Maximum Entropy, Scoring Rules, and Prediction Markets

September 27, 2013

1 Scoring Rule

* Discrete or continuous outcome space.

* Want to elicit statistics of the distribution (not full distribution).

* Single-agent setting: scoring rules for expectations of random variables.

* Can also elicit one-to-one-functions of those expectations (e.g., mean and variance from first and second uncentered moments). Need to understand possible technical subtleties: see "functional invariance" property of maximum likelihood estimate.

* Strictly proper scoring rule implicitly defined via solution of a maximum entropy problem.

* Take log of maxent distribution with mean parameters given by the agent. Generalizes the usual log scoring rule for complete probability distributions.

* Although it's a generalization of the log scoring rule, may not look anything like log scoring rule: e.g. normal, exponential, beta distributions.

* Informational advantage: the agent doesn't need to know full probability density to be incentivized to reveal the statistics truthfully. Only needs to know the statistics.

* Does not have to agree that underlying density if from an exponential family.

* Technical subtlety (need to understand better): agent must agree on the *support* of the distribution. This is related to the "base measure" of the

exponential family.

* Important aspect of this approach: the primitive is now the *uncertainty function* (e.g., negative entropy) defined over the space of probability densities.

* Other uncertainty functions can be used, they just have to be convex (is that all?). But the point is that they're defined over the space of densities, and then the convex function over the space of statistics is derived from that primitive via an optimization problem.

* It would be nice to show how many properties of the market follow from properties of the uncertainty function used.

* Do all cost functions arise in this way? From an optimization problem having objective function an uncertainty function over probability densities? This is straightforward for proper scoring rules but not as much for *strictly* proper scoring rules.

* Connection between log scoring rule and maximum likelihood. This can give scoring rules for other kinds of parameters besides expectations? (Bounds on the support for the uniform distribution, for example). But it assumes that the agent's belief distribution takes a specific form (e.g., exponential, uniform, etc.) The maxent approach works for statistics of any underlying distribution, but is limited to expectations.

* Can we broaden the maxent scoring rule approach beyond expectations? Can we develop a scoring rule for the median with this approach?

* More generally, rather than just the maximum likelihood estimator, we can look at m-estimators to construct these kinds of scoring rules.

* Need to understand the consistency proofs of these kinds of estimators well.

2 Prediction Market

* Re-interpretation of the exponential family elements: sufficient statistic becomes payoff function, natural parameter is share vector, normalizer becomes cost function.

* Go over desirable properties of a price/cost function: monotonicity, path independence, etc.

* Important technical detail when looking at statistics: the domain of the cost function. For instance with first and second uncentered moments as statistics, can't have negative oustanding shares for the second security. How odd is this for practice?

* Insight: LMSR prediction market is estimating the natural parameter of an exponential family.

* Sequence of traders seeking to maximize profit can be viewed as MLE estimation of natural parameters.

* With budgets, adversarial traders can cause limited damage (in what formal sense again?)

* Budgets of informative traders grow in expectation over time

* This assumes a market that occurs over rounds, i.e., there is not just a single ultimate outcome (like a Presidential election) but several outcomes over time so that securities pay off.

* Need to think more about the model of bidder behavior: Bayesian updating? Budget constrained? Risk neutral or risk averse?

* Connection to mirror descent: the update with Bayesian traders is the mirror descent update. Frongillo et al. also observed this but in their model they have budget-constrained bidders with log utility, so there are some differences. Need to understand this better.

* What is the advantage of the mirror descent perspective/connection? Can we talk about the regret of the market and its convergence rate with Bayesian traders? How does it compare to the optimal rate (usually something like $\sqrt{T}$)?

* Note: The kind of aggregation performed by Bayesian update bidders is also very reminiscent of *exponential smoothing* (not related to "exponential" families, at least not directly as far as I'm aware). It's used for forecasting from time series data and has some nice properties.

* Many of the results for prediction market bidders currently assume that they believe the underlying distribution to be an exponential family. It would be nice to do away with that as much as possible to try to match the mild informational requirements from the scoring rule section.

3 Duality

* The duality between scoring and market scoring rules (i.e., cost function prediction markets) is understood from Hanson's work.

* Exponential families perspective make the duality in even more concrete: share vector and agent belief live in dual spaces. Scoring rules work on mean parametrization and prediction markets work on natural parametrization.

