JMLR: Workshop and Conference Proceedings vol 23 (2012) 27.1–27.13 25th Annual Conference on Learning Theory

A Characterization of Scoring Rules for Linear Properties

Jacob D. Abernethy JABER@SEAS.UPENN.EDU*
Computer and Information Science, University of Pennsylvania, Philadelphia, PA

Rafael M. Frongillo RAF@CS.BERKELEY.EDU†
Computer Science Division, University of California at Berkeley

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

We consider the design of *proper scoring rules*, equivalently *proper losses*, when the goal is to elicit some function, known as a *property*, of the underlying distribution. We provide a full characterization of the class of proper scoring rules when the property is linear as a function of the input distribution. A key conclusion is that any such scoring rule can be written in the form of a *Bregman divergence* for some convex function. We also apply our results to the design of prediction market mechanisms, showing a strong equivalence between scoring rules for linear properties and automated prediction market makers.

1. Introduction

In Machine Learning, we use loss functions as a way to measure the performance of a hypothesis on a set of data. In another sense, loss functions are a way to incentivize "correctness" in a learning algorithm. If we consider algorithms which output probability predictions, then we say that a loss is *proper* if, given any distribution P on an outcome y, an algorithm (forecaster) minimizes expected loss by actually reporting P. In the setting of multiclass classification, where our goal is to predict a distribution $\hat{P}$ from the n-dimensional probability simplex Δ_n, $\ell(\hat{P}, y)$ is a proper loss if, given any $P \in \Delta_n$, $P \in \arg\min_{\hat{P}} \mathbb{E}_{y\sim P}\, \ell(\hat{P}, y)$.

In the multiclass setting our goal is to incentivize a report of the full distribution of y, but there are several reasons why we might go further. Indeed, there are any number of scenarios in which we are given data from some distribution and we would like our forecaster not to report an entire distribution, which can be computationally impossible anyway, but rather some simpler object – e.g. a set of summary statistics, or even a data classifier. Formally, given a space of distributions $\mathcal{P}$, we shall call $\Gamma : \mathcal{P} \to \mathbb{R}^n$ a *property*, and $\Gamma(P)$ will contain all of the information about P for the task at hand. As a very pertinent example, if $P \in \mathcal{P}$ is a distributions on pairs $(\mathbf{x}, y)$, with $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$, then one interesting choice of Γ is the linear regressor $\mathbf{w} \in \mathbb{R}^d$ that minimizes the expected squared loss $(\mathbf{w} \cdot \mathbf{x} - y)^2$.

† Supported by a Simons Foundation Postdoctoral Fellowship
† Supported by NSF grant CC-0964033, a Google University Research Award, and the National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a

In the present paper we use the terminology of a *scoring rule*, terminology drawn from the economics literature which, in essence, denotes the same concept as a loss function but in which the goal is to maximize rather than minimize. A scoring rule $s[\cdot](\cdot)$ is said to be *proper* for property Γ if $\Gamma(P) \in \arg\max_{r \in \mathbb{R}^d} \mathbb{E}_{\omega \sim P}\, s[r](\omega)$ for every distribution P; s is *strictly proper* if $\Gamma(P)$ is the only member of the $\arg\max$ for every P. A natural question to pose is, given any Γ, under what conditions can we design a score s that is (strictly) proper for Γ? Moreover, can we provide a full classification of all such scores? These relatively broad questions were perhaps first proposed by Lambert et al. (2008).

In the present paper, we provide a fully complete answer to the latter questions for a particular class of properties, namely those Γ which are *linear* in the distribution. The central conclusion of the present paper is that any scoring rule for a linear property Γ must take the form of a *Bregman divergence*.

Scoring rules have been studied for some time, but there has been renewed focus on the topic particularly due to their relationship to the design of *prediction markets*. A prediction market is a financial mechanism whose purpose, given some uncertain future outcome, is to aggregate the subjective probability beliefs of this outcome from a large crowd of individuals. Hanson (2003) showed how scoring rules can be used in a very simple fashion to construct such markets, and there has since been much work strengthening this connection. As our results relate quite strongly to this body of work we shall take a tour through some of the recent literature to emphasize these connections.

2. Previous Work and a Discussion of Results

Many authors point to the paper of Brier (1950) as the earliest mention of what we now call a *proper scoring rule*, or a *proper loss*. The paper, published in the *Monthly Weather Review*, observed that verifying probability forecasts can be tricky and proposed a tool, now known as the Brier score, to measure predictions once an outcome is known. Twenty years later, in what is widely considered to be the seminal work on the topic, Leonard Savage published a very general treatise on eliciting "personal probabilities," i.e. subjective probability estimates Savage (1971). In its simplest incarnation, a proper scoring rule is simply a function $S(\cdot;\cdot)$ taking two inputs, a probability distribution forecast $\hat{P}$ and an outcome x from a finite set, with the property that if x is sampled according to some "true" distribution P then for any $\hat{P}$

$$\mathbb{E}_P[S(P,x)] \geq \mathbb{E}_P[S(\hat{P},x)].$$

In other words, the forecaster maximizes the expected value of S by reporting the true distribution. One of the most well-known examples is the *logarithmic scoring rule* defined by $S(P,x) := \log P(x)$, and it is an easy exercise to see that this satisfies the desired property. Among many useful observations, Savage makes a point that one can construct a scoring rule using any strictly convex function. A more modern discussion of the topic can be found in Gneiting and Raftery (2007).

