*)\}| = 1 \ .$$

Proposition 5. *If an expert i knows the true distribution of Nature P and supplies forecasts accordingly, then, for P-almost every sequence of states, the following obtains:*

- *the performance measure of expert i is always asymptotically as good as that of any other expert j, and*
- *if the performance measure of some expert j is asymptotically as good as that of expert i, then the sequence of forecasts of expert j is essentially correct.*

Proof. The proof makes use of the following lemmas.

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 forecasts $\theta, \hat{\theta}$, and all distributions over states $p \in \Delta\Omega$ such that θ is correct under p, if $S(\hat{\theta}, p) \geq S(\theta, p) - \epsilon$, then there exists $q \in \Delta\Omega$ with $\|p - q\| < \delta$ such that $\hat{\theta}$ is correct under q.*

Lemma 3. *For P-almost every sequence of states, for all experts j,*

$$\lim_{T \to +\infty} \frac{1}{T} \sum_{t=1}^{T} [S(\theta_t^j, \omega_t^*) - E[S(\theta_t^j, \omega_t)|\omega_1^*, \ldots, \omega_{t-1}^*]) = 0 \ .$$

The first two lemmas are elementary and their proofs relegated to the Online Appendix. For Lemma 3, see Shiryaev (1996), chapter 7, section 3, corollary 2.

As S is strictly proper and expert i provides correct forecasts, for all other experts j,

$$\mathsf{E}[S(\theta_t^i, \omega_t)|\omega_1^*, \ldots, \omega_{t-1}^*] \geq \mathsf{E}[S(\theta_t^j, \omega_t)|\omega_1^*, \ldots, \omega_{t-1}^*] \ .$$

Besides, noting that the performance measure at period t equals the average score over the first t periods, a direct application of Lemma 3 yields

$$\liminf_{t \to +\infty} (\pi_t^i - \pi_t^j) \geq 0 \ .$$

for P-almost every sequence of states. This proves the first point.

Now take any $\delta > 0$. From Lemma 2, there exists $\epsilon > 0$, and some distribution $Q \in \Delta\Omega^\infty$, such that $Q \in \mathcal{P}_\delta$ and also such that, whenever

$$\mathsf{E}[S(\theta_t^j, \omega_t)|\omega_1^*, \ldots, \omega_{t-1}^*] \geq \mathsf{E}[S(\theta_t^i, \omega_t)|\omega_1^*, \ldots, \omega_{t-1}^*] - \epsilon$$

it is the case that θ_t^j is correct under Q, conditionally on the first $t-1$ realized states.

Applying Lemma 3 to both experts, we get that, for P-almost every sequence of states,

$$\lim_{T \to +\infty} \frac{1}{T} \sum_{t=1}^{T} [E[S(\theta_t^i, \omega_t)|\omega_1^*, \ldots, \omega_{t-1}^*] - E[S(\theta_t^j, \omega_t)|\omega_1^*, \ldots, \omega_{t-1}^*]] = 0$$

whenever expert j's performance is asymptotically as good as expert i's. Then Lemma 1 implies

$$\lim_{T \to \infty} \frac{1}{T} |\{\mathsf{E}[S(\theta_t^j, \omega_t)|\omega_1^*, \ldots, \omega_{t-1}^*] \geq \mathsf{E}[S(\theta_t^i, \omega_t)|\omega_1^*, \ldots, \omega_{t-1}^*] - \epsilon\}| = 1 \ ,$$

which makes the sequence of forecasts of expert j essentially correct. This proves the second point. □

This result formalizes two important facts. First, any true expert is guaranteed to maximize his performance measure in the long run. Second, whenever another expert has a performance that asymptotically matches that of a true expert, it must be the case that the predictions of this expert are nearly as accurate as those of the true expert, for all periods except possibly a proportion that vanish to zero.

Evaluating experts in this fashion gives a test that separates the “true” from the “false” experts, as do the tests proposed by Al-Najjar and Weinstein (2008); Feinberg and Stewart (2006) for the case of reports of the full distribution. As in these papers, I find that the presence of a true expert in a group makes it possible to detect the charlatans. Note that, because the cross-calibration test of Feinberg and Stewart relies mainly on a convexity argument, any discrete statistic whose level-sets form a convex partition of distribution can be comparative-tested by cross-calibration. In particular, the cross-calibration test applies to any statistic that can be elicited via a strictly proper scoring rule. As in these papers, in absence of true experts the scoring method does not permit the detection of the “false” experts. Indeed, if false experts were to coordinate and provide the exact same forecast, we would essentially face a single expert. The impossibility results hold even in absence of coordination, as demonstrated by Olszewski and Sandroni (2009b). It is known that, when forecasting probability estimates, no reasonable test can separate the “true” from the “false” experts (Olszewski and Sandroni, 2008, 2009c; Shmaya, 2008). The technical arguments of these papers also carry over to statistics with convex level-sets.

6 Concluding Remarks

This paper studies the problem of eliciting (or evaluating) expert forecasts regarding some statistic of interest. The expert’s payoff is controlled through a (generalized) scoring rule, whose inputs are the announced statistical forecast and the state drawn by Nature. I focus on two classes of statistics: the statistics that are continuous and take a continuum of values, and the statistics that are discrete and take a finite number of values. In both cases, I provide a geometric characterization of the statistics that can be elicited via strictly proper scoring rules. For those statistics, I also describe the entire collection of the proper and strictly proper scoring rules. For statistics that take value in a naturally ordered set, the characterizations are extended to account for order sensitivity.