A Maximum Entropy Market Making

The purpose of a prediction market is to elicit and aggregate the beliefs (i.e., subjective probabilities) of agents over a space of *outcomes*. In a single-agent setting, a scoring rule is used to elicit the agent's beliefs. In a multi-agent setting, an information market is used to aggregate the agent's beliefs. Hanson [] introduced the idea of a market scoring rule, which inherits the appealing elicitation and aggregation properties of both so that they can perform well in both thin and thick markets. In this work we derive a wide-ranging generalization of the logarithmic market scoring rule. Using a maximum entropy approach, we explain how a market scoring rule can be developed for outcome spaces both discrete and continuous, and for generic properties of the underlying distribution (e.g., mean and variance), rather than just the probabilities of individual outcomes.

A.1 The Model

Let $\mathcal{X}$ be the outcome space, which may be discrete or continuous. Let $\mathcal{P}$ denote the set of probability distributions over the outcome space. We represent a probability distribution as a density p absolutely continuous with respect to some base measure ν (e.g., counting for discrete outcomes, or Lebesgue for continuous outcomes).

There is an unknown distribution p over outcomes. We are interested in estimating the expected value of different outcome *statistics* under this distribution by aggregating the beliefs of agents. A statistic is a real-valued function $\phi_s : \mathcal{X} \to \mathbf{R}$ over outcomes, independent of p; here s belongs to a finite index set $\mathcal{S}$ of size d. The collection of functions $\phi = (\phi_s,\, s \in \mathcal{S})$ can be viewed as a vector-valued statistic mapping outcomes into $\mathbf{R}^d$. Formally, the aim is to estimate

$$\mathbf{E}_p[\phi(x)] = \mu. \tag{1}$$

We stress that we are only seek to elicit $\mu \in \mathbf{R}^d$, not the full distribution p. As an example, suppose that $\mathcal{X} = \mathbf{R}$. If we were interested in eliciting the first two uncentered moments of p, we would include $\phi_1(x) = x$ and $\phi_2(x) = x^2$ as statistics. Note that the variance cannot be directly obtained through any single statistic: it corresponds to the expectation of $(x - \mathbf{E}_p[x])$, which depends on p, violating our definition of a statistic. Of course, the variance

can be indirectly obtained by eliciting the first two moments separately.[1]

There are two approaches to eliciting an agent's estimated μ in (1). The first approach is to incentivize the agent to directly report μ via a *scoring rule*. The second approach is to form an *information market* by issuing securities for each statistic $s \in \mathcal{S}$, with payoff depending on the realized value of the statistic. (Such securities are known as contingent claims.) The amount of shares of each security acquired by the agent indirectly reveals it estimate μ. In a market scoring rule [**?**], these two approaches are dual to each other in a formal sense.

A.2 Scoring Rule

We first consider how to directly elicit μ via a scoring rule. The space of feasible reports from an agent corresponds to

$$\mathcal{M} = \left\{ \mu \in \mathbf{R}^d : \mathbf{E}_p[\phi(x)] = \mu, \text{ for some } p \in \mathcal{P} \right\}. \tag{2}$$

A scoring rule is a function $S : \mathcal{M} \times \mathcal{X} \to \mathbf{R}$ that rewards an agent with $S(\mu, x)$ based on how its report μ agreed with the eventual outcome x. A scoring rule is *proper* if for all $\mu \in \mathcal{M}$, and $p \in \mathcal{P}$ such that $\mathbf{E}_p[\phi(x)] = \mu$, we have

$$\mathbf{E}_p[S(\mu, x)] \geq \mathbf{E}_p[S(\hat{\mu}, x)] \tag{3}$$

for all $\hat{\mu} \neq \mu$; the rule is *strictly proper* if the inequality is strict. Note that in the literature, scoring rules are defined over entire probability distributions, not just properties of those distributions. Our definition imposes a minimum of informational requirements on agents, because an agent does not need to assess the full distribution p that might have lead to its estimates μ in order to realize that truthful reporting is an optimal strategy.

Our scoring rule is implicitly defined via the solution of a mathematical program. The program computes the maximum entropy distribution consistent

[1] This is related to the notion of *elicitation complexity* of statistical properties introduced by Lambert et al. [**?**], but we do not pursue this connection here.

with the agent's reported beliefs μ.

$$\max_{p \geq 0} \quad -\int_{x \in \mathcal{X}} p(x) \log p(x) \nu(dx) \tag{4}$$

$$\text{s.t.} \quad \int_{x \in \mathcal{X}} \phi(x) p(x) \nu(dx) = \mu \tag{5}$$

$$\int_{x \in \mathcal{X}} p(x) \nu(dx) = 1 \tag{6}$$

The non-negativity constraints together with (6) ensure that the solution is a probability distribution. Constraints (5), which correspond to d separate constraints, ensure consistency with the agent's report. The objective (4) is the entropy of the solution p with respect to the measure ν. The program has a convex objective and linear constraints.