2.1. Design of Prediction Markets

It was observed by Hanson (2003) that one can use a scoring rule not only to elicit correct forecasts from a single individual but also to design a *prediction market*. In such a market, traders would have the ability to place bets with a central authority, known as a *market maker*, and the market maker would continue to publish a joint forecast representing the "consensus hypothesis" of the distribution. The framework is remarkably simple, and we describe it here. The market maker publishes a proper scoring rule S and an initial probability estimate P_0. On each round t in a sequence, the current consensus probability P_t is posted, and any trader can place a bet by *modifying* the probability to any desired value P_{t+1}. In the end, the true outcome x is revealed to the world, each trader receives a (potentially negative) profit of

$$S(P_{t+1}, x) - S(P_t, x). \tag{1}$$

Notice two facts about this framework: (a) if a trader at time t knows the true probability P^* then he always maximizes expected profit by setting $P_{t+1} = P^*$ and (b) because of the telescoping sum, if P_T is the final estimated probability then the market maker needs only to pay out a total of $S(P_0, x) - S(P_T, x)$. Hanson referred to this form of prediction market as a *market scoring rule* and, when the log scoring rule from above is used, this was called the Logarithmic Market Scoring Rule (LMSR).

Hanson's prediction market framework, which requires traders to make probability estimates and judges them according to a scoring rule, does not fit into our typical understanding of betting markets, as well as other financial markets, in which parties buy and sell contracts whose payoff is contingent on future outcomes. One such type of contract is the *Arrow-Debreu security*, which pays off a unit of currency if a given event occurs, or it pays off nil. A question one might ask is whether we can convert the market scoring rule betting language, in which traders are asked to report probability predictions, to one in which traders simply purchase bundles of Arrow-Debreu securities at prices set by the market maker. Indeed, Chen and Pennock (2007) showed that this is possible for a certain class of market scoring rules and proposed a market formulation based on a "cost function", which we sketch here:

- Some future outcome $i \in \{1, \ldots, n\}$ will occur, and the market maker sells an Arrow-Debreu security for each outcome: contract j pays \$1 if and only if outcome j occurs. Market maker has a convex differentiable $C : \mathbb{R}^n \to \mathbb{R}$ and maintains a "quantity vector" $\mathbf{q} \in \mathbb{R}^n$, where initially $\mathbf{q}$ is set to be the vector of 0's.
- At any point in time, a trader may purchase a "bundle" of shares described by $\mathbf{r} \in \mathbb{R}^n_{\geq 0}$; that is, r_i is the number of shares purchased for outcome i. Given current quantity vector $\mathbf{q}$, the price for bundle $\mathbf{r}$ is $C(\mathbf{q} + \mathbf{r}) - C(\mathbf{q})$. After selling $\mathbf{r}$ to the trader, the market maker then updates $\mathbf{q} \leftarrow \mathbf{q} + \mathbf{r}$.
- At the close of the market, when the outcome i is revealed, the market maker has to make a payout to all of the winning contracts, which is a total cost of q_i.

Notice that the derivative $\nabla C(\mathbf{q})$ is essentially the market estimate of the true distribution on the outcome, since $\nabla_i C(\mathbf{q})$ is the marginal cost of a tiny purchase of contract i, which in equilibrium

would be the expected return of the contract; that is, the probability. Indeed, to avoid arbitrage opportunities the market maker must ensure that $\nabla C(\mathbf{q})$ is always a distribution. Chen and Pennock (2007) showed that one can replicate the LMSR by using this cost-function framework, with $C(\mathbf{q}) := \alpha^{-1} \log\left(\sum_i \exp(\alpha q_i)\right)$, for any parameter $\alpha > 0$.

2.2. Markets for Large Outcome Spaces

An important problem with the prediction market frameworks we have proposed thus far is that they are not practical for large outcome spaces. Imagine a scenario where the outcome is a combinatorial object, like the joint outcome of a single-elimination tournament with n teams. In the case of the market scoring rule, we must ask each participant to submit the *entire distribution* over the outcome space according to his belief. In the cost-function framework, the market maker is required to sell an Arrow-Debreu security for each of the possible outcomes. Clearly neither of these will be feasible for large n. One natural solution is to consider a small set of marginal probabilities, and to have the betting language depend only on these values. It has been considered whether a market maker can efficiently simulate LMSR pricing within this betting language, yet a large number of these results have been negative Chen et al. (2007, 2008).

Abernethy et al. (2011) proposed a new framework for combinatorial prediction market design which avoids some of these hardness issues. The idea is best explained by way of example. Imagine a round-robin tournament which ends up with a ranking of all n teams. Rather than have a single contract corresponding to each of the $n!$ outcomes, a market maker can sell only $\binom{n}{2}$ contracts, one for each pair i, j corresponding to the predicate "does team i rank higher than team j?" This is often called an *incomplete* market, as the traders can only express beliefs in this lower-dimensional contract space. Nevertheless, we can still use a cost function $C : \mathbb{R}^{\binom{n}{2}} \to \mathbb{R}$ to price these contracts as we did in the complete market setting. The market maker will maintain a quantity vector $\mathbf{q} \in \mathbb{R}^{\binom{n}{2}}$, and will price a bundle of contracts $\mathbf{r} \in \mathbb{R}^{\binom{n}{2}}$ according to the rule $C(\mathbf{q} + \mathbf{r}) - C(\mathbf{q})$. Given any final ranking of the n teams, we can describe the payoffs of all contracts by some $\mathbf{x} \in \{0, 1\}^{\binom{n}{2}}$. The trader who previously purchased bundle $\mathbf{r}$ will receive $\mathbf{r} \cdot \mathbf{x}$.

