Nonparametric Scoring Rules

Erik Zawadzki
Carnegie Mellon University
Pittsburgh, PA 15213
epz@cs.cmu.edu

Sébastien Lahaie
Microsoft Research
New York, NY 10011
slahaie@microsoft.com

Abstract

A scoring rule is a device for eliciting and assessing probabilistic forecasts from an agent. When dealing with continuous outcome spaces, and absent any prior insights into the structure of the agent's beliefs, the rule should allow for a flexible reporting interface that can accurately represent complicated, multi-modal distributions. In this paper, we provide such a scoring rule based on a nonparametric approach of eliciting a set of samples from the agent and efficiently evaluating the score using kernel methods. We prove that sampled reports of increasing size converge rapidly to the true score, and that sampled reports are approximately optimal. We also demonstrate a connection between the scoring rule and the maximum mean discrepancy divergence. Experimental results are provided that confirm rapid convergence and that the expected score correlates well with standard notions of divergence, both important considerations for ensuring that agents are incentivized to report accurate information.

Introduction

A scoring rule is a device for eliciting and assessing probabilistic forecasts from an agent. The earliest scoring rules were designed for probability distributions over a finite, exhaustive set of outcomes, or expectations of random variables (Savage 1971), and the theory has since been extended to continuous outcome spaces and general statistics (Lambert, Pennock, and Shoham 2008; Matheson and Winkler 1976). The broader field of elicitation now spans several disciplines, and includes the study of user interfaces and the effectiveness of eliciting various probability summaries from experts; for an extensive review see (O'Hagan et al. 2006). Encouraging selfish agents to accurately report their beliefs is a fundamental problem in multi-agent systems, and a key building block in systems that aggregate beliefs in order to form a consensus view of uncertain events.

In this paper we are concerned with eliciting a *density function* over a continuous outcome space. Density forecasting finds application in surveys of experts on macroeconomic indicators, and in the estimation of asset and portfolio returns in finance, among other areas (Boero, Smith, and Wallis 2011; Tay and Wallis 2002). In practice, the main difficulty stems from the simple fact that, generically, the agent cannot specify a density function in full precision with a finite-length report. Much of the literature therefore focuses on eliciting summaries such as mean and variance, median, or quantiles, and the challenging case of correlations in the case of multivariate data (Clemen, Fischer, and Winkler 2000; Garthwaite, Kadane, and O'Hagan 2005). Another common approach is to bin the domain and effectively elicit a histogram (Goldstein, Johnson, and Sharpe 2008).

These interfaces can all be construed as 'parametric' in the choice of summaries they choose to elicit. We propose instead a 'nonparametric' approach. Our central contribution is the design and analysis of a proper scoring rule that scores an agent's forecast communicated in terms of a *representative sample* from its private density estimate. We argue that this kind of scoring rule provides a versatile, modular method for assessing forecasts that can complement the variety of interfaces used for eliciting summaries of densities, and that can also serve as an interface in its own right.

Sample-based reports are extremely flexible, and this allows the forecast elicitor to offer the agent a large menu of different interfaces that all hook into the same underlying score. Our only requirement is that each of these interfaces be associated with a sampling mechanism. For instance, histograms have a straightforward sampling mechanism associated with them. In a typical elicitation experiment, once summaries are obtained, a common strategy is to then fit a parametric density function to the summaries, which immediately lends itself to sampling (Kadane et al. 1980; O'Hagan 1998). Another important use case is if the agent's belief is a posterior predictive distribution approximated via Markov chain Monte Carlo (MCMC). In this case the agent can directly submit its MCMC samples to our scoring mechanism.

Our sample-based scoring rule is derived as an empirical estimate of the *kernel score* (Eaton 1982), which in fact defines a family of scoring rules that depend on the choice of a kernel function. A kernel function can have several interpretations but we view it mainly as a density centered at a point in outcome space (e.g., a Gaussian). We give conditions on the kernel that make the score strictly proper, and relate it to a measure of divergence between probability measures known as maximum mean discrepancy (Gretton et al.

2006). In terms of the direct sampling interface, the choice of representative sample to report is in principle a nontrivial optimization. Our main incentive results show that simple random sampling is an approximately optimal strategy. We also show how to calibrate a per-sample reporting fee to control the size of the agent's reported set of samples.

We conducted experiments to investigate the empirical properties of the sample-based kernel score. Two experiments are described. The first investigates how rapidly the score converges as the reported set of samples increases in size. The experiment confirms our theoretical result that convergence is rapid and reliable. Our second experiment investigates the incentive properties of the sample-based method. We found that the kernel-based score does a better job of awarding high scores to accurate predictions than an alternative score based on histograms. Moreover, we found that a subsampled version of the sample-based kernel score achieved nearly the same performance as the original variant, but had a better asymptotic run time.