Two points deserve mention. First, many statistics fail to admit strictly proper scoring rules. But the difference between what can be elicited and what cannot is more a matter of degree than a dichotomy. Eventually, all statistics can be elicited in some indirect fashion. We can always elicit forecasts of the full probability distribution and subsequently compute the value of any statistic. The results of this paper assert that, sometimes, a statistic does not convey enough information to create strict incentives. In such cases, one must rely on a finer statistic. Some statistics provide more information than others. In that respect, the full distribution conveys the most information. A trade-off appears: we want enough information to be able to offer strict incentives, but we don't want too much information either, to remain within reasonable bounds of communication. And indeed the full distribution is not always required. For example, forecasts of the variance only cannot be elicited, but forecasts of the mean *and* the variance together can be elicited. This is a mere consequence of the fact that the mean and the variance are isomorphic to the first and second moments, which both admit strictly proper scoring rules. This paper focuses on the statistics that can be elicited directly. But clearly, there is also a lot to be said for the statistics that cannot.

Second, the paper embraces the setting of the seminal work by Savage (1971). Results are presented in their full generality, with no tie to any particular economic environment. What happens when two or more experts compete with one another? When the expert bears some cost to gain information? When he may find himself unable to acquire proper knowledge? What happens when the expert and the decision-maker have conflicting interests? Or when the observability of states is contingent upon the decision? In all these questions, strictly proper scoring rules remain central in the design of appropriate incentive schemes. But not all should be retained. Each question elaborates its own set of constraints, which has the effect of keeping only a small subset of all strictly proper scoring rules, as the preceding section suggests. Knowing which ones to keep and which ones to discard is an important, and often difficult matter, that constitutes a natural venue for future research.

References

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

N.I. Al-Najjar, A. Sandroni, R. Smorodinsky, and J. Weinstein. Testing theories with learnable and predictive representations. *Journal of Economic Theory (forthcoming)*, 2010.

F. Aurenhammer. Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16:78, 1987.

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

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

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

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

A.P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007. ISSN 0020-3157.

M. De Berg, O. Cheong, and M. van Kreveld. *Computational geometry: algorithms and applications*. Springer, 2008.

B. De Finetti. Does it make sense to speak of "good probability appraisers". *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, pages 257–364, 1962.

E. Dekel and Y. Feinberg. Non-bayesian testing of a stochastic prediction. *Review of Economic Studies*, 73(4):893–906, 2006.

I.I. Eremin. *Theory of linear optimization*. VSP International Science Publishers, 2002.

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

Y. Feinberg and C. Stewart. Testing multiple forecasters. *Econometrica*, 76:561–582, 2006.

L. Fortnow and R. Vohra. The complexity of testing forecasts. *Econometrica*, 77: 93–105, 2009.

D.P. Foster and R.V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379, 1998.

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

D. Fudenberg and D.K. Levine. An easier way to calibrate. *Games and Economic Behavior*, 29(1-2):131–137, 1999.

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

IJ Good. *Rational decisions*. Springer-Verlag, 1997.

A.D. Hendrickson and R.J. Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42(6):1916–1921, 1971.

H. Imai, M. Iri, and K. Murota. Voronoi diagram in the laguerre geometry and its applications. *SIAM Journal on Computing*, 14:93, 1985.

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

E. Karni. A mechanism for eliciting probabilities. *Econometrica*, 77(2):603–606, 2009.

E. Lehrer. Any inspection is manipulable. *Econometrica*, 69(5):1333–1347, 2001.

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

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

R.E. Megginson. *An introduction to Banach space theory.* Springer Verlag, 1998.

T. Offerman, J. Sonnemans, G. Van De Kuilen, and P.P. Wakker. A truth serum for non-Bayesians: Correcting proper scoring rules for risk attitudes. *Review of Economic Studies*, 76(4):1461–1489, 2009. ISSN 1467-937X.

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

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

W. Olszewski and A. Sandroni. A non-manipulable test. *Annals of Statistics*, 37(2): 1013–1039, 2009a.

W. Olszewski and A. Sandroni. Manipulability of comparative tests. *Proceedings of the National Academy of Sciences*, 106(13):5029, 2009b.

W. Olszewski and A. Sandroni. Strategic manipulation of empirical tests. *Mathematics of Operations Resesearch*, 34:57–70, 2009c.

W. Olszewski and A. Sandroni. Falsifiability. *American Economic Review (forthcoming)*, 2010.

K. Osband. Optimal forecasting incentives. *The Journal of Political Economy*, 97(5): 1091–1112, 1989.

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

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

A. Sandroni, R. Smorodinsky, and R.V. Vohra. Calibration with many checking rules. *Mathematics of Operations Research*, 28(1):141–153, 2003.

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

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

R. Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1):43–62, 1998. ISSN 1386-4157.

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

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

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

R.L. Winkler. Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, pages 1073–1078, 1969.

R.L. Winkler and A.H. Murphy. "Good" probability assessors. *Journal of Applied Meteorology*, 7(5):751–758, 1968.