Recent Trends in Information Elicitation

RAFAEL FRONGILLO
University of Colorado Boulder
and
BO WAGGONER
University of Colorado Boulder

This note provides a survey for the Economics and Computation community of some recent trends in the field of information elicitation. At its core, the field concerns the design of incentives for strategic agents to provide accurate and truthful information. Such incentives are formalized as *proper scoring rules*, and turn out to be the same object as *loss functions* in machine-learning settings, providing many connections. More broadly, the field concerns the design of mechanisms to obtain information from groups of agents and aggregate it or use it for decision making. Recently, work on information elicitation has expanded and been connected to online no-regret learning, mechanism design, fair division, and more.

1. BACKGROUND: SCORING RULES AND PROPERTY ELICITATION

1.1 Introduction

This note surveys recent trends in *information elicitation*, the design of incentives for strategic agents to provide information. Before discussing these recent developments, we recall the key tools of the field, proper scoring rules and property elicitation. These can be viewed as *single-agent* incentives; we then recall multi-agent settings including prediction markets, wagering mechanisms, and peer prediction.

1.2 Single-agent elicitation

The original single-agent elicitation problem is the design of *proper scoring rules* [Brier 1950; Good 1952; Savage 1971; Gneiting and Raftery 2007]. An agent predicts a probability distribution p, then an outcome y is observed. A scoring rule S is a function assigning a score $S(p, y)$ to the prediction p on the obeservation y. S is *proper* if the agent maximizes expected score, over the randomness in y, by predicting their true belief. For example, the *log score* $S(p, y) = \log p(y)$ is proper.[1]

Scoring rules have been extended to the case where the agent predicts the mean, mode, variance, or any other *property* $\Gamma(p)$ of the distribution p of y [Osband 1985; Lambert et al. 2008]; see Gneiting [2011] for the origins of these ideas. The score $S(r, y)$ is said to *elicit* the property Γ if, when the agent's true belief is the distribution p, they uniquely maximize expected score by reporting $r = \Gamma(p)$. For example, the quadratic score $S(r, y) = -\|r - y\|_2^2$, where y and r take values in Euclidean space, elicits the *mean*: the score is maximized by setting $r = \mathbb{E}_p Y$. As

[1]To see that the log score is proper, note that the regret of reporting q instead of the true p is $\mathbb{E}_p S(p, Y) - \mathbb{E}_p S(q, Y) = \mathrm{KL}(p\|q)$, which is nonnegative and uniquely zero at the report $q = p$.

Authors' addresses: raf@colorado.edu, bwag@colorado.edu

is common in machine learning, the scoring rule can be rephrased as a *loss function* by taking the negative, $\ell(r, y) = \|r - y\|_2^2$.

In both the classic scoring rule and the real-valued property elicitation setting, we have general characterizations of which scoring rules are proper and how they can be generated, relating to convex functions [Gneiting and Raftery 2007; Osband 1985; Lambert et al. 2008; Lambert 2018; Steinwart et al. 2014]. It turns out that some properties, such as the variance, cannot be directly elicited by any scoring rule; for example, for a property to be elicitable, its *level sets* $\{p : \Gamma(p) = r\}$, i.e., the set of distributions sharing a particular a property value, must be a convex set. When a property is not elicitable, one often considers "indirect" elicitation, where the agent provides some other information from which one can compute the property of interest. How much other information is required to indirectly elicit a property Γ is known as its *elicitation complexity* [Lambert et al. 2008; Fissler et al. 2016; Frongillo and Kash 2021a]. For example, the variance has elicitation complexity 2, because it can be computed from the mean $\mathbb{E}\, Y$ and the second moment $\mathbb{E}\, Y^2$, both of which are elicitable.

1.3 Multi-agent elicitation

There is an extensive literature on eliciting information from groups of agents, often utilizing proper scoring rules as a key tool. Prediction markets [Hanson 2003; Chen and Pennock 2007; Abernethy et al. 2013; Frongillo and Waggoner 2018] and wagering mechanisms [Lambert et al. 2008; Lambert et al. 2015; Freeman and Pennock 2018] are two of the most common. In a prediction market, agents arrive dynamically and make updates to a consensus forecast (equivalently, purchase shares corresponding to predicted events). In wagering mechansims, agents simultaneously submit "sealed-bid" predictions and wagers. In each case, final payoffs are assigned based on the eventual observed outcome, according to a function typically based on a proper scoring rule. When ground truth is not available—that is, when the outcome y cannot be observed—information can still be elicited by comparing agent reports against each other. This large area of research is referred to as *information elicitation without verification* or the *peer prediction* literature after an eponymous paper [Miller et al. 2005].

2. SINGLE-AGENT ELICITATION

2.1 Elicitability

The term "elicitability", common in statistics and finance, refers to understanding whether or not a particular property/statistic is elicitable, and if not, what its elicitation complexity is. The fact that elicitable properties have convex level sets has been the main tool to rule out elicitability, such as for the variance and many financial risk measures of interest (see below). Yet since the beginning, one common statistic stood out: the mode, meaning $\Gamma(p) = \operatorname{argmax}_y p(y)$ when p is a probability density function (and suitable generalizations). The mode is interesting because it was widely thought not to be elicitable, yet it does have convex level sets: mixing two distributions with mode r gives a distribution with mode r. Heinrich [2014] showed that indeed the mode is not elicitable, even for some restricted classes of distributions. Even worse, its elicitation complexity is infinite [Dearborn and Frongillo

2020]. More recent work shows that the mode is "asymptotically elicitable", as the limit of the midpoint of modal intervals [Dimitriadis et al. 2019], and that the mode still fails to be elicitable even when restricting to *strongly unimodal* densities, which have a unique local maximum [Heinrich-Mertsching and Fissler 2022].

A similar story has unfolded in the literature on financial risk measures. Gneiting [2011] caught the attention of this community by observing that Expected Shortfall (ES), a popular risk measure for the regulation of banks, is not elicitable [Embrechts et al. 2014]. A few years later, Fissler and Ziegel [2016] proved that the pair (VaR,ES) is elicitable, where VaR (Value at Risk) is simply a quantile. Their result shows that the elicitation complexity of ES is at most 2; a corresponding lower bound was shown by Frongillo and Kash [2021a]. These results are considered positive, in that they enable "backtesting" procedures to verify that banks are holding enough capital in reserve to compensate for their risk [Fissler et al. 2016]. In parallel to this study of ES, there has been a flurry of research in the statistics and finance communities on the elicitability of various risk and uncertainty measures [Wang and Ziegel 2015; Wang and Wei 2018; Fissler and Ziegel 2021; Fissler et al. 2024; Fissler and Ziegel 2021; Fissler et al. 2021].

Various extensions or generalizations of elicitability have also appeared, such as the "asymptotic elicitability" above. Another notion is *multi-observation elicitation* where one assumes access to multiple independent copies of the outcome Y [Casalaina-Martin et al. 2017; Frongillo et al. 2019]; an example result is that the squared 2-norm of a distribution $\|p\|_2^2$ is elicitable with 2 observations, despite having large elicitation complexity in the usual setting with 1 observation. Finally, the notion of *conditional elicitation* allows one to first elicit a property Γ_1, and then elicit Γ_2 assuming knowledge of $\Gamma_1(p)$ [Emmer et al. 2015; Frongillo and Kash 2015b; Fissler and Hoga 2024].

A few open problems stand out. First, perhaps the main question remaining in the study of the mode is its elicitation complexity with respect to strongly unimodal distributions. Second, there are many financial risk measures whose elicitation complexity is unknown. One interesting example is the Gini coefficient, given by $(\mathbb{E}|Y_1 - Y_2|)/(2\mathbb{E}Y)$, where Y_1, Y_2 are independent copies of Y [Bellini et al. 2022]. (We note that this property is elicitable with 2 observations, using a scoring rule for the ratio of expectations.) Finally, a nice step toward a general characterization of elicitable vector-valued properties [Frongillo and Kash 2015a] would be to understand the elicitation complexity of the nth central moment, $\Gamma_n(p) = \mathbb{E}_p(Y - \mathbb{E}_p Y)^n$. The main tools for elicitation complexity lower bounds fail for this example, since it is 2-identifiable (a first-order condition) yet the best known upper bound is n, by eliciting the first n moments. The fact that Γ_n is conditionally elicitable conditioned on the mean allows for some partial progress [Frongillo and Kash 2015b].

2.2 Incentives for acquiring information or exerting effort

While proper scoring rules incent agents to provide information they already have, a natural question is how to incentivize acquisition of new information, or exertion of effort to produce more accurate predictions. A proper scoring rule will generally pay agents more (in expectation) for better information, so a common approach is to consider the optimal shape of such a scoring rule. A significant amount of recent work has considered aspects of this problem. This work includes Neyman et.

al [2021]; Li et. al [2022] and Hartline et. al [2023] with a motivation of incentivizing effort on the part of e.g. students to learn material; Li and Libgober [2023]; Chen and Yu [2023]; Carrol [2019] and Papireddygari and Waggoner [2022] in the context of contracts; Zhang and Schoenebeck [2023] in a peer-prediction context; and Schoenebeck et. al [2021] in a prediction-markets context.

3. MULTI-AGENT ELICITATION

3.1 Aggregation of forecasts

Given a set of forecasts, how should they be aggregated into a single prediction? This problem has received significant recent theoretical attention, starting with Arieli et. al [2018]. Although the problem does not necessarily involve incentives, it arises naturally in conjunction with e.g. wagering mechanisms and forecasting competitions, which elicit such a set of forecasts.

In the aggregation problem, a set of signals $(S_1, \ldots, S_n, Y)$ are drawn jointly from a prior distribution. Here Y is the outcome to be predicted, a binary or real-valued random variable. Each expert $i \in \{1, \ldots, n\}$ observes the realization of their signal S_i and updates to a posterior belief with updated expectation $X_i = \mathbb{E}[Y \mid S_i]$. The Bayes-optimal prediction would be $R^* = \mathbb{E}[Y \mid S_1, \ldots, S_n]$, the Bayesian aggreagtion of the information available to all agents. An aggregator collects the individual predictions and produces an estimate $R = R(X_1, \ldots, X_n)$. Performance is typically measured as the difference in expected squared loss compared to the optimal aggregator, i.e. $\mathbb{E}\left[(R - Y)^2 - (R^* - Y)^2\right]$.

While Arieli et. al [2018] used an additive regret notion to measure performance, Neyman and Roughgarden [2022] considered a competitive-ratio approach. In each case, it is generally impossible to give nontrivial results for arbitrary information structures,[2] so research focuses on classes of information structures for which positive results are possible. Arieli et. al [2018] considers structures such as conditionally independent signals and Blackwell-ordered experts; Neyman and Roughgarden [2022] considers, in particular, a substitutes condition. Follow-up and other recent work, which generally focuses on improving the bounds for different classes of structures, includes [Levy and Razin 2021; 2022; De Oliveira et al. 2021; Lin and Chen 2023; Guo et al. 2024].

3.2 Information elicitation without verification

The IEWV or "peer prediction" literature is large and generally well-known to the Economics and Computation community; one recent survey in the area is Faltings [2023]. This brief section will be incomplete, but we highlight some important recent progress and interesting recent ideas.

Since Dasgupta and Ghosh [2013], an important paradigm has been the *multi-task* setting in which a group of agents are each asked for their answers to a set of questions, with rewards determined by comparing their sets of answers. Recently, Kong [2020; 2024] gave the *Determiant Mutual Information (DMI)* mechanism which achieves the strongest possible notion of truthfulness using a constant number

[2]For example, even averaging the agents' predictions can be an arbitrarily bad idea as compared to e.g. selecting one of them at random.

of questions.[3] Zheng et al. [2021] uses theory of property elicitation to give lower bounds (impossibility results) for multi-task elicitation mechanisms.

In the even more challenging *single-question* setting, Schoenebeck and Yu [2023] give strongly-truthful mechanisms with just three agents, each of which answers just a single question. The mechanism is inspired by the Bayesian truth serum of Prelec [2004]; a similar result is given independently by Prelec [2021].

Of other recent work on peer prediction with an elicitation focus, we note Wang et al. [2021], which considers aggregation of forecasts via a peer-prediction style perspective; Liu et al. [2023], which designs "surrogate scoring rules" which can, in some settings, use peer reports to replicate the incentives of a proper scoring rule; and Zhang and Schoenebeck [2023], which considers incentives in peer prediction to exert effort or acquire information.

3.3 Forecasting competitions

In a forecasting competition, a group of agents provide forecasts for a set of events, then the outcome is observed and a single winner is selected. (A *wagering mechanism*, in contrast, may assign various levels of payouts to any or all of the participants.) The Kaggle machine-learning platform, for example, implements forecasting competitions. The "winner-take-all" structure of such competitions distorts the incentives for accurate forecasting. A truthful mechanism was given by Witkowski, et al. [2018], with more results provided in the journal version [Witkowski et al. 2023]. Frongillo, et al. [2021] give a mechanism that is only approximately truthful, but can select the most accurate forecaster with a smaller set of forecasted events. Guarantees have also recently been extended to the setting of correlated events [Frongillo et al. 2023].

3.4 Prediction markets and decentralized finance

The design of automated market makers, algorithms that offer prices to buy or sell assets, originated in the prediction market community as a solution to thin market problems [Hanson 2003]. There has been a significant literature on the theory of prediction markets (see e.g. [Abernethy et al. 2013; Frongillo and Waggoner 2018] and references therein) and efficient implementation in combinatorial settings (e.g. [Dudík et al. 2013; Wang et al. 2021] and references therein). One recent work, Schoenebeck, et al. [2021], considers incentives in prediction markets for exerting effort and acquiring information. Kong and Schoenebeck [2023] examines when information in prediction markets is fully aggregated, relating to when signals are substitutes; Frongillo, et al. [2023] considers a similar question in the case of general communication protocols inspired by prediction markets.

Despite the large literature on the design of automated market makers for prediction markets, their adoption in practice has been limited. By contrast, automated market makers are quite popular mechanisms to run decentralized exchanges in blockchain settings, having traded billions of dollars of assets in recent years [Angeris and Chitra 2020; Angeris et al. 2022]. The dominant paradigm in the blockchain context is the class of *constant-function market makers* (CFMMs). It turns out that CFMMs, while not designed to elicit information per se, are equiv-

[3]Details of the setting, such as i.i.d. assumptions, are of interest but omitted here.

alent in a strong sense to prediction markets [Frongillo et al. 2024]. We expect this connection to spark a lively exchange of ideas between these two previously independent literatures.

3.5 Mechanism design and fair division

While information elicitation concerns incentives for an agent or group of agents to report honest information, it has deep connections to mechanism design and more generally the problem of allocating goods or services. These connections were perhaps first observed by Fiat, et al. [2013] and later expanded by Frongillo and Kash [2014; 2021b]; these papers give constructions to convert scoring rules to single-agent mechanisms and vice versa, among other connections and results.

The points of contact between these literatures continue to broaden. Particularly emblematic is the recent work showing the equivalence between wagering mechanisms and fair division mechanisms [Freeman et al. 2019; Freeman et al. 2023]. The authors apply this equivalence to a popular wagering mechanism to give the first nontrivial mechanism which is incentive-compatible, proportional, and envy-free. We anticipate more results of this type for multi-agent mechanism design, as researchers continue to leverage the perspective and techniques of information elicitation.

3.6 Online learning from strategic experts

A recent line of work, initiated by [Roughgarden and Schrijvers 2017], asks how to conduct online no-regret learning from expert advice when the experts have incentives to be "chosen" and may misreport their predictions. Exactly truthful mechanisms, built on connections to forecasting competitions, are studied by [Freeman et al. 2020; Mortazavi et al. 2024]. Mechanisms that are only approximately truthful but satisfy good regret guarantees are studied by [Frongillo et al. 2021] and then in the "m-experts" setting by [Sadeghi and Fazel 2023].

4. CONNECTIONS TO MACHINE LEARNING

The design of scoring rules or loss functions is also an active area of research in machine learning. We first review the basic framework of supervised machine learning, and the role of property elicitation.

4.1 Indirect elicitation in machine learning

In supervised machine learning, we wish to learn a model or hypothesis which makes a prediction given some feature vector $x \in \mathcal{X}$. Most algorithms employ *empirical risk minimization (ERM)*, which simply chooses a hypothesis $h : \mathcal{X} \to \mathbb{R}^d$ from a class $\mathcal{H}$ that minimizes loss over a data set:

$$h^* \in \operatorname*{argmin}_{h \in \mathcal{H}} \sum_{(x,y) \in \text{data}} L(h(x), y) \ . \tag{1}$$

For example, in ordinary least-squares (OLS) regression, $\mathcal{H}$ consists of linear functions from $\mathcal{X}$ to $\mathbb{R}$, and $L_2(r, y) = (r - y)^2$ is squared error. In regression problems, the statistical consistency ("correctness") of ERM boils down to a question of property elicitation for the conditional distributions $\Pr[Y|X = x]$: whether L elicits the desired conditional statistic (Fig. 1(a)). As squared error elicits the mean, OLS

therefore fits to the conditional mean $h^*(x) = \mathbb{E}[Y|X=x]$ so long as $h^* \in \mathcal{H}$ (see below). Similarly, absolute loss $L_1(r, y) = |r - y|$ yields median regression.

In discrete prediction problems such as classification, ranking, and structured prediction, we are instead given a target discrete loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ for finite sets $\mathcal{R}$ (predictions) and $\mathcal{Y}$ (labels), and wish to learn a hypothesis $h_{\text{targ}} : \mathcal{X} \to \mathcal{R}$ achieving low expected loss $\mathbb{E}_D \ell(h_{\text{targ}}(X), Y)$ over the underlying distribution D. For example, traditional classification has $\mathcal{R} = \mathcal{Y}$ with $\ell(r, y) = \mathbb{1}\{r \neq y\}$ being 0-1 loss (penalty 0 if correct, 1 if incorrect). Solving ERM (1) is generally NP-hard for discrete losses ℓ, so instead we seek a *surrogate* loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ which is convex in the first argument. This d is called the *prediction dimension*, and plays a key role in structured prediction, where it can be exponentially large in the natural dimension of the problem [Osokin et al. 2017]. The hypothesis $h_{\text{surg}} : \mathcal{X} \to \mathbb{R}^d$ is then converted to one answering the target problem via a *link function* $\psi : \mathbb{R}^d \to \mathcal{R}$ mapping back to target predictions.

$$h_{\text{surg}} \in \operatorname*{argmin}_{h \in \mathcal{H}} \sum_{(x,y) \in \text{data}} L(h(x), y) \ , \quad h_{\text{targ}} = \psi \circ h_{\text{surg}} \ . \tag{2}$$

Many algorithms follow this paradigm, including support vector machines (SVMs), logistic regression, boosting, and deep neural networks.

For surrogate ERM (2), consistency means that low surrogate (L) loss of h_{surg} should imply low target (ℓ) loss of the linked hypothesis $h_{\text{targ}} = \psi \circ h_{\text{surg}}$ given more and more data. Consistency is a precursor to *rates*, which quantify how fast target loss is minimized. When the class $\mathcal{H}$ is sufficiently rich, consistency of surrogate minimization reduces to *calibration*, a condition stating that predictions cannot approach the optimal surrogate loss while linking to incorrect target predictions [Bartlett et al. 2006; Tewari and Bartlett 2007; Agarwal and Agarwal 2015]. Just as with regression problems, calibration only depends on the loss with respect to conditional distributions on $\mathcal{Y}$. Moreover, calibration implies "in-

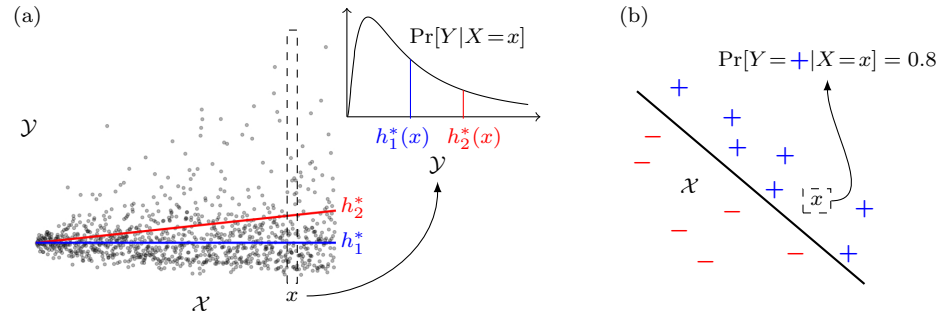

Fig. 1. For rich hypothesis classes $\mathcal{H}$, consistency of a learning algorithm reduces to its behavior on the conditional distributions $Y|X=x$ for each $x \in \mathcal{X}$. **(a)** In linear regression, minimizing absolute loss L_1 yields hypothesis h_1^*, where $h_1^*(x)$ tracks the median of $Y|X=x$. Squared loss L_2 instead tracks the conditional mean. **(b)** In SVMs for classification, minimizing hinge loss conditioned on $X=x$, and taking the sign of the result (the link function), will minimize conditional 0-1 loss.

direct" property elicitation, meaning $\gamma = \psi \circ \Gamma$ where ℓ elicits γ and L elicits Γ.[4] For example, 0-1 loss elicits the mode (most likely label), and the SVM hinge loss $L_{\text{hinge}}(r, y) = \max(0, 1 - ry)$ indirectly elicits the mode with the link $\psi(r) = \text{sign}(r)$ (Fig. 1(b)), where here $r \in \mathbb{R}$ and $y \in \mathcal{Y} = \{-1, 1\}$.

4.2 Surrogate loss design

Compared to the property elicitation results discussed in § 2.1, perhaps the most significant constraint in machine learning settings is that the surrogate loss should ideally be convex and thus efficient to optimize. Somewhat surprisingly, convexity of the loss comes for free for continuous real-valued properties [Finocchiaro and Frongillo 2018], i.e., in prediction dimension 1, but for higher prediction dimensions this is no longer the case. By analogy to elicitation complexity, recent work has tried to understand the lowest prediction dimension needed for a consistent surrogate loss to exist [Agarwal and Agarwal 2015; Ramaswamy and Agarwal 2016; Finocchiaro et al. 2020]. One interesting example is the *abstain property*, which takes the value of abstain ($\perp$) if no label has probability at least 0.5, and is the most likely label otherwise. Naively a convex loss for this property would require prediction dimension $n - 1$ for n labels, but a clever construction due to Ramaswamy, et al. [2018] uses only $\log n$ dimensions. Motivated by this work, Finocchiaro et al. [2019; 2024] give a general framework to design consistent convex surrogate loss functions for any target. Despite this progress, however, many important questions remain; perhaps most glaring is the lack of tools to bound the prediction dimension required for most target problems.

4.3 Multicalibration and decision robustness

A particular trend of interest is the problem of predicting without knowing which scoring rule (or loss function) one is predicting for. Constructing such "omnipredictors" is the focus of Gopalan et. al [2022] and Hu et. al [2023]. We are given a dataset and, instead of a single target loss, a family of loss functions $\mathcal{L}$. The goal is to learn a hypothesis h that performs well on every loss in $\mathcal{L}$ simultaneously, possibly with loss-specific post-processing. A conceptually similar problem is studied in an online learning setting by Kleinberg et. al [2023]; see also Ehm et al. [2016].

These works are closely related to *multicalibration* [Hebert-Johnson et al. 2018], a concept of interest in fair machine learning. In the binary classification setting, a hypothesis $h : \mathcal{X} \to [0, 1]$ is called calibrated if, among the subset of pairs (x, y) for which $h(x) = c$, a c-fraction have true label $y = 1$. In multicalibration, h must be calibrated even when conditioning on particular subgroups; this turns out to be useful for achieving omniprediction. In the spirit of property elicitation, Jung et. al [2021] extends multicalibration from means to properties such as the variance.

Acknowledgements and requests for suggestions

We thank Tilmann Gneiting, Ian Kash, and Jens Witkowski. We appreciate and welcome any additional comments or suggestions of works that we have missed.

[4]It remains an interesting open problem to study exactly when and how calibration and indirect elicitation differ in practical examples.

REFERENCES

Abernethy, J., Chen, Y., and Vaughan, J. W. 2013. Efficient market making via convex optimization, and a connection to online learning. *ACM Transactions on Economics and Computation 1,* 2, 12.

Agarwal, A. and Agarwal, S. 2015. On consistent surrogate risk minimization and property elicitation. In *Conference on Learning Theory*. 4–22.

Angeris, G., Agrawal, A., Evans, A., Chitra, T., and Boyd, S. 2022. Constant function market makers: Multi-asset trades via convex optimization. In *Handbook on Blockchain*. Springer, 415–444.

Angeris, G. and Chitra, T. 2020. Improved price oracles: Constant function market makers. In *Proceedings of the 2nd ACM Conference on Advances in Financial Technologies*. 80–91.

Arieli, I., Babichenko, Y., and Smorodinsky, R. 2018. Robust forecast aggregation. *Proceedings of the National Academy of Sciences 115,* 52, E12135–E12143.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. 2006. Convexity, classification, and risk bounds. *Journal of the American Statistical Association 101,* 473, 138–156.

Bellini, F., Fadina, T., Wang, R., and Wei, Y. 2022. Parametric measures of variability induced by risk measures. *Insurance: Mathematics and Economics 106*, 270–284.

Brier, G. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review 78,* 1, 1–3.

Carroll, G. 2019. Robust incentives for information acquisition. *Journal of Economic Theory 181*, 382–420.

Casalaina-Martin, S., Frongillo, R., Morgan, T., and Waggoner, B. 2017. Multi-observation elicitation. In *Conference on Learning Theory*. PMLR, 449–464.

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

Chen, Y. and Yu, F.-Y. 2023. Optimal scoring rule design under partial knowledge.

Dasgupta, A. and Ghosh, A. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web*. 319–330.

De Oliveira, H., Ishii, Y., and Lin, X. 2021. Robust merging of information. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. EC '21. Association for Computing Machinery, New York, NY, USA, 341–342.

Dearborn, K. and Frongillo, R. 2020. On the indirect elicitability of the mode and modal interval. *Annals of the Institute of Statistical Mathematics 72,* 5, 1095–1108.

Dimitriadis, T., Patton, A. J., and Schmidt, P. W. 2019. Testing forecast rationality for measures of central tendency. *arXiv preprint arXiv:1910.12545* .

Dudík, M., Lahaie, S., Pennock, D. M., and Rothschild, D. 2013. A combinatorial prediction market for the us elections. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*. 341–358.

Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. 2016. Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society Series B: Statistical Methodology 78,* 3, 505–562.

Embrechts, P., Puccetti, G., Rüschendorf, L., Wang, R., and Beleraj, A. 2014. An Academic Response to Basel 3.5. *Risks 2,* 1 (Feb.), 25–48.

Emmer, S., Kratz, M., and Tasche, D. 2015. What is the best risk measure in practice? a comparison of standard measures. *Journal of Risk 18,* 2, 31–60.

Faltings, B. 2023. Game-theoretic mechanisms for eliciting accurate information. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. IJCAI '23.

Fiat, A., Karlin, A., Koutsoupias, E., and Vidali, A. 2013. Approaching Utopia: strong truthfulness and externality-resistant mechanisms. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*. 221–230.

Finocchiaro, J. and Frongillo, R. 2018. Convex elicitation of continuous properties. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 10425–10434.

Finocchiaro, J., Frongillo, R., and Waggoner, B. 2019. An embedding framework for consistent polyhedral surrogates. In *Proceedings of Advances In Neural Information Processing Systems (NeurIPS)*. 10780–10790.

Finocchiaro, J., Frongillo, R., and Waggoner, B. 2020. Embedding dimension of polyhedral losses. In *Proceedings of Conference on Learning Theory (COLT)*.

Finocchiaro, J., Frongillo, R. M., and Waggoner, B. 2024. An embedding framework for the design and analysis of consistent polyhedral surrogates. *Journal of Machine Learning Research 25,* 63, 1–60.

Fissler, T., Frongillo, R., Hlavinová, J., and Rudloff, B. 2021. Forecast evaluation of quantiles, prediction intervals, and other set-valued functionals. *Electronic Journal of Statistics 15*, 1034–1084.

Fissler, T. and Hoga, Y. 2024. Backtesting systemic risk forecasts using multi-objective elicitability. *Journal of Business & Economic Statistics 42,* 2, 485–498.

Fissler, T., Liu, F., Wang, R., and Wei, L. 2024. Elicitability and identifiability of tail risk measures. *arXiv preprint arXiv:2404.14136*.

Fissler, T., Ziegel, J., and Gneiting, T. 2016. Expected Shortfall is jointly elicitable with Value at Risk-Implications for backtesting. *Risk Magazine*.

Fissler, T. and Ziegel, J. F. 2016. Higher order elicitability and Osband's principle. *The Annals of Statistics 44,* 4, 1680–1707.

Fissler, T. and Ziegel, J. F. 2021. On the elicitability of range value at risk. *Statistics & risk modeling 38,* 1-2, 25–46.

Freeman, R., Pennock, D., Podimata, C., and Vaughan, J. W. 2020. No-regret and incentive-compatible online learning. In *International Conference on Machine Learning*. PMLR, 3270–3279.

Freeman, R. and Pennock, D. M. 2018. An axiomatic view of the parimutuel consensus wagering mechanism. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1936–1938.

Freeman, R., Pennock, D. M., and Vaughan, J. W. 2019. An equivalence between wagering and fair-division mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 1957–1964.

Freeman, R., Witkowski, J., Wortman Vaughan, J., and Pennock, D. M. 2023. An equivalence between fair division and wagering mechanisms. *Management Science 0,* 0.

Frongillo, R., Gomez, R., Thilagar, A., and Waggoner, B. 2021. Efficient competitions and online learning with strategic forecasters. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 479–496.

Frongillo, R. and Kash, I. 2014. General truthfulness characterizations via convex analysis. In *Web and Internet Economics*. Springer, 354–370.

Frongillo, R. and Kash, I. 2015a. Vector-Valued Property Elicitation. In *Proceedings of the 28th Conference on Learning Theory*. 1–18.

Frongillo, R. and Kash, I. A. 2015b. On Elicitation Complexity and Conditional Elicitation. *arXiv preprint arXiv:1506.07212*.

Frongillo, R., Lladser, M., Thilagar, A., and Waggoner, B. 2023. Forecasting competitions with correlated events. *arXiv preprint arXiv:2303.13793*.

Frongillo, R., Mehta, N. A., Morgan, T., and Waggoner, B. 2019. Multi-Observation Regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*. 2691–2700.

Frongillo, R., Neyman, E., and Waggoner, B. 2023. Agreement implies accuracy for substitutable signals. In *Proceedings of the 2023 ACM Conference on Economics and Computation*. EC. ACM.

Frongillo, R., Papireddygari, M., and Waggoner, B. 2024. An Axiomatic Characterization of CFMMs and Equivalence to Prediction Markets. In *15th Innovations in Theoretical Computer Science Conference (ITCS 2024)*. Vol. 287. 51:1–51:21.

Frongillo, R. and Waggoner, B. 2018. An axiomatic study of scoring rule markets. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science (ITCS)*. 15:1–15:20.

Frongillo, R. M. and Kash, I. A. 2021a. Elicitation complexity of statistical properties. *Biometrika 108,* 4, 857–879.

Frongillo, R. M. and Kash, I. A. 2021b. General truthfulness characterizations via convex analysis. *Games and Economic Behavior 130*, 636–662.

Gneiting, T. 2011. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association 106,* 494, 746–762.

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

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

Gopalan, P., Kalai, A. T., Reingold, O., Sharan, V., and Wieder, U. 2022. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, M. Braverman, Ed. Leibniz International Proceedings in Informatics (LIPIcs), vol. 215. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 79:1–79:21.

Guo, Y., Hartline, J. D., Huang, Z., Kong, Y., Shah, A., and Yu, F.-Y. 2024. Algorithmic robust forecast aggregation.

Hanson, R. 2003. Combinatorial Information Market Design. *Information Systems Frontiers 5,* 1, 107–119.

Hartline, J. D., Shan, L., Li, Y., and Wu, Y. 2023. Optimal scoring rules for multi-dimensional effort. In *Proceedings of Thirty Sixth Conference on Learning Theory*, G. Neu and L. Rosasco, Eds. Proceedings of Machine Learning Research, vol. 195. PMLR, 2624–2650.

Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. 2018. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds. Proceedings of Machine Learning Research, vol. 80. PMLR, 1939–1948.

Heinrich, C. 2014. The mode functional is not elicitable. *Biometrika 101,* 1, 245–251.

Heinrich-Mertsching, C. and Fissler, T. 2022. Is the mode elicitable relative to unimodal distributions? *Biometrika 109,* 4, 1157–1164.

Hu, L., Navon, I. R. L., Reingold, O., and Yang, C. 2023. Omnipredictors for constrained optimization. In *International Conference on Machine Learning*. PMLR, 13497–13527.

Jung, C., Lee, C., Pai, M., Roth, A., and Vohra, R. 2021. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*. PMLR, 2634–2678.

Kleinberg, B., Leme, R. P., Schneider, J., and Teng, Y. 2023. U-calibration: Forecasting