Scoring rules can serve as a basis for a variety of multi-agent elicitation and aggregation procedures. In the area of multi-agent systems, scoring rules have been applied to demand prediction for energy resource planning (Robu et al. 2012; Rose, Rogers, and Gerding 2012), and in the context of distributed information systems (Papakonstantinou et al. 2011). In the final section of the paper we discuss how our scoring rule could form the basis of aggregation procedures such as market scoring rules, wagering and escrow mechanisms, and cost-function based prediction markets.

Background

We consider probabilistic forecasts over a sample space Ω. Formally, let Ω be a compact metric space with an associated Borel σ-algebra of events. A *probabilistic forecast* corresponds to a probability measure on Ω. We restrict our attention to forecasts that are absolutely continuous with respect to a σ-finite reference measure ν so that they have probability densities. Throughout the paper it is understood that densities and their integrals (including expectations) are taken with respect to ν.

As an example the sample space could be a box in $\mathbb{R}^n$ with the standard Euclidean metric and Lebesgue as the reference measure. Our setting and results also apply to finite (perhaps exponentially large) outcome spaces with the counting measure, but we will focus on continuous sample spaces in our motivation and experiments.

Forecasts can be evaluated by means of a scoring rule, which intuitively assesses how well a forecast agrees with the eventual outcome. An agent seeks to report a forecast that maximizes its expected score. Let $\mathcal{P}$ be a class of probability measures on Ω. Formally, a *scoring rule* is a real-valued function $S : \mathcal{P} \times \Omega \to \mathbb{R}$ such that for all $P, Q \in \mathcal{P}$, the score $S(P, \cdot)$ is integrable with respect to Q.

Following several authors we overload the scoring rule notation and write

$$S(P, Q) = \mathbb{E}_{\omega \sim Q} S(P, \omega)$$

to denote the expected score of probabilistic forecast P under belief Q. A scoring rule S is *proper* relative to the class of probability measures $\mathcal{P}$ if for all $P, Q \in \mathcal{P}$ we have $S(Q, Q) \geq S(P, Q)$. The score is *strictly proper* if the inequality is strict whenever $P \neq Q$. Proper scoring rules are important from an elicitation standpoint because they incentivize an agent to truthfully report its subjective distribution. Two prominent examples of scoring rules are the quadratic score and the logarithmic score.

Def. 1 (Quadratic score)**.** *For any $P \in \mathcal{P}$ with density p its* quadratic score *is*

$$S_Q(P, \omega) = p(\omega) - \tfrac{1}{2}\|p\|_2^2. \tag{1}$$

Def. 2 (Logarithmic score)**.** *For any $P \in \mathcal{P}$ with density p its* logarithmic score *is*

$$S_L(P, \omega) = \log p(\omega). \tag{2}$$

Both of these scores are proper scores for restricted classes of probability measures (Gneiting and Raftery 2007). The quadratic score is strictly proper with respect to measures with square-integrable densities, and the log score is strictly proper with respect to integrable densities (recall, this means integrable with respect to ν).

Proper scores are closely related to statistical notions of *divergence* (Dawid 1998). Given a strictly proper scoring rule S, define

$$d(P, Q) = S(Q, Q) - S(P, Q), \tag{3}$$

which yields a non-negative, possibly asymmetric function such that $d(P, Q) = 0$ if and only if $P = Q$. This can be viewed as a 'distance' between measures. In particular, the quadratic score induces the squared $\mathcal{L}_2$ divergence, and the log score induces the Kullback-Leibler divergence.

Our work is motivated by the fact that in practice an agent may not be able to communicate its belief with full accuracy over a continuous outcome space, and will necessarily have to resort to some finite-length report. We consider reports that take the form of representative samples $X = \{x_1, \ldots, x_m\}$. One approach to scoring such a report would be for the center to first infer a probability density from the sample, and then apply a proper score such as (1) or (2) to the density. For a generic inference procedure the difficulty is in understanding how this incentivizes the agent to form its reported sample, both in terms of its size and the choice of data points. For certain scoring rules, this approach may also be more computationally intensive than necessary—for instance, the log score (2) only requires a probabilistic inference at the realized outcome, not over the entire sample space.

We will see that our nonparametric approach addresses both the incentive and computation concerns, but it can nevertheless be construed as an (implicit) application of an inference procedure followed by scoring. The inference procedure is kernel density estimation, a standard workhorse of nonparametric statistics. Let $K : \Omega \times \Omega \to \mathbb{R}$ be a symmetric *kernel* function over the sample space such that $K(\omega, \cdot)$ integrates to 1 for all $\omega \in \Omega$. Given a sample X, the *kernel density estimate* (KDE) is defined as the average kernel value for the sample points:

$$p_{K,X}(\omega) = \frac{1}{m} \sum_{i=1}^{m} K(x_i, \omega). \tag{4}$$

The KDE can be seen as 'smoothing' the empirical measure of the sample

$$\hat{P}_X(E) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}[x_i \in E] \quad (5)$$

to form a smooth density estimate; here $E \subset \Omega$ is an event and $\mathbb{I}$ is the 0-1 truth predicate. See (Wand and Jones 1995) for a standard reference on KDE.