Theorem 1 *Let $p(x; \mu)$ be the optimal solution to the maximum entropy program, given report $\mu \in \mathbf{R}^d$. The scoring rule defined by*

$$S(\mu, x) = a_x + b \log p(x; \mu), \tag{7}$$

where $b, a_x \in \mathbf{R}$ and $b > 0$, is strictly proper.

This rule represents a generalization of the logarithmic scoring rule for probability distributions to any properties of distributions that can be captured as expectations of statistics. As explained in the next section, the solution to the maximum entropy program takes the form of an *exponential family* distribution. For many common properties of interest, these distributions are familiar; for example, the maximum entropy distribution with a given mean is an exponential distribution, and the maximum entropy distribution with a given mean and variance is a normal distribution []. We provide here two examples in more depth.

Multinomial Distribution As a first example, consider the problem of estimating the individual probabilities of a finite set of outcomes. We have $\mathcal{X} = \{1, \ldots, k\}$. The relevant statistics indicate which outcome actually occurs, so the index set is $\mathcal{S} = \mathcal{X}$. Statistic $\phi_x(x')$ is 1 if $x' = x$ and 0 otherwise. Observe that if p is the distribution over outcomes, then $\mathbf{E}_p[\phi(x)] = p$. A feasible report from an agent is any non-negative $\pi = (\pi_1, \ldots, \pi_k)$ such that $\sum_{i=1}^k \pi_i = 1$. Given a report of $\mu = \pi$, the unique solution to the maximum entropy program is $p(x; \pi) = \pi$, and (7) corresponds to the classic logarithmic scoring rule of $S(\pi, x) = \log \pi_x$. □

Normal Distribution As a second example, suppose the outcome space is $\mathcal{X} = \mathbf{R}$ and that we are interested in estimating the mean μ and variance σ^2 of the underlying distribution. We choose the first and second (uncentered) moments, $\kappa_1 = \mu$ and $\kappa_2 = \mu^2 + \sigma^2$ as our statistics. It is well-known that the maximum entropy distribution with a given mean and variance is the normal distribution, so the solution to the maximum entropy program (using Lebesgue as the base measure ν) is

$$p(x; \kappa) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

Applying Theorem 1 and choosing appropriate constants, we find that the following scoring rule is strictly proper in this context.

$$S(\kappa, x) = -\frac{(x-\mu)^2}{\sigma^2} - \log \sigma^2.$$

Note that the fact that this scoring rule is proper for the mean amounts to the well-known fact that reporting the mean is a Bayes act under quadratic loss [**?**]. □

A brief observation on parametrizations is in order. Note that the maximum entropy distribution here is parametrized by the uncentered moments $\kappa = (\kappa_1, \kappa_2)$ rather than (μ, σ^2) because the variance is not a valid statistic. However, it is clear that the agent does not have to directly report κ: the market maker itself can do the translation from reported mean and variance to uncentered moments. This is relevant because the κ parametrization may be unintuitive for an agent, especially given the constraints $\kappa_2 - \kappa_1^2 \geq 0$ that must be enforced. In the (μ, σ^2) parametrization each parameter is unrestricted, and the parameters have intuitive interpretations.

Our approach so far has been to identify relevant properties of the unknown distribution, and elicit them via a scoring rule computed through a maximum entropy program. However, it is the case that a distribution is a maximum entropy distribution if and only if it is an exponential family []. Therefore, we see the following alternative approach. The market designer can begin by considering the outcome space $\mathcal{X}$, and assessing what family of distributions over $\mathcal{X}$ the unknown p might come from. If this family is an exponential family, then Theorem 1 immediately gives a proper scoring rule to elicit the parameters of the distribution. Many of the most familiar distributions are exponential families, including the normal, exponential, gamma, beta, binomial, Poisson, Weibull, and Dirichlet. To illustrate, we consider two examples in more depth.

Exponential Suppose the designer would like an estimate of the rate at which servers fail in a data center. An outcome is the time between failures, so $\mathcal{X} = [0, +\infty)$. The usual distribution for failure times is the exponential distribution $p(x; \lambda) = \lambda e^{-\lambda x}$, which is parametrized by the rate λ and has mean $\mu = 1/\lambda$. The exponential distribution is the maximum entropy distribution given a fixed mean μ. From Theorem 1 we obtain the following proper scoring rule for this setting:

$$S(\mu, x) = \log \lambda - \lambda x.$$

It is simple to check using basic calculus that an agent maximizes the expected score by reporting $\lambda = 1/\hat{\mu}$, where here $\hat{\mu}$ is its estimate of the mean. In this case, it is unclear a priori which of the mean or rate parametrizations is most intuitive, but the maximum entropy approach easily allows for either. □