In this setting, how ought we design C? Previously, in the complete market setting, we noted that ∇C should always be a distribution. Abernethy et al. (2011) showed that, in a similar vein, C must have the property that $\{\nabla C(\mathbf{q}) : \mathbf{q} \in \mathbb{R}^{\binom{n}{2}}\}$ must identically be the convex hull of all payout vectors $\mathbf{x}$ over all the $n!$ possible outcomes. If we let H denote this convex hull, then we can construct C via *conjugate duality*. Let R be some strictly convex function with domain H, then we can set $C(\mathbf{q}) := \sup_{\mathbf{z} \in H} \mathbf{z} \cdot \mathbf{q} - R(\mathbf{z})$. It is shown by Abernethy et al. (2011) that this construction is sufficient to guarantee the desired properties of C.

2.3. Eliciting Expectations using a Bregman Divergence

Staying with the framework of Abernethy et al. (2011) for a moment, let us now consider what is a trader's objective. We will assume this trader wants to optimize his profit, and we can imagine that the trader maintains some belief distribution D on the set of all achievable outcome vectors $\mathbf{x} \in \{0, 1\}^{\binom{n}{2}}$. If the current quantity vector is $\mathbf{q}$, then a trader's expected profit of a bundle purchase

$\mathbf{r}$ under the presumed distribution D is

$$\text{Profit}(\mathbf{q} \to \mathbf{q}+\mathbf{r}|D) = \mathbb{E}_{\mathbf{x}\sim D}[\mathbf{r}\cdot\mathbf{x} - (C(\mathbf{q}+\mathbf{r}) - C(\mathbf{q}))].$$

With some work, this can be written in terms of *Bregman divergence* with respect to the conjugate function R of C:

$$\text{Profit}(\mathbf{q} \to \mathbf{q}+\mathbf{r}|D) = D_R(\mathbb{E}_D\mathbf{x}, \nabla C(\mathbf{q})) - D_R(\mathbb{E}_D\mathbf{x}, \nabla C(\mathbf{q}+\mathbf{r})), \tag{2}$$

where we define[1] $D_f(\mathbf{x},\mathbf{y}) := f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})\cdot(\mathbf{x}-\mathbf{y})$ for any convex function f, and observe that $D_f(\mathbf{x},\mathbf{y}) \geq 0$ via the convexity assumption. Notice that a profit-maximizing choice of $\mathbf{r}$ will bring $\nabla C(\mathbf{q}+\mathbf{r})$ to be identically $\mathbb{E}_D\mathbf{x}$, causing the second divergence term to vanish. It is then no wonder that the gradient space of C should be identified with the convex hull of the possible payoffs $\mathbf{x}$.

The profit description in (2) looks quite similar to the market scoring rule profit in (1) – and this relationship has been discussed in Chen and Vaughan (2010) for the case of complete markets. Indeed, the connection between Bregman divergences and proper scoring rules goes back to Savage (1971) who showed that one can design a scoring rule to elicit probability predictions from the divergence with respect to an arbitrary differentiable convex function (although he did not use the Bregman terminology). Further discussion can be found in Grünwald and Dawid (2004), Dawid (2006), Gneiting and Raftery (2007), and Banerjee et al. (2005). While using a somewhat different terminology, one can find a very thorough treatment in Reid and Williamson (2010) and Vernet et al. (2011).

2.4. General Elicitation and Our Results

It was perhaps Lambert et al. (2008) who first considered the following general problem: given an outcome space Ω and an *arbitrary* map $\Gamma : \Delta_\Omega \to \mathbb{R}$, under what circumstances can we construct a proper scoring rule $s : \mathbb{R}\times\omega \to \mathbb{R}$ for Γ, i.e. where

$$\Gamma(P) \in \arg\min_{r\in\mathbb{R}} \mathbb{E}_{\omega\sim P}[s(r,\omega)]$$

for every $P \in \Delta_\Omega$? In other words, under what circumstances can we construct an *incentive-compatible* payment scheme for Γ, so that a forecaster knowing the true distribution P maximizes expected utility by reporting $\Gamma(P)$? Moreover, can we provide a full classification of which functions exhibit this property? Lambert et al. (2008) makes a number of significant contributions towards these goals, although their results are largely focused on scalar properties, that is where Γ is real valued.

In the present paper we consider vector-valued properties. One of our main contributions is to give a full and very general characterization of proper scoring rules for the case when Γ is *linear* in the distribution – this is equivalent to saying that $\Gamma(P)$ is an expectation according to P. We show that, in a very strong sense, any proper scoring rule $s(\mathbf{r},\omega)$ for a linear Γ can be written identically in terms of some Bregman divergence. A similar result appeared in Banerjee et al. (2005); our method more closely follows that of Gneiting and Raftery (2007) and hence requires fewer regularity conditions and is more general.

1. A more precise definition is given later

3. Definitions

We use Ω to denote the (possibily infinite) set of possible "outcomes" of some future event, and we let Σ be a sigma algebra over Ω. We let $\mathcal{P} = \Delta_{|\Omega|}$ denote the set of probability measures on (Ω, Σ). We denote by δ_ω the probability measure placing all weight on $\omega \in \Omega$.

Following Gneiting and Raftery (2007), we say that a function $g : \Omega \to \mathbb{R}$ is $\mathcal{P}$-integrable if g is measurable with respect to Σ and integrable with respect to every $\mu \in \mathcal{P}$. Similarly, a function $g : \Omega \to \overline{\mathbb{R}}$ is *$\mathcal{P}$-quasi-integrable* if g is measurable with respect to Σ and *quasi-integrable* (having possibly infinite but not indeterminate integrals – see Bauer (2001)) with respect to all $\mu \in \mathcal{P}$. Here $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ denotes the extended real numbers.