We specifically make use of *positive definite* (p.d.) kernels, drawing on work in machine learning that uses such kernels for nonparametric comparison of probability distributions (Smola et al. 2007; Sriperumbudur et al. 2010). A kernel is p.d. if, given any sample $\{x_1, \ldots, x_m\}$, the matrix formed from the entries $K(x_i, x_j)$ is positive definite. Associated with a p.d. kernel K is a reproducing kernel Hilbert space (RKHS) of functions over Ω, denoted $\mathcal{H}_K$. The RKHS has a reproducing property in the sense that $K(x, \cdot) \in \mathcal{H}_K$ for each $x \in \Omega$ and the inner product satisfies $\langle f, K(x, \cdot)\rangle = f(x)$ for each $f \in \mathcal{H}_K$. It is common to view $\phi(x) \equiv K(x, \cdot)$ as a 'feature map' sending x to a (possibly infinite-dimensional) vector of features $\phi(x)$. We refer to (Schölkopf and Smola 2002) for a comprehensive treatment of kernel methods.

Theory

In this section we develop the ideas leading to our sample-based scoring rule and examine its incentive properties. A p.d. kernel can be used to directly define a proper scoring rule called the *kernel score* (Eaton 1982; Gneiting and Raftery 2007).

Def. 3 (Kernel score)**.** *For p.d. kernel K on $\Omega \times \Omega$ and any probability measure $P \in \mathcal{P}$, the* kernel score *is*

$$S_K(P, \omega) = \mathbb{E}_{x\sim P} K(x, \omega) - \tfrac{1}{2}\mathbb{E}_{x,x'\sim P} K(x, x'). \quad (6)$$

The kernel score can be represented concisely via inner-products in the RKHS $\mathcal{H}_K$ associated with K using the mean map $\mu : \mathcal{P} \to \mathcal{H}_K$. This function μ maps probability measures P to *expected evaluation functions*—these are elements of $\mathcal{H}_K$ such that, for any f in $\mathcal{H}_K$, $\mathbb{E}_P f(x) = \langle \mu[P], f\rangle$. The mean map exists for P if $\mathbb{E}_{x,x'\sim P} K(x, x') < \infty$ (Borgwardt et al. 2006). With this notation the kernel score can be rewritten as

$$S_K(P, \omega) = \langle \mu[P], \phi(\omega)\rangle - \tfrac{1}{2}\langle \mu[P], \mu[P]\rangle . \quad (7)$$

Like other scoring rules this gives rise to a divergence: in this case a particular squared *maximum mean discrepancy* (Gretton et al. 2006).

Def. 4. *Let $\mathcal{F}$ be a class of functions $f : \Omega \to \mathbb{R}$, and let P, Q be probability measures on Ω. The* maximum mean discrepancy (MMD) *is*

$$\mathrm{MMD}[\mathcal{F}, P, Q] = \sup_{f\in\mathcal{F}} \mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(y)].$$

This definition of distance makes intuitive sense: $P = Q$ if and only if $\mathbb{E}_P f(x) = \mathbb{E}_Q f(x)$ for all $f \in \mathcal{F}$. The MMD finds a 'witness function' f that highlights the difference between P and Q.

The connection between the kernel score and the MMD becomes apparent from (7).

Lem. 1. *The divergence associated with the kernel score is the squared MMD based on $\mathcal{B}$, the unit ball in $\mathcal{H}_K$.*

Proof. Let $\mathcal{B} = \{f \,:\, \|f\|_{\mathcal{H}_K} \le 1\}$.

$$\begin{aligned} d_K(P, Q) &= S_K(P, P) - S_K(Q, P) \\ &= \tfrac{1}{2}\langle \mu[P], \mu[P]\rangle - \langle \mu[P], \mu[Q]\rangle + \tfrac{1}{2}\langle \mu[Q], \mu[Q]\rangle \\ &= \tfrac{1}{2}\|\mu[P] - \mu[Q]\|^2_{\mathcal{H}_K} \\ &= \tfrac{1}{2}\mathrm{MMD}^2[\mathcal{B}, P, Q]. \end{aligned}$$

The final line follows from (Borgwardt et al. 2006). □

We see from (3) that a proper scoring rule incentivizes the agent to minimize the associated divergence to its subjective distribution. If the reporting interface does not allow the agent to exactly describe its belief (e.g., an interface based on summaries or intervals), then Lemma 1 implies that the agent will report the closest permissible density according to MMD. This shows that the kernel score is *effective* in the sense of Friedman (1983): the expected score is a strictly decreasing function of the distance between reported and subjective distributions, in the RKHS metric.