Beta Suppose the designer would like an estimate of the click-through rate of an online advertisement. The outcome space here is $\mathcal{X} = (0, 1)$. A typical prior over a probability like this is the beta distribution, which has two parameters $\alpha, \beta > 0$ and density function $p(x; \alpha, \beta) = x^{\alpha-1}(1 - x)^{\beta-1}/B(\alpha, \beta)$, where B is the beta function. By theorem 1 this leads to the proper scoring rule

$$S((\alpha, \beta), x) = (\alpha - 1) \log x + (\beta - 1) \log(1 - x) - \log B(\alpha, \beta).$$

Although this parametrization is standard for the beta, it is perhaps more intuitive to use the parametrization $\mu = \alpha/(\alpha + \beta) \in (0, 1)$ and $n = \alpha + \beta > 0$. Here μ is the mean of the distribution—corresponding to the agent's actual estimate of the click-through rate—and n is a measure of the agent's confidence in the estimate. □

A.3 Information Market

We next consider how to indirectly elicit μ by setting up a market of contingent claim securities. Under this approach, the elements of the index set $\mathcal{S}$ are interpreted not as statistics but as securities. The payoff from one share of security s when outcome x occurs is given by the mapping ϕ_s. Thus if the vector of shares held by the agent is $\theta \in \mathbf{R}^d$, where entry θ_s corresponds the number of shares of security s, then the payoff to the agent when x occurs is

evaluated by taking the inner product $\langle \theta, \phi(x) \rangle$. As a concrete example, recall our multinomial distribution example from the previous section, where $\phi_x(x') = 1$ if $x' = x$ and 0 otherwise. This means that security $x \in \mathcal{S}$ pays 1 dollar if outcome $x \in \mathcal{X}$ occurs, and nothing otherwise. (Such securities are known as Arrow-Debreu securities.) In our normal distribution example, the statistic for the mean had the mapping $\phi_1(x) = x$. Therefore a share of the corresponding security has a payoff that is linear in the outcome. (Such securities amount to futures contracts.)

To extract useful information from such a securities market, we set a pricing scheme and examine the number of shares the agent chooses to acquire. Let Ω be the set of all portfolios (i.e., vectors of shares) that the agent can feasibly hold; we will see how this domain is determined in an instant. Each security s has an associated price function $c_s : \Omega \to \mathbf{R}$, which gives the marginal price $c_s(\theta)$ of security s when the agent holds portfolio θ. (Note that the marginal price depends on the entire portfolio θ, not just the number of shares θ_s.) A risk-neutral agent will choose to acquire shares up to the point where, for each share, expected payoff equals marginal price. Formally, if the agent acquires portfolio θ, then for each $s \in \mathcal{S}$ we must have

$$\mathbf{E}_p[\phi_s(x)] = c_s(\theta). \tag{8}$$

In this way, by its choice of θ, the agent reveals that its belief is $\mu = c(\theta)$. There are several important properties that the price function c should have. To simplify the agent's portfolio acquisition process, it should not be the case that the total cost of acquiring θ depends on the order in which shares are bought. This means that there should exist a *cost function* $C : \Omega \to \mathbf{R}$ such that $c = \nabla C$. The gradient ∇C must be onto $\mathcal{M}$ to ensure that (8) always has a solution—this is the most important but technical condition to realize. To ensure that the solution is unique, it would also be convenient if the cost function were strictly convex.

Proposition 1 *The following cost function is monotone, convex, and onto* $\mathcal{M}$*:*

$$C(\theta) = \log \int_{x \in \mathcal{X}} \exp\left[\langle \theta, \phi(x) \rangle\right] \nu(dx). \tag{9}$$

Observe that in the context of our earlier multinomial example, cost function (9) is exactly the cost function for Hanson's logarithmic market scoring rule. Our approach here provides a generalization of this market scoring rule to markets with contingent claim securities with arbitrary payoffs, designed

to elicit specific properties of distributions, beyond just the probabilities of different outcomes.

Because an agent would never select a portfolio with infinite cost, the effective domain of C is $\Omega = \{\theta \,|\, C(\theta) < +\infty\}$. In the context of the multinomial distribution example, $\Omega = \mathbf{R}^d$, so that shares of each security can always be bought or sold short. In the context of the normal distribution example the effective domain is $\Omega = \{(\theta_1, \theta_2) \in \mathbf{R}^2 \mid \theta_2 > 0\}$ if we re-define the second statistic to be $\phi_2(x) = -x^2$; this means that the security has a negative payoff for each share, and consequently the agent must be compensated to acquire such shares. Evaluating (9), the cost function takes the form $C(\theta) = \frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(2\theta_2)$.