We let U denote the *information space*, and we shall assume that $U \subseteq \mathbb{R}^k$ and that U is convex. A distributional *property* will simply be a function $\Gamma : \mathcal{P} \to U$. We will denote the range of Γ by $R_\Gamma \subseteq U$. One could think of $\Gamma(\mu)$ for some $\mu \in \mathcal{P}$ to be the "relevant information about μ." We will refer to the *level set* of Γ to mean $\{\mu \,|\, \Gamma(\mu) = r\}$ for some r, also denoted $\Gamma^{-1}(r)$.

Definition 1 *Assume we are given a sample space Ω with sigma algebra Σ, a set of distributions $\mathcal{P}$, and an information space U. A* scoring rule *is any function $s : U \to (\Omega \to \overline{\mathbb{R}})$ such that $s[r]$ is $\mathcal{P}$-quasi-integrable for all $r \in U$. For simplicity, we will write $s[r](\mu)$ to denote $\int_\Omega s[r]\, d\mu$.*

Definition 2 *We say that a scoring rule s is* proper *for a property Γ if for all $\mu \in \mathcal{P}$,*

$$\Gamma(\mu) \in \operatorname*{argmin}_{r \in R_\Gamma} \left\{ \int_\Omega s[r]\, d\mu \right\}. \tag{3}$$

If the argmin *in (3) is always unique for every μ, then we say s is* strictly proper*, and that s* elicits *Γ. Conversely, a property Γ is* elicitable *if there is some strictly proper scoring rule s for Γ.*

Given any scoring rule s, one can construct a modified scoring rule s' via a simple rule: for all $r \in U, \omega \in \Omega$, let $s'[r](\omega) := s[r](\omega) + g(\omega)$, where $g : \Omega \to \mathbb{R}$ is arbitrary. By adding a term which depends solely on ω (and not r), we have not changed the elicitation property of s, as $\operatorname{argmin}_{r \in U} s[r](\mu) = \operatorname{argmin}_{r \in U} s'[r](\mu)$ for all distributions μ. This leads us to the following definition.

Definition 3 *Given a pair of scoring rules $s, s' : U \to (\Omega \to \overline{\mathbb{R}})$ and any subset $V \subset U$, we say that s and s' are* equivalent on V*, $s \cong_V s'$, when there exists a $\mathcal{P}$-integrable function $g : \Omega \to \mathbb{R}$ such that $s[r](\omega) = s'[r](\omega) + g(\omega)$ for all $r \in V$ and all $\omega \in \Omega$. When $V = U$, then we simply say that s and s' are* equivalent *and write $s \cong s'$.*

4. Linear properties

Throughout the paper, we will assume our convex functions to be closed and proper.

Definition 4 *Let $f : U \to \mathbb{R}$ be convex, and let df be a function $U \to \mathit{Linear}(U)$, where $\mathit{Linear}(U)$ is the set of linear operators on U. For convenience, we shall typically write df_r instead of $df(r)$. Then df is a* subderivative *of f if for all $r, r' \in U$ we have*

$$f(r') - f(r) \geq df_r \cdot (r' - r). \tag{4}$$

Definition 5 *The* Bregman divergence *with respect to a convex function f with subderivative df is the function*

$$D_{f,df}(x, y) = f(x) - f(y) - df_y \cdot (x - y). \tag{5}$$

When f is differentiable, and hence there is only one unique subderivative $df := \nabla f$, we will suppress df and simply write $D_f(x, y)$.

We note one crucial property, which is that a Bregman divergence is always nonnegative, and when f is strictly convex then we have that $D_f(x, y) = 0$ if an only if $x = y$. We would now like to take a given Bregman divergence $D_{f,df}$ and construct a scoring rule from it, an object we will call a *Bregman score*, following Grünwald and Dawid (2004).

Definition 6 *Given any $\mathcal{P}$-integrable function $\rho : \Omega \to U$, we can associate a* Bregman score *$s : U \to (\Omega \to \overline{\mathbb{R}})$ to the triple (f, df, ρ), where $f : U \to \mathbb{R}$ is convex and df is an associated subderivative, where*

$$s[r](\omega) := -D_{f,df}(\rho(\omega), r) + f(\rho(\omega)) = f(r) + df_r \cdot (\rho(\omega) - r) \tag{6}$$

Note that the term $f(\rho(\omega))$ was chosen to simplify the expression for s, but as observed above, does not change the elicitation properties: s is equivalent to the score $s'[r](\omega) = -D_{f,df}(\rho(\omega), r)$. It is natural to ask what property is elicited by this Bregman score, and we answer this via the following simple computation.

$$\begin{aligned} s[r](\mu) = \int_\Omega \Big(f(r) + df_r \cdot (\rho(\omega) - r) \Big)\, d\mu(\omega) &= f(r) + df_r \cdot \left(\textstyle\int_\Omega \rho \, d\mu - r\right) \\ &= -D_{f,df}\Big(\textstyle\int_\Omega \rho \, d\mu, r\Big) + f\Big(\textstyle\int_\Omega \rho \, d\mu\Big). \end{aligned} \tag{7}$$

We have written the expression this way to single out the variable r, which only occurs in the divergence term. By the observation above about the minima of divergences, it becomes immediately clear that this function is maximized at $r = \int_\Omega \rho \, d\mu$. Hence, s is proper for the property $\Gamma_s(\mu) = \int_\Omega \rho \, d\mu$. Moreover, f is strictly convex if and only if s is strictly proper.