Lemma 1 also allows us extend Theorem 4 from Gneiting and Raftery (2007) to identify conditions under which the kernel scoring rule is *strictly* proper. A kernel K is *universal* if its RKHS is dense in $\mathcal{C}(\Omega)$, the set of continuous real functions on Ω (Steinwart 2002). Gaussian and Laplace radial basis functions are both universal, for instance.

Prop. 1. *The kernel score associated with p.d. K is strictly proper for the class of probability measures $\mathcal{P}_0$ where $\mathbb{E}_{x,x'\sim P} K(x, x') < \infty$ if K is universal.*

The key property that makes the kernel score appealing is that it can be estimated using m samples $X = \{x_1, \ldots, x_m\}$ drawn from P. The empirical estimate of (6) leads to our proposed scoring rule for a reported sample from the agent.

Def. 5 (Sample score)**.** *For a p.d. kernel K on $\Omega \times \Omega$ and any finite sample $X \in \Omega^m$ $(m > 0)$, the* sample score *is*

$$\hat{S}_K(X, \omega) = \frac{1}{m}\sum_{i=1}^{m} K(x_i, \omega) - \frac{1}{2m^2}\sum_{i,j=1}^{m} K(x_i, x_j). \quad (8)$$

Using properties of RKHS, one can confirm that this estimator is exactly the quadratic score applied to (4), the KDE base on the samples. Note also that $\hat{S}_K(X, \omega) = S_K(\hat{P}_X, \omega)$, meaning that the kernel score is implicitly doing KDE, but density estimation is not needed as an intermediate step in order to evaluate the score.

The main properties of the sample score stem from the fact that the estimate (8) converges quickly to the kernel score.

Lem. 2. *Let K be a p.d. kernel with $K(\omega, \omega) \le \kappa$ for all $\omega \in \Omega$. For any probability measure $P \in \mathcal{P}$, unit ball $\mathcal{B} \subset \mathcal{H}_K$ and any set $X = \{x_1, \ldots, x_m\}$ drawn i.i.d. from P,*

$$\mathrm{MMD}^2_b[\mathcal{B}, P, X] = \sup_{f\in\mathcal{B}} \left| \mathbb{E}_P[f(x)] - \frac{1}{m}\sum_{i=1}^{m} f(x_i) \right|^2 > \epsilon$$

with probability at most $\delta = 2\exp\left(-m\epsilon/2\kappa\right)$.

Proof sketch. The kernel is bounded; changing a single sample has a bounded effect on the MMD. Therefore, we apply McDiarmid's inequality to obtain the stated bound. □

Thm. 1. *For any probability measure $P \in \mathcal{P}$, and set of m samples X, $S_K(P,P) - \hat{S}_K(X,P) > \epsilon$ with probability at most $\delta = \exp(-m\epsilon/2\kappa)$.*

Proof. Let $\phi(x) = K(x,\cdot) \in \mathcal{H}_K$ be the point evaluation function. The mean map embedding for $\hat{P}_X$ is $\frac{1}{m}\sum_i^m \phi(x_i)$.

$$|S_K(P,P) - \hat{S}_K(X,P)| = \frac{1}{2}\left\|\mu[p] - \frac{1}{m}\sum_i^m \phi(x_i)\right\|^2_{\mathcal{H}_K}$$
$$= \tfrac{1}{2}\mathrm{MMD}_b^2[\mathcal{B}, P, X].$$

Result follows from Lem. 2. □

This result is useful for reasoning about the agent's incentives. If sampling is free—there are no costs associated with it—then a rational strategy for an agent is to sample from its belief. This is a consequence of the above convergence and the fact that the kernel score is strictly proper. Unfortunately, for general P there will always be an incentive to add an additional sample to the report. We would like to encourage reports of manageable (and certainly finite) size.

Suppose that there is a cost $\tau > 0$ for reporting a data point, leading to the score

$$\tilde{S}_K(X,P) = \hat{S}_K(X,P) - \tau|X|.$$

The cost may stem from exogenous introspective costs associated with sampling, an endogenous per-sample fee levied by the center, or a combination thereof. The addition of τ distorts the scoring rule and muddies the incentive story; there might now be some alternative report that does better than straightforward sampling.

If τ is small relative to $S(P,P)$ then we expect that the principal effect will be as a sparsifying 'regularizer' that induces the agent to submit a finite-sample report. We justify this claim in the next two results. We first show that Thm. 1 bounds the marginal benefit of reporting $m+1$ samples over m samples *ex-ante*—prior to the revelation of either agent belief or particular sample points.