Now, in our previous elicitation approach that used a scoring rule, the process of computing the scoring rule also provided a complete distribution $p(x;\mu)$ over outcomes, which could be used to infer other properties beyond just the agent beliefs μ. In the current market-based approach, the agent's chosen portfolio θ can also form the basis of a distribution over outcomes. Consider the following distribution, represented as a density with respect to a base measure ν:

$$p(x;\theta) = \exp\left[\langle\theta, \phi(x)\rangle - C(\theta)\right]. \tag{10}$$

Meaning, the probability that a subset $X \subseteq \mathcal{X}$ of the outcomes occurs is $\int_{x\in X} p(x;\theta)\nu(dx)$. Observe that by definition (9), this probability density indeed integrates to 1 over $\mathcal{X}$, and it is clear that the density is non-negative as required.

In the statistical literature a distribution that takes the form (10) is known as an *exponential family*. The mapping ϕ is known as the *sufficient statistic*, θ is the *natural parameter*, and C is the log-partition or *cumulant* function.

A.4 Duality

There is a well-known duality between the maximum entropy approach and exponential families, which translates into a duality between the scoring rule and information market just developed. The duality implies that the approach leads to a *market scoring rule*, applicable to both thin and thick multi-agent settings.

It is known that a distribution is a maximum entropy distribution if and only if it is an exponential family []. To see this, let $\theta(\mu) \in \mathbf{R}^d$ be the

Lagrange multiplier corresponding to constraints (5) when solving the maximum entropy program given the agent report μ. (Our choice of notation is deliberately suggestive.) Let $A(\mu)$ be the Lagrange multiplier corresponding to (6). From the first order necessary conditions for optimality, we find that

$$p(x;\mu) = \exp\left[\langle\theta(\mu), \phi(x)\rangle - A(\mu) - 1\right],$$

so that the solution is indeed in exponential family form. Conversely, given a cumulant C from an exponential family and a parameter θ, we see that $p(x;\theta)$ defined according to (10) is the maximum entropy distribution under constraints $\mathbf{E}_p[\phi(x)] = \mu$ where $\mu = \nabla C(\theta)$. We obtain the following.

Proposition 2 *Let C be the cumulant (i.e, cost function) for the exponential family corresponding to ϕ, and let $\theta(\mu)$ be the optimal Lagrange multiplier for the mean constraints given an agent report of $\mu \in \mathcal{M}$. Then the following scoring rule is equivalent to (7):*

$$S(\mu, x) = \langle\theta(\mu), \phi(x)\rangle - C(\theta(\mu)). \tag{11}$$

The scoring rule (11) decomposes neatly into payoff and cost functions. The first term $\langle\theta(\mu), \phi(x)\rangle$ defines the agent's outcome-contingent reward, while the second term $C(\theta(\mu))$ is the cost of acquiring a portfolio $\theta(\mu)$. We see here the duality between the approach of directly reporting μ, or acquiring shares θ: an agent reporting beliefs μ under scoring rule (7) would choose to acquire shares $\theta(\mu)$ in the information market with cost function (9). Since (7) is a proper scoring rule, the information market with cost function (9) is based on a proper market scoring rule.

B Sindhu's Outline

In the notes on Maximum Entropy Market Maker, Lahaie constructs a scoring rule (and its sequential form: a market maker) based on maximum entropy distributions. Here is the problem: Suppose you want to elicit sufficient statistics from a trader (or group of traders: in this latter case you want to aggregate their information). Then if you find a probability distribution that is consistent with their belief of the expected sufficient statistics, you can construct a proper scoring rule (that is, the true distribution maximizes payoff). However, the solution to this consistent distribution is not unique. Hence, to pick amongst the distributions, you could choose the one that maximizes entropy (which is a notion of uncertainty) – intuitively, this assumes the least about the distribution generating the data. Now, it turns out that the distribution that maximizes entropy is an exponential family distribution. So, if it is known that the true probability is an exponential family distribution, eliciting just the statistics allows you to characterize the actual probability distribution.

In the prediction markets and exponential families write-up, we have shown that that the various parts of the exponential family correspond to securities, payoffs and cost in a prediction market. We have shown that a sequence of traders seeking to maximize profit will essentially be performing a MLE of the natural parameters of an exponential family, when the beliefs of the traders take the form of an exponential family distribution (it turns out that this last assumption is not necesary in light of the above result). Further, in a budget-limited setting, adversarial traders can cause limited damage and informative traders' budgets grow in expectation over time when trading in multiple markets. The budget-limited results have the following issues. First, the movement of the share vector is not completely justified. Second, as it stands currently, the growth in budget is shown in expectation over a belief distribution of the informative trader, and this distribution is assumed to be an exponential family (this may be fixable).

So broadly, the outline is as follows: The main goal is to construct a prediction market that aggregates information from traders to determine the sufficient statistics of some unknown distribution. First, we can show that in a thin market with a single trader this can be done by constructing a (proper) scoring rule using a maximum entropy approach. This approach can then be extended to define a market scoring rule for a prediction market to aggregate information from multiple traders. This follows from the du-

ality that exists between the maximum entropy approach and exponential families. We know that the updates to the mean parameters in a Bayesian setting are a convex combination of the empirical mean and the prior. We can then show that the share vectors of this prediction market follow the Bayesian updates of the mean parameters when each trader treats the current market state as the prior and his beliefs as the empirical mean. Since we can interpret the share vector as the natural parameters of the exponential family, the updates essentially happen in a dual space. This process of mapping to a dual space to perform an update is reminiscent of the online mirror descent algorithm; what exactly is the mapping? Also, we would like to analyze the prediction market so defined, when the traders are budget limited. Is the step size of the algorithm somehow related to the budgets of the traders?

C Conjugate Priors

In Sindhu's notes it's assumed that the agent has a belief $\mathbf{q}$, and moves the market state from initial state $\mathbf{q}_{init}$ to a convex combination $\tilde{\mathbf{q}} = \lambda\mathbf{q} + (1 - \lambda)\mathbf{q}_{init}$. This note tries to justify this based on Bayesian updating, rather than budget limitations. (But it doesn't quite succeed.)

Let $p(x;\theta)$ denote a probability density drawn from an exponential family with sufficient statistic $\phi : \mathcal{X} \to \mathbf{R}^d$, where θ is the natural parameter:

$$p(x;\theta) = \exp\left[\langle\theta, \phi(x)\rangle\right],$$

and

$$g(\theta) = \log \int_{\mathcal{X}} \exp\langle\phi(x), \theta\rangle dx.$$

Recall that $\nabla g(\theta) = \mathbf{E}[\phi(x)]$ and $\nabla^2 g(\theta) = \mathrm{Var}[\phi(x)]$. The family of conjugate priors is also an exponential family and takes the form

$$p(\theta; n, \nu) = \exp\left[\langle n\nu, \theta\rangle - ng(\theta) - h(\nu, n)\right].$$

Here the feature map is $\psi(\theta) = (\theta, -g(\theta))$, the natural parameter is $(n\nu, n)$ where $n \in \mathbf{R}$ and $\nu \in \mathbf{R}^d$. The normalizer $h(\nu, n)$ is convex in $(n\nu, n)$. It is helpful to think of the prior as being based on a 'phantom' sample of size n and mean ν. The justification for this is that

$$\mathbf{E}_\theta\left[\mathbf{E}_x\left[\phi(x)|\theta\right]\right] = \mathbf{E}_\theta\left[\nabla g(\theta)\right] = \nu.$$

(The proof is straightforward but not obvious. I have a reference for this but it's impossible to read—it's a mathematical statistics paper.)

Suppose we draw a sample $X = (x_1, \ldots, x_m)$ of size m, and denote the empirical mean by $\mu[X] = \sum_{i=1}^m \phi(x_i)$. The posterior distribution is then

$$p(\theta|X) \propto p(X|\theta)p(\theta|n,\nu) \propto \exp\left[\langle\mu[X] + n\nu, \theta\rangle - (m+n)g(\theta)\right],$$

and so the posterior mean is

$$\frac{m\mu[X] + n\nu}{m+n}. \tag{12}$$

Thus the posterior mean is a convex combination of the prior and posterior means, and their relative weights depend on the phantom and empirical sample sizes.

In the context of prediction markets, one could imagine an agent who uses the current market estimate as its prior, and draws a empirical sample of size m. Its belief then takes the form (12), where n depends on the importance the agent places on the market estimate (perhaps based on how long the market has been running). However, note that the *mean parameter* becomes a convex combination of market state and empirical belief, and this does not translate to the *natural parameter*, which is what we would have liked. The new natural parameter is

$$\nabla g^{-1}\left(\frac{m\mu[X]+n\nu}{m+n}\right) = \nabla g^{*}\left(\frac{m\mu[X]+n\nu}{m+n}\right).$$

where g^* is the convex conjugate of g.

D Maximum Likelihood Scoring Rules

D.1 Preliminaries

We consider a state space $\mathcal{X}$ which may be discrete or continuous, and a measure ν over sets of states. We will typically take the state space to be $\mathbf{R}$ or $\mathbf{R}^d$ together with the Lebesgue measure (perhaps restricted to some subset of states), or some finite set together with the counting measure.[2] A *probability density* is a ν-measurable function $p : \mathcal{X} \to \mathbf{R}_+$ that integrates to 1 over $\mathcal{X}$.

Let $\mathcal{P}$ denote the set of all probability densities over the given measure space. We are interested in eliciting information about some unknown density $p \in \mathcal{P}$ which corresponds to the beliefs of an expert. The expert may be familiar with p in its entirety, or only with certain properties of p. Given a set of densities $\mathcal{D} \subseteq \mathcal{P}$, a *property* is a real (possibly vector-valued) function $\Gamma : \mathcal{D} \to \mathbf{R}^k$. With a slight abuse of terminology we call elements of its image $\Theta = \Gamma(\mathcal{D})$ *properties* and we say that p has property θ when $\Gamma(p) = \theta$. If Γ is one-to-one, we instead call Γ a *parametrization* and refer to the elements of its image Θ as *parameters*.

To be concrete, our outcome space could for instance be $\mathbf{R}$ with ν the Lebesgue measure restricted to $[0, +\infty)$. The set $\mathcal{P}$ then consists of all probability densities with support contained in the non-negative reals, and $\mathcal{D} \subset \mathcal{P}$ might consist of the densities of exponential distributions. An example of a property over $\mathcal{P}$ is the mapping from a density to its mean.[3] When restricted to the exponential densities $\mathcal{D}$, this mapping is a parametrization. In much of this work, as in this example, the purpose of the measure ν is to implicitly encode bounds on the support of possible densities. The reason we use this approach, rather than explicitly stating bounds, is that it allows for clearer connections with maximum entropy estimation later on.

Given the set of relevant densities $\mathcal{D} \subseteq \mathcal{P}$, which contains the expert's belief p, and a property Γ over this set, the expert is asked to report $\theta = \Gamma(p)$. The expert is rewarded according to a *scoring rule* $S : \Theta \times \mathcal{X} \to \mathbf{R} \cup \{-\infty\}$ which pays it $S(\hat{\theta}, x)$ according to how well its report $\hat{\theta} \in \Theta$ agrees with the

[2] We will not need to make reference to the underlying σ-algebra, but to complete the specification of the measure space we take it to be the collection of Borel sets in the case of $\mathbf{R}$ and $\mathbf{R}^d$, and the power set in the case of a discrete set.

[3] The property may not be defined for all $p \in \mathcal{P}$. In this case we implicitly restrict it to the subset of $\mathcal{P}$ for which it is defined whenever its domain $\mathcal{D}$ is not made explicit.

eventual outcome $x \in \mathcal{X}$. The idea is to design the scoring rule so that the expert is incentivized to report its true belief, setting $\hat{\theta} = \theta$.

Definition 1 *Given a property $\Gamma : \mathcal{D} \to \mathbf{R}^k$, a scoring rule S is* proper *over domain $\mathcal{D}$ if for each $\theta \in \Theta$ and $p \in \mathcal{D}$ with property θ, we have*

$$\mathbf{E}_p[S(\theta, x)] > \mathbf{E}_p[S(\hat{\theta}, x)] \tag{13}$$

for all $\hat{\theta} \neq \theta$. If the domain is $\mathcal{D} = \mathcal{P}$, the set of all possible densities, then we say the scoring rule is strongly proper.

- Note that the inequality is strict, in the literature these rules are called *strictly* proper, but we only consider strictly proper rules here so don't make this qualification.
- Lambert et al.'s rules are actually strongly proper, however we consider scoring rules over vector valued properties. Lambert et al only develop these by adding together different scoring rules for each property. We go further in our constructions and develop rules that are not necessarily separable.
- It is useful to consider rules that are not strongly proper because we might have constructions for these for properties where no strongly proper rule is known.
- It is important to understand the informational requirements placed on the expert. Under a proper scoring rule, the expert does not need to be aware of the entire density p to be incentivized to report its property θ, since θ is an optimal report no matter what the density might be. The agent only needs to agree that the density indeed lies in $\mathcal{D}$.
- For instance, if we are considering densities with support on $[0, +\infty)$, and $\mathcal{D}$ is the set of lognormal densities, then a proper scoring rule over $\mathcal{D}$ for the mean does not require the agent to know anything beyond the mean (e.g., it does not need to know the variance).
- A strongly proper scoring rule goes further; with such a rule we do not require the expert to believe the density is drawn from some subset $\mathcal{D}$, only that it agrees with the support implied by the base measure ν. Therefore, a strongly proper scoring rule for the mean here would elicit the mean no matter what kind of density over the positive reals might apply (e.g., exponential, Pareto, lognormal).

- In practical scenarios this seems like an important extension, because while it might be clear which states are possible, it might be difficult to ensure the expert agrees that the probability density for states falls within some restricted family $\mathcal{D}$.

D.2 Maximum Likelihood

Theorem 2 *The logarithmic scoring rule defined by*

$$S(\theta, x) = \log p(x; \theta)$$

is weakly proper for $\mathcal{F}$ over Θ if and only if the maximum likelihood estimate for θ is consistent. Furthermore, if $\mathcal{M}$ is an alternative parameter space such that there exists a bijection $t : \mathcal{M} \to \Theta$, then $\bar{S}(\mu, x) = S(t(\mu), x)$ is weakly proper for $\mathcal{F}$ over $\mathcal{M}$.

Example 1. Suppose X is supported on $[0, +\infty)$ and follows an exponential distribution with density $f(x; \lambda) = \lambda e^{-\lambda x}$ parametrized by a rate $\lambda > 0$. The mean μ is related to the rate via the one-to-one mapping $\mu = 1/\lambda$, so the density can alternatively be parametrized by the mean. According to Theorem 2 the following is a weakly proper scoring rule for the mean of an exponential distribution:

$$S(\mu, x) = -\frac{x}{\mu} - \log \mu. \tag{14}$$

To verify this, suppose the agent believes X follows an exponential distribution with rate λ_0, equivalently mean $\mu_0 = 1/\lambda_0$. By the one-to-one relation between rate and mean, choosing μ to maximize the expected score (14) is equivalent to choosing λ to maximize $\mathbf{E} - \lambda x + \log \lambda = -\lambda \mu_0 + \log \lambda$. The latter is strictly concave in λ, and using basic calculus its unique global maximum is $\lambda^* = 1/\mu_0$, which therefore leads the agent to report $\mu = \mu_0 = 1/\lambda^*$ under scoring rule (14).

Example 2. Suppose that X is supported on $[1, +\infty)$ and follows a Pareto distribution with density $f(x; \alpha) = \alpha / x^{\alpha+1}$ parametrized by an index $\alpha > 0$. The mean μ is related to the index via the one-to-one mapping $\mu = \frac{\alpha}{\alpha - 1}$, so the density can alternatively be parametrized by the mean. Theorem 2 gives the following weakly proper scoring rule for the mean of a Pareto distribution with support on $[1, +\infty)$:

$$S(\mu, x) = \log \frac{\mu}{\mu - 1} - \left(\frac{\mu}{\mu - 1} + 1 \right) \log x. \tag{15}$$

To verify this, suppose the agent believes X follows a Pareto distribution with index α_0, equivalently mean $\mu_0 = \frac{\alpha_0}{\alpha_0 - 1}$. By the one-to-one relation between the index and mean, choosing μ to maximize the expected score (15) is equivalent to choosing α to maximize $\mathbf{E}\log\alpha - (\alpha+1)\log x = \log\alpha - (\alpha+1)\mathbf{E}\log x$. The latter is strictly concave in α, and using basic calculus its unique global maximum is $\alpha^* = 1/\mathbf{E}\log X = \alpha_0$, where the last equality follows from the fact that $\log X$ has an exponential distribution with rate α when X is Pareto with rate α. Therefore (15) leads the agent to report $\mu = \mu_0 = \frac{\alpha^*}{\alpha^* - 1}$.

- Explain that the scoring rule from Example 2 does not elicit the mean when the agent's distribution is exponential (with change of variable so that it lies in $[1, +\infty)$.
- Observe that scoring rule from Example does elicit the mean when the agent's distribution is exponential, Pareto, or in fact any other distribution with support on $[0, +\infty)$ such as the lognormal.
- This occurs because the scoring rule is linear in x, so its expectation is the same for any distribution with mean μ.
- Another feature to note is that under an appropriate parametrization (rate instead of mean), the scoring rule (1) is convex in the parameter being elicited.
- These two observations taken together lay the groundwork for the characterization of strongly proper scoring rules for distribution statistics.
- Before we proceed should note that the arguments behind Theorem 2 apply not just to maximum likelihood but to any *m-estimator*, which are notions from robust statistics. Explain what an m-estimator looks like and how there are general consistency results for those as well.
- The point of these alternative estimators is to be more robust to outliers in the sample, and so these kinds of estimators could translate into scoring rule more robust to inaccuracies in the agent's assessment of the underlying probability distribution or its parameters. (Just a hunch.)