We have just shown how to construct a Bregman score, as well as describe the property Γ that it elicits. We shall refer to this property as *linear*, as the map $\mu \mapsto \int_\Omega \rho \, d\mu$ is linear in the input μ. As it turns out, given any linear property Γ, we now have a useful tool to design a large family of scoring rules that elicit Γ. If we set $\rho(\omega) := \Gamma(\delta_\omega)$ where δ_ω is the distribution with a point mass on ω, then it is easy to see that $\Gamma(\mu) = \int_\Omega \rho \, d\mu$ via the linearity of Γ.

From the above discussion, we have shown the following Lemma.

Lemma 7 *Let s be a (f, df, ρ) Bregman score. Then s is a proper scoring rule for $\Gamma : \mu \mapsto \int_\Omega \rho \, d\mu$, and is strictly proper if and only if f is strictly convex.*

The following definition captures a particular notion of differentiability which will be central to our characterization. It essentially implies smoothness within R_Γ, the range of Γ.

Definition 8 *Let $s : U \to (\Omega \to \overline{\mathbb{R}})$ be a proper scoring rule for property Γ. Given any $r \in R_\Gamma$ and $\mu \in \mathcal{P}$, we say that s is* Γ-differentiable *at (r, μ) if the directional derivative $\frac{d}{dt} s[r + tv](\mu)$ exists for all $v \in \mathbb{R}^k$ such that $r + \epsilon v \in R_\Gamma$ for sufficiently small $\epsilon > 0$. If this holds for all $\mu \in \Gamma^{-1}(r)$ and all $r \in \text{relint}(R_\Gamma)$, the relative interior[2] of R_Γ, we simply say that s is* Γ-differentiable.

Γ-differentiability is a weaker notion than differentiability, and allows for a broader class of scoring rules. The definition has two features: (a) we restrict our attention to how s behaves only within R_Γ and (b) we require that $s[r](\mu)$ is differentiable at a point $r = r_0$ *only* when $r_0 = \Gamma(\mu)$. Put another way, whenever the scoring rule is minimized, it must be differentiable, within R_Γ, at the minimizer.

Why provide such a weak definition of differentiability for scoring rules? It turns out that, for an arbitrary (non-smooth) convex function f with subderivative df, it is simply not the case that a (f, df, ρ) Bregman score is differentiable in general. On the other hand, every Bregman score does satisfy Γ-differentiability.

Lemma 9 *Any (f, df, ρ) Bregman score is Γ-differentiable for $\Gamma : \mu \mapsto \int_\Omega \rho \, d\mu$.*

Proof For any convex $f : U \to \mathbb{R}$ with subderivative df, and some $\mathcal{P}$-integrable function $\rho : \Omega \to U$, it is clear that the (f, df, ρ) Bregman score elicits the property $\Gamma(\mu) := \int_\Omega \rho \, d\mu$. Let us call this scoring rule s, and we recall from (7) that $s[r](\mu) = -D_{f,df}\Big(\Gamma(\mu), r\Big) + f\Big(\Gamma(\mu)\Big)$. To establish Γ-differentiability it is sufficient to show that $s[r](\mu)$ is differentiable in r at each pair (r, μ) for which $r = \Gamma(\mu)$. More generally, we will show that the function $g(y) := D_{f,df}(x, y)$ is differentiable at $y = x$ and, moreover, that $\nabla|_{y=x} g(y) \equiv 0$. Going a step further, it suffices to prove this statement in one dimension, i.e. when $U = \mathbb{R}$. This is because we can check the differentiability of $D_{f,df}(x, y)$ by checking the differentiability of the scalar function $h_v(t) := D_{f,df}(x, x + tv)$, at $t = 0$, for all unit vectors v. Of course, $h_v(t)$ is simply a divergence on $\mathbb{R}$ with respect to the convex function f and subderivative df when restricted to the set $\{x + tv : t \in \mathbb{R}\}$.

Let us now look at

$$\lim_{\epsilon \downarrow 0} \frac{D_{f,df}(x, x + \epsilon)}{\epsilon} = \lim_{\epsilon \downarrow 0} \frac{f(x) - f(x + \epsilon) - \epsilon df_{x+\epsilon}}{\epsilon} = f'_+(x) - \lim_{\epsilon \downarrow 0} df_{x+\epsilon} \tag{8}$$

where $f'_+(x)$ is the right derivative of f. It is clear that $f'_-(z) \leq df_z \leq f'_+(z)$ for all z, and we have from Rockafellar (1997) (Theorem 24.1) that $f'_+(x) \leq f'_+(x + \epsilon)$ and $\lim_{\epsilon \downarrow 0} f'_+(x + \epsilon) = f'_+(x)$. Combining these we see that $\lim_{\epsilon \downarrow 0} df_{x+\epsilon} = f'_+(x)$. We have now established that the limit in (8) vanishes. The argument holds for the left derivative as well by symmetry. ■

Γ-differentiability gives scoring rules for linear properties a lot of structure along level sets of a linear Γ. In essence, the scoring rule must behave the same way along every level set. This observation will allow us to write a scoring rule $s[r](\omega)$ in terms of r and $\Gamma(\delta_\omega)$, which will be crucial in our main proof.

2. We recall that the *relative interior* of a convex set $S \subset \mathbb{R}^n$ is the interior of the set after restricting to the smallest affine subspace containing S.

Lemma 10 *Let $s : U \to (\Omega \to \overline{\mathbb{R}})$ be a Γ-differentiable proper scoring rule for a linear property Γ. Then there exists a function $\sigma : \Omega \to \mathbb{R}$ such that for all $\mu_1, \mu_2 \in \mathcal{P}$ with $\Gamma(\mu_1) = \Gamma(\mu_2)$ such that for all $r \in$ relint(R_Γ),*

$$s[r](\mu_1 - \mu_2) = \int_\Omega \sigma \, d(\mu_1 - \mu_2). \tag{9}$$

Proof Let $\hat{r} = \Gamma(\mu_1) = \Gamma(\mu_2)$ and let $\mu_d = \mu_1 - \mu_2$. Note that $\Gamma(\mu_d) = 0$ by linearity, and hence for any $\mu \in \mathcal{P}$ we have $\Gamma(\mu + \mu_d) = \Gamma(\mu)$. Thus, these "level-set differences" μ_d are independent of the Γ-value of the level set; in light of this, define $D = \{\mu_d | \Gamma(\mu_d) = 0\}$. Now by assumption, s is Γ-differentiable at $(\hat{r}, \mu_1)$ and at $(\hat{r}, \mu_2)$, so for all appropriate $v \in \mathbb{R}^k$ we have

$$\frac{d}{dt} s[\hat{r} + tv](\mu_d) = \frac{d}{dt} s[\hat{r} + tv](\mu_1) - \frac{d}{dt} s[\hat{r} + tv](\mu_2) = 0,$$

meaning s is Γ-differentiable at $(\hat{r}, \mu_d)$ as well. But as noted above, μ_d is independent of the Γ value $\hat{r}$, and so s is Γ-differentiable at (r, μ_d) for all $r \in$ relint(R_Γ) and all $\mu_d \in D$. Moreover, all defined directional derivatives of s are 0 at any such (r, μ_d).

We now note that R_Γ is a convex set, being the range of a linear function with a convex domain. Thus, for any $r \in$ relint(R_Γ), the path $p(t) := \hat{r} + t(r - \hat{r})$, satisfies $p(t) + \epsilon \frac{dp}{dt} \in$ relint(R_Γ) for sufficiently small $\epsilon > 0$, for all $0 \leq t < 1$. Using the observation above, it follows that s is Γ-differentiable at $(p(t), \mu_d)$ for all $0 \leq t < 1$ and $\mu_d \in D$. We can now integrate along p to obtain

$$s[\hat{r}](\mu_d) = s[p(0)](\mu_d) = s[p(1)](\mu_d) = s[r](\mu_d)$$

for all $r \in$ relint(R_Γ). Now define $\sigma(\omega) = s[\hat{r}](\omega)$ to complete the proof. ■

Theorem 11 *Let scoring rule $s : U \to (\Omega \to \overline{\mathbb{R}})$ be given. If s is Γ-differentiable and proper for a linear property Γ, then s is equivalent to some (f, df, ρ) Bregman score on* relint(R_Γ).

Proof We of course take ρ such that $\Gamma(\mu) = \int_\Omega \rho \, d\mu$ for all $\mu \in \mathcal{P}$. Note that Γ may be defined on finite linear combinations of $\mathcal{P}$ by

$$\Gamma\left(\sum_i^N \alpha_i \mu_i\right) = \int_\Omega \rho \, d\left(\sum_i^N \alpha_i \mu_i\right) = \sum_i^N \alpha_i \int_\Omega \rho \, d\mu_i. \tag{10}$$

We will denote the space of these linear combinations by $\mathrm{span}(\mathcal{P})$.

We now define a linear inverse $\hat{\mu}$ of Γ whose range lies in $\mathrm{span}(\mathcal{P})$. Choose a basis $\mathcal{B} = \{b_1, \ldots, b_k\}$ of R_Γ and for all i choose $\mu_i \in \Gamma^{-1}(b_i)$. Now define $\hat{\mu} : R_\Gamma \to \mathrm{span}(\mathcal{P})$ by $\hat{\mu}\left[\sum_i \alpha_i b_i\right] = \sum_i \alpha_i \mu_i$. By linearity of Γ then, for all $r \in R_\Gamma$ we have

$$\Gamma(\hat{\mu}[r]) = \Gamma\left(\sum_i \alpha_i \hat{\mu}[b_i]\right) = \sum_i \alpha_i \Gamma\left(\hat{\mu}[b_i]\right) = \sum_i \alpha_i b_i = r.$$

Now let M be a change of basis matrix from the standard basis $\{e_i\}_{i\in[d]}$ to $\mathcal{B}$. Let $v[r] \in \mathbb{R}^k$ be the vector with $v[r]_i = s[r]\left(\hat{\mu}\left[b_i\right]\right)$, and define $df_r = v[r]^\top M$. Then for $r' = \sum_i \alpha_i' b_i$, we have

$$df_r \cdot r' = v[r]^\top M r' = \sum_i \alpha_i' s[r]\left(\hat{\mu}\left[b_i\right]\right) = s[r]\left(\hat{\mu}\left[\sum_i \alpha_i' b_i\right]\right) = s[r]\left(\hat{\mu}\left[r'\right]\right). \tag{11}$$

Now define $f : R_\Gamma \to \mathbb{R}$ by $f(r) = df_r \cdot r$. Using (11) and the fact that s is proper for Γ, we can show that df is a subderivative of f:

$$\begin{aligned} f(r) + df_r \cdot (r' - r) &= df_r \cdot r + df_r \cdot (r' - r) = df_r \cdot r' \\ &= s[r]\left(\hat{\mu}\left[r'\right]\right) \le s[r']\left(\hat{\mu}\left[r'\right]\right) = f(r'), \end{aligned}$$

for all $r', r \in R_\Gamma$. Thus, f is convex, and we can consider a (f, df, ρ) Bregman score s':

$$s'[r](\omega) = f(r) + df_r \cdot (\rho(\omega) - r) = df_r \cdot \rho(\omega) = s[r]\left(\hat{\mu}\left[\rho(\omega)\right]\right).$$

Now by Lemma 10 there exists a function σ such that

$$s[r](\delta_\omega - \hat{\mu}[\rho(\omega)]) = \sigma(\omega) - \int_\Omega \sigma d\hat{\mu}[\rho(\omega)],$$

for all $\omega \in \Omega$ and all $r \in \text{relint}(R_\Gamma)$. Thus,

$$s[r](\omega) - s'[r](\omega) = s[r](\omega) - s[r]\left(\hat{\mu}\left[\rho(\omega)\right]\right) = \sigma(\omega) - \int_\Omega \sigma d\hat{\mu}[\rho(\omega)],$$

which is a function of ω alone and hence s is equivalent to the Bregman score s'. ■

We have now characterized how a scoring rule s for a linear property Γ can behave on $\text{relint}(R_\Gamma)$, but what if an elicitation falls out of this range? The only requirement of s is that such "inadmissible" values not be maxima for any distribution μ. Formally, for any $r \notin R_\Gamma$, we must have

$$s[r](\mu) \le s[\Gamma(\mu)](\mu) \tag{12}$$

for all $\mu \in \mathcal{P}$, with a strict inequality if s is to be strictly proper. Note that the right-hand side of (12) does not depend on r. Letting $F(\mu) = s[\Gamma(\mu)](\mu)$, the analog of our generalized entropy function on the probability space $\mathcal{P}$, we can express our condition (12) quite simply: $s[r]$ must be a subderivative of F, or a *strict subderivative* if s is to be strictly proper. We now have the following result, which summarizes our work on linear properties.

Theorem 12 *Let scoring rule $s : U \to (\Omega \to \overline{\mathbb{R}})$ and linear property $\Gamma : \mu \mapsto \int_\Omega \rho d\mu$ be given, and define $F(\mu) = s[\Gamma(\mu)](\mu)$. Then s is Γ-differentiable and proper for Γ if and only if s is equivalent to a (f, df, ρ) Bregman score on* $\text{relint}(R_\Gamma)$ *and $s[r]$ is a subderivative of F for all $r \notin R_\Gamma$.*

Moreover, s is strictly proper if and only if f is strictly convex and $s[r]$ is a strict subderivative of F for all $r \notin R_\Gamma$.

5. Bregman Scores, Prediction Markets, and Learning Mechanisms

Let us recall the market scoring rule concept mentioned in Section 2.1. Once we have a scoring rule $S(P, i)$ for eliciting full probability distributions $P \in \Delta_n$, Hanson observed that we can construct a prediction market for forecasting the true outcome i. The market maker simply needs to propose an inital estimate P_0 and let traders propose a series of "modifications" to this hypothesis. A trader that proposes $P_t \to P_{t+1}$, where $P_t, P_{t+1} \in \Delta_n$, earns as profit $S(P_{t+1}, i) - S(P_t, i)$ when the true outcome i is revealed.

Drawing from Hanson's market scoring rule, which created a financial mechanism to learn probability distributions "from the crowd," Abernethy and Frongillo (2011) proposed taking this idea further, to learn from more complex classes of hypothesis spaces. Their mechanism, called a *Crowdsourced Learning Mechanism* (CLM), may be described as follows. The mechanism organizers choose a hypothesis space $\mathcal{H}$, and outcome space $\mathcal{O}$, and a measure of relative performance $\texttt{Profit} : \mathcal{H} \times \mathcal{H} \times \mathcal{O} \to \mathbb{R}$. The organizers post an initial public hypothesis $\mathbf{w}_0 \in \mathcal{H}$, after which participants propose updates $\mathbf{w}_t \mapsto \mathbf{w}_{t+1}$ to the posted hypothesis, one at a time. The mechanism terminates by releasing a final outcome $X \in \mathcal{O}$, at which point the participants are paid $\texttt{Profit}(\mathbf{w}_t, \mathbf{w}_{t+1}; X)$ for each update $\mathbf{w}_t \mapsto \mathbf{w}_{t+1}$ they were responsible for.

Of course, the task of choosing this `Profit` function is of crucial importance. The authors of Abernethy and Frongillo (2011) consider machine learning problems which aim to minimize some loss function $L : \mathcal{H} \times \mathcal{O} \to \mathbb{R}$, where $\mathcal{H}$ is some arbitrary hypothesis space, and $\mathcal{O}$ is the test data space. If a data point $X \in \mathcal{O}$ is drawn according to some true distribution $P \in \Delta_{\mathcal{O}}$, then we hope to obtain a hypothesis in $\operatorname{argmin}_{\mathbf{w} \in \mathcal{H}} \mathbb{E}_{X \sim P}[L(\mathbf{w}; X)]$. Thus, to incentivize CLM participants to minimize this loss, we may simply choose

$$\texttt{Profit}(\mathbf{w}_t, \mathbf{w}_{t+1}; X) \propto L(\mathbf{w}_t; X) - L(\mathbf{w}_{t+1}; X). \tag{13}$$

This gives a crowdsourcing mechanism in the same vein of Hanson's market scoring rule. We will refer to this particular CLM, with `Profit` defined as the drop in a loss function L, as the *L-incentivized CLM*.

In Section 2 we also discussed the notion of "share-based" markets, which use securities whose payout is dependent on some outcome $X \in \mathcal{O}$. In this setting, traders purchase share bundles $\mathbf{r}$ for a price of $C(\mathbf{q} + \mathbf{r}) - C(\mathbf{q})$, where $\mathbf{q}$ is the current quantity vector. The payoff of these shares is determined by a function $\rho : \mathcal{O} \to \mathbb{R}$, where $\rho(X)_i$ is the payout of 1 share of security i when $X \in \mathcal{O}$ occurs. We will call such mechanisms *Automated Prediction Market Makers (APMM)*.

Viewing these shares purchases as moving the current quantity vector $\mathbf{q} \mapsto \mathbf{q} + \mathbf{r}$, and letting $\mathcal{H}$ denote the space of possible quantity vectors, it is clear that APMMs are just special cases of CLMs with $\texttt{Profit}(\mathbf{q}, \mathbf{q}'; X) = C(\mathbf{q}') - C(\mathbf{q}) + \rho(X) \cdot (\mathbf{q}' - \mathbf{q})$. Moreover, equation (2) implies that any APMM $(\mathcal{H}, \mathcal{O}, \rho, C)$ is an L-incentivized CLM, where $L(\mathbf{q}; X) = D_R(\rho(X), \nabla C(\mathbf{q}))$. This loss function looks like a Bregman score; to make a more formal statement, we will need a notion of *equivalence* between CLMs.

Definition 13 *We say that* CLM $A = (\mathcal{H}_A, \mathcal{O}, \texttt{Profit}_A)$ reduces to CLM $B = (\mathcal{H}_B, \mathcal{O}, \texttt{Profit}_B)$ *if there exists a map $\varphi : \mathcal{H}_A \to \mathcal{H}_B$ such that for each bet $w_1 \mapsto w_2 \in (\mathcal{H}_A \mapsto \mathcal{H}_A)$, we have*

$$\texttt{Profit}_A(\mathbf{w}_1, \mathbf{w}_2; X) = \texttt{Profit}_B(\varphi(\mathbf{w}_1), \varphi(\mathbf{w}_2); X). \tag{14}$$

If in addition B reduces to A, we say A and B are equivalent.

The central result of Abernethy and Frongillo (2011) is a strong connection between APMMs and L-incentivized CLMs for loss functions L based on a Bregman divergence. To relate these loss functions to our Bregman scores, we extend our definition slightly: for f differentiable let a (f, ρ, ψ) Bregman score be defined by $s[r](\omega) = -D_f(\rho(\omega), \psi(r))$. Now using our notion of equivalence we can restate their results as follows.

Theorem 14 (Abernethy and Frongillo (2011)) *Let $\mathcal{H}, \mathcal{H}'$ be hypothesis spaces with $\mathcal{H}' = \text{relint}(\mathcal{H}')$, and let $\psi : \mathcal{H} \to \mathcal{H}'$ such that $\rho(\mathcal{O}) \subseteq \psi(\mathcal{H})$. Then APMM $(\mathcal{H}, \mathcal{O}, \rho, C)$ is equivalent to an L-incentivized CLM for some differentiable C if and only if $s[\mathbf{w}](X) = -L(\mathbf{w}; X)$ is a (f, ρ, ψ) Bregman score for some differentiable f.*

Combining this result with that of Section 4, we have the following.

Theorem 15 *Let $\mathcal{H}, \mathcal{H}', \psi, \rho$ be as in Theorem 14, and let ψ^{-1} be a right inverse of ψ. Then the following are equivalent:*

1. *APMM $(\mathcal{H}, \mathcal{O}, \rho, C)$ is equivalent to an L-incentivized CLM for some differentiable C*
2. *$s[\mathbf{w}](X) = -L(\mathbf{w}; X)$ is a (f, ρ, ψ) Bregman score for some differentiable f*
3. *Scoring rule $s[r](X) = -L(\psi^{-1}(r); X)$ is proper for linear property $\Gamma : P \mapsto \mathbb{E}_{X \sim P}\, \rho(X)$*

References

J. Abernethy, Y. Chen, and J. Wortman Vaughan. An optimization-based framework for automated market-making. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 297–306, 2011.

J. D. Abernethy and R. M. Frongillo. A collaborative mechanism for crowdsourcing prediction problems. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2600–2608. 2011.

A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

H. Bauer. *Measure and integration theory*, volume 26. Walter de Gruyter, 2001.

G. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1): 1–3, 1950. ISSN 1520-0493.

Y. Chen and D. Pennock. A utility framework for bounded-loss market makers. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, pages 49–56, 2007.

Y. Chen and J. Vaughan. A new understanding of prediction markets via no-regret learning. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 189–198, 2010.

Y. Chen, L. Fortnow, E. Nikolova, and D. Pennock. Betting on permutations. In *Proceedings of the 8th ACM Conference on Electronic Commerce*, pages 326–335, 2007.

Y. Chen, L. Fortnow, N. Lambert, D. M. Pennock, and J. Wortman. Complexity of combinatorial market makers. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 190–199, 2008.

A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, Dec. 2006. ISSN 0020-3157, 1572-9052. doi: 10.1007/s10463-006-0099-8. URL http://www.springerlink.com/index/10.1007/s10463-006-0099-8.

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

P. Grünwald and A. Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *the Annals of Statistics*, 32(4):1367–433, 2004.

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

N. S. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008.

M. Reid and R. Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 11:2387–2422, 2010.

R. Rockafellar. *Convex analysis*, volume 28. Princeton Univ Pr, 1997.

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

E. Vernet, R. Williamson, and M. Reid. Composite multiclass losses. *NIPS*, 2011.