Thm. 2. *If the cost per sample is τ, then with probability at least $1-\delta$ a rational agent will not report more than $(2\kappa/\tau)\log(2/\delta)$ samples.*

Proof. Let $n > m$, $X \sim P^m$, and $Y \sim P^n$. The larger report Y is preferred if

$$\tilde{S}(X,P) < \tilde{S}(Y,P)$$
$$\Rightarrow\ \tau < \hat{S}(Y,P) - \hat{S}(X,P) \le S(P,P) - \hat{S}(X,P)$$

From Thm. 1, $|S(P,P) - \hat{S}(X,P)|$ is more extreme than τ with probability less than δ if $m \ge 2\kappa/\tau\log(2/\delta)$. Therefore, with probability at least $1-\delta$ the difference is less than or equal to τ, so $\tilde{S}(X,P) \ge \tilde{S}(Y,P)$. □

Since this bound does not use any prior information about P, the result can be used by the center prior to interaction with the agent to roughly calibrate transaction costs and obtain appropriately sized reports.

Finally, we justify our focus on sampling by the fact that it is a good strategy when costs are small. This follows from Prop. 1 when $\tau = 0$, and from Thm. 1 when $\tau > 0$.

Cor. 1. *Let $X \sim P^m$ where $m > 0$. Then*

$$\sup_{Y \in \Omega^m} \tilde{S}_K(Y,P) - \tilde{S}_K(X,P) \le \epsilon$$

with probability at least $1 - 2\exp(-\epsilon m/2\kappa)$.

For large reports the above corollary states that an optimal choice of Y will not be much better than reporting $|Y|$ i.i.d. samples, while potentially being much more difficult to figure out. On the other hand, the bound will be weak if there is a small report Y that attains a sample score close to $S_K(P,P)$. However, in this regime it becomes easier for the agent to find an optimal representative sample, and we are not as concerned with providing an alternative approximately optimal strategy. Recall that in any case an optimal report will minimize the MMD to the agent's true belief.

Experiments

Our experiments fall into two main categories: characterizing kernel score convergence and establishing the quality of the kernel score by comparing it to $\mathcal{L}_2$ divergence. These experiments were repeated in $D = 1, 2, 3$ dimensions to indicated scaling behavior.

We used the Gaussian radial basis function (RBF) kernel for the kernel score. Our benchmark was to uniformly bin samples in $[-1,1]^D$, and use the standard quadratic score on the resulting piecewise constant function. We also evaluated a 'stochastic' version of the kernel score that subsampled $5m$ of the m^2 elements from the summation in (8). This is an asymptotically faster version of the kernel score that does not have a $\Theta(m^2)$ dependence on the number of samples.

Our experiments require a way to generate a distribution P for the ground truth of an event and a distribution Q for the agent's beliefs. We generate P and Q independently from the same distribution of distributions. We used a mixture of between five and ten isotropic Laplace densities with bandwidth 0.05. Centers for the individual Laplace densities were located in $[-1,1]^D$ uniformly at random, and had weights drawn uniformly from $[0,1]$. The Laplace density was used because the Gaussian RBF kernel is not a 'perfect fit' for it—the Gaussian is too smooth and its tails are too light. A synthetic distribution was used in part to remain agnostic to particular applications, and in part because current surveys of density forecasts appear to be limited to one-dimensional binned reports (Boero, Smith, and Wallis 2011, for example). The nonparametric approach advocated in this paper could allow experts to submit more sophisticated forecasts.

Our primary interest is how scores motivate an agent to reveal information. Therefore we are agnostic to affine transformations of the score. The score $S(P,\omega)$ is equivalent to $aS(P,\omega) + b$ for any $a > 0$ and b in terms of incentives. Because of this, we avoided evaluating experiments using metrics that distinguished between equivalent scores.

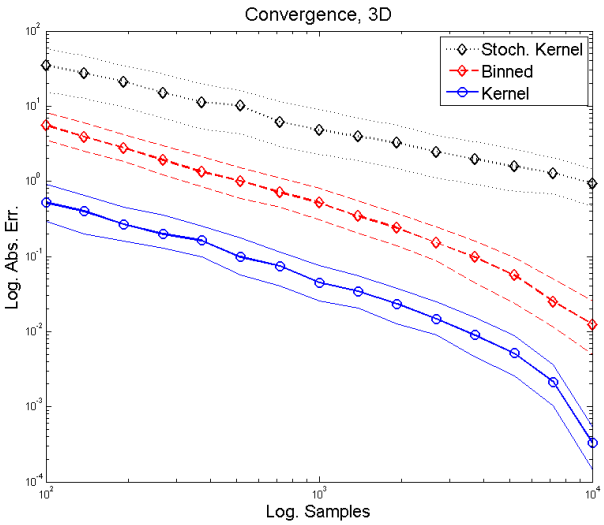

Figure 1: Convergence in median log absolute error, as a function of sample size in 3D. Thin lines indicate the 25%-75% interquartile range. Kernel bandwidth was 0.25, binning used 4^3 bins.

All experiments were coded in MATLAB, and run on a 3.40GHz i5-3570K with 8GB RAM.

Convergence

In this set of experiments we measured the convergence of kernel and binned scores $\hat{S}(X, P)$ for $X \sim P^m$ to their true value as the number of samples $m = |X|$ increased. We incrementally generated $N = 100$ different Laplace mixtures $(P_1, \ldots, P_N)$. For each P_i, we ran $M = 250$ test instances. Each instance $j \leq M$ consisted of generating two $m = 10,000$ sample sets $X_{i,j}, X'_{i,j} \sim P_i^m$. For each instance the agent reported a prefix of k elements of $X_{i,j}$ and was evaluated against the set $X'_{i,j}$. We tracked the absolute error between the scores for the prefixes and the scores for reporting the full sample:

$$\text{AbsError}(i,j) = \left| \hat{S}(X_{i,j}, X'_{i,j}) - \hat{S}(X_{i,j}^{(1:k)}, X'_{i,j}) \right|.$$

We used absolute error rather than relative error because of our indifference to affine transformation.

Both scoring rules converged rapidly and reliably. Figure 1 shows convergence for one of the N distributions, and is representative of the convergence that we saw in the other distributions. Again, due to our indifference to affine transformation, the y-intercept is unimportant.[1] We are primarily concerned with the slope of the convergence.

For most instances both scores exhibited similar linear rates of convergence. Least-square fitting in log-space showed that binned scoring had a slope of -1.18, while kernel scoring had a slope of -1.21. The linear convergence for the kernel series was anticipated by Thm. 1. Figure 1 also shows that the stochastic variant of the kernel score also converged reliably, but at a noticeably slower rate; least-square fitting found a slope of -0.78.

[1] In fact, scores were scaled to tease apart the series and make Figure 1 visually clearer.

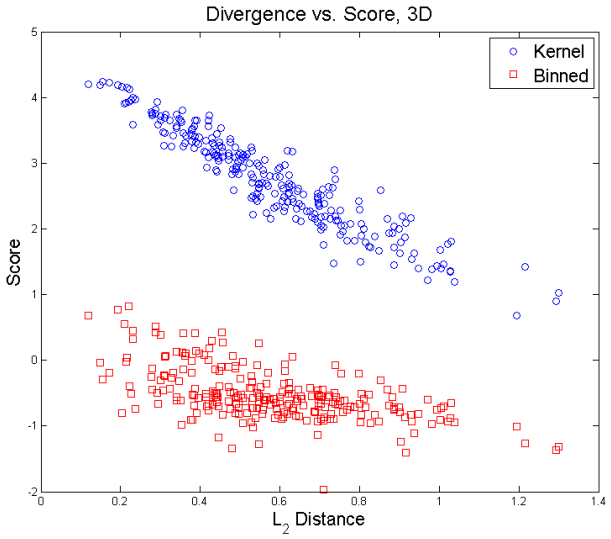

Figure 2: How score correlates with $\mathcal{L}_2$ divergence.

These convergence plots also give an empirical way of calibrating per-sample costs for a particular distribution. For example, based on these experiments, a per-sample cost of 0.014 will prevent roughly 90% of agents from reporting more than 1000 samples to the kernel scoring rule.

Incentive Properties

The next set of experiments explored how the scores rewarded accurate beliefs. We want beliefs that are closer to the ground truth to obtain higher scores.

We investigated this by generating $N = 160$ Laplace mixtures $(P_1, \ldots, P_N)$ to serve as our true distributions, and for each distribution instance P_i generating another $M = 250$ mixtures $(Q_{i,1}, \ldots, Q_{i,M})$ for agent beliefs. Since both the binned and kernel scores are based on the quadratic score, we used squared $\mathcal{L}_2$ distance (evaluated numerically) to measure the divergence between the ground truth and the various beliefs mixtures. For each belief instance $Q_{i,j}$ we evaluated the sampled kernel score and binned score using a $m_1 = 1500$ sample report $X_{i,j} \sim Q_{i,j}^{m_1}$ and a $m_2 = 10,000$ sample evaluation set $Y_{i,j} \sim P_i^{m_2}$. The setting of $m_1 = 1500$ was intended to represent a moderately sized report.

Each belief instance $Q_{i,j}$ generated an ordered triple $(BS_{i,j}, KS_{i,j}, D_{i,j})$ of binned score, kernel score and divergence. Figure 2 shows a scatter plot of the pairs $(D_{i,j}, BS_{i,j})$ and $(D_{i,j}, KS_{i,j})$ for a particular ground truth distribution.. Visually, the kernel score correlates better with divergence.

For each triple we examined how often worse beliefs (with respect to divergence) obtained higher scores. We called this metric *unfairness*. For the kernel score unfairness is defined as:

$$\text{Unfair}_{KS}(j|i) = \frac{\sum_k \mathbb{I}[(KS_{i,k} > KS_{i,j}) \wedge (D_{i,k} > D_{i,j})]}{\sum_k \mathbb{I}[D_{i,k} > D_{i,j}]}.$$

Unfair_{BS} can be defined similarly for the binned score. These capture situations where the incentives are opposite to what one would expect. One way of aggregating unfairness over multiple belief instances is to report the *area under*

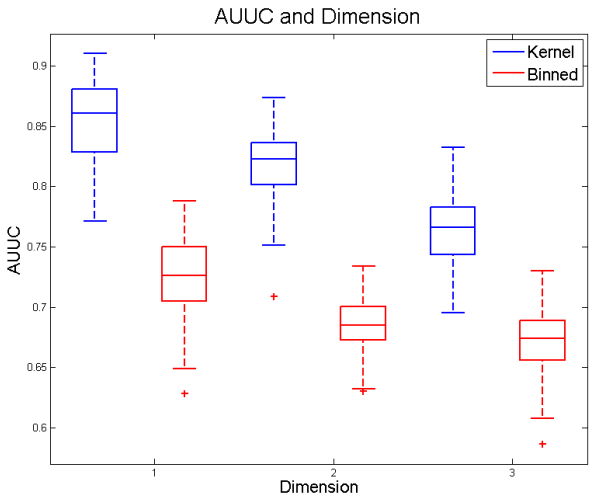

Figure 3: Box and whisker plot for AUUC for each dimension, alternating score type.

the unfairness CDF (AUUC). The AUUC is 1 if the score is never unfair with respect to the divergence, and about 0.5 if the score is random. We also examined Kendall's τ which is a more standard rank correlation coefficient.

The quality of the binned score—both in terms of AUUC and Kendall's τ—depended strongly on the number of bins used. The settings used in this experiment were found through an initial set of calibration experiments. In 1D roughly 45 bins seemed reasonable, whereas 6^2 bins was the best in 2D, and 4^3 bins was the best in three dimensions. The quality of the kernel score was less sensitive to bandwidth. Experiments indicated that a bandwidth between 0.1 and 0.7 seemed reasonable for all three dimensions, with a slight benefit to broader kernels in 3D. We used 0.25 as our bandwidth in all three dimensions.

In all three dimensions, the kernel scoring rule had a significantly higher median AUUC and τ than the binned score.[2] More emphatically, 90% confidence intervals (over the 160 true distribution instances) did not overlap for either AUUC or τ in any dimension. Figure 3 summarizes the AUUC distribution as a function of dimension. The kernel score exhibits a slightly stronger dimensional dependence than binning, but this is likely due to our use of the same bandwidth for all three dimension—allowing the bandwidth to grow with dimension may ameliorate this. Despite this handicap the kernel score consistently outperforms the more carefully tuned binned approach in all dimensions.

Analyzing each distribution instance more closely we found that the kernel score's unfairness CDF stochastically dominated the binned score in a number of instances. This means that the elicitor would prefer the kernel score over the binned score as long as their utility function for an agent's report decreases with increasing $\mathcal{L}_2$ distance. Dominance occurred in 77% of the 1D instances, 71% of the 2D distributions and 48% of the 3D instances.

Interestingly, the stochastic version of the kernel scoring rule was not significantly different than the deterministic kernel score in terms of either median AUUC or median τ,[2] despite the aggressive subsampling. The 90% confidence intervals for AUUC and τ were nearly identical for the two variants, and the CDFs for AUUC and τ were not significantly different in any dimension.[3] This suggests that while stochastic sampling adds noise to the kernel score (as seen in the convergence experiments), it could be a useful approximation in applications where the $\Theta(m^2)$ kernel scoring rule proves too slow.

Discussion

In this paper we have been concerned with the fundamental problem of eliciting a single agents' subjective density. A promising avenue for future work is to use our kernel-based sample score as the basis for multi-agent elicitation and aggregation procedures. We discuss several alternatives.

In a *wagering mechanism* agents place a wager along with their forecasts. Lambert et al. (2008) have proposed a wagering mechanism with several desirable properties including truthfulness and budget-balance; see also (Chen et al. 2014). Their mechanism is based on a proper scoring rule: the payoff to an agent is proportional to its wager according to the difference between its score and the (wager-weighted) score of the group. Our scoring rule can in principle form the basis of these mechanisms, but the implications of sampling in this context must still be worked out.

Under a *market scoring rule*, introduced by (Hanson 2003), agents arrive sequentially and an agent's payoff is the difference between the score of its forecast and the previous agent's forecast, according to a proper scoring rule. If an agent reports P' while the previous report was P, the agent's payoff if ω occurs is simply $S(P',\omega) - S(P,\omega)$, which clearly inherits the incentive properties of the underlying scoring rule. As described the payoffs all occur when the outcome materializes, but for our rule there is also a decomposition that takes an amount in 'escrow':

$$\underbrace{\langle \mu[P'] - \mu[P], \phi(\omega) \rangle}_{\text{payout}} - \underbrace{\tfrac{1}{2}\left(\|\mu[P']\|^2 - \|\mu[P]\|^2 \right)}_{\text{escrow}}.$$

The escrow (which may be negative) depends on the extent to which the agent changes the "complexity" of the forecast distribution, measured in terms of RKHS norm.

Cost-function based prediction markets represent a more sophisticated decomposition of market scoring rules (Abernethy, Chen, and Wortman Vaughan 2011). An agent buys a portfolio of shares and is charged according to a convex cost function. Prediction markets for continuous outcome spaces are an active area of research (Chen, Ruberry, and Vaughan 2013; Gao, Chen, and Pennock 2009). It may be possible to design markets based on the nonparametric ideas in this paper using the known duality relationships between scoring rules and cost functions.

[2] Mann-Whitney, $p < 0.01$, two-sided.

[3] Kolmogorov-Smirnov test, $p < 0.01$

References

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

Boero, G.; Smith, J.; and Wallis, K. F. 2011. Scoring rules and survey density forecasts. *International Journal of Forecasting* 27(2):379–393.

Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.-P.; Schölkopf, B.; and Smola, A. J. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22(14):e49–e57.

Chen, Y.; Devanur, N. R.; Pennock, D. M.; and Vaughan, J. W. 2014. Removing arbitrage from wagering mechanisms. In *Proceedings of the 15th ACM conference on Economics and Computation (EC)*, 377–394.

Chen, Y.; Ruberry, M.; and Vaughan, J. W. 2013. Cost function market makers for measurable spaces. In *Proceedings of the 14th ACM Conference on Electronic Commerce (EC)*, 785–802.

Clemen, R. T.; Fischer, G. W.; and Winkler, R. L. 2000. Assessing dependence: Some experimental results. *Management Science* 46(8):1100–1115.

Dawid, A. P. 1998. Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design. Technical Report 139, Department of Statistical Science, University College London.

Eaton, M. L. 1982. A method for evaluating improper prior distributions. *Statistical Decision Theory and Related Topics III* 1:329–352.

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

Gao, X.; Chen, Y.; and Pennock, D. M. 2009. Betting on the real line. In *Proceedings of the 5th Workshop on Internet and Network Economics (WINE)*. 553–560.

Garthwaite, P. H.; Kadane, J. B.; and O'Hagan, A. 2005. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association* 100(470):680–701.

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

Goldstein, D. G.; Johnson, E. J.; and Sharpe, W. F. 2008. Choosing outcomes versus choosing products: Consumer-focused retirement investment advice. *Journal of Consumer Research* 35(3):440–456.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2006. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems (NIPS)*, 513–520.

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

Kadane, J. B.; Dickey, J. M.; Winkler, R. L.; Smith, W. S.; and Peters, S. C. 1980. Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association* 75(372):845–854.

Lambert, N. S.; Langford, J.; Wortman, J.; Chen, Y.; Reeves, D.; Shoham, Y.; et al. 2008. Self-financed wagering mechanisms for forecasting. In *Proceedings of the 9th ACM conference on Electronic Commerce (EC)*, 170–179. ACM.

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

Matheson, J. E., and Winkler, R. L. 1976. Scoring rules for continuous probability distributions. *Management Science* 22(10):1087–1096.

O'Hagan, A.; Buck, C. E.; Daneshkhah, A.; Eiser, J. R.; Garthwaite, P. H.; Jenkinson, D. J.; Oakley, J. E.; and Rakow, T. 2006. *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons.

O'Hagan, A. 1998. Eliciting expert beliefs in substantial practical applications. *Journal of the Royal Statistical Society: Series D (The Statistician)* 47(1):21–35.

Papakonstantinou, A.; Rogers, A.; Gerding, E. H.; and Jennings, N. R. 2011. Mechanism design for the truthful elicitation of costly probabilistic estimates in distributed information systems. *Artificial Intelligence* 175(2):648–672.

Robu, V.; Kota, R.; Chalkiadakis, G.; Rogers, A.; and Jennings, N. R. 2012. Cooperative virtual power plant formation using scoring rules. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI)*.

Rose, H.; Rogers, A.; and Gerding, E. H. 2012. A scoring rule-based mechanism for aggregate demand prediction in the smart grid. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 661–668.

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

Schölkopf, B., and Smola, A. J. 2002. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

Smola, A.; Gretton, A.; Song, L.; and Schölkopf, B. 2007. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, 13–31.

Sriperumbudur, B. K.; Gretton, A.; Fukumizu, K.; Schölkopf, B.; and Lanckriet, G. R. 2010. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research* 11:1517–1561.

Steinwart, I. 2002. On the influence of the kernel on the consistency of support vector machines. *The Journal of Machine Learning Research* 2:67–93.

Tay, A. S., and Wallis, K. F. 2002. Density forecasting: A survey. *Companion to Economic Forecasting* 45–68.

Wand, M. P., and Jones, M. C. 1995. *Kernel smoothing*. CRC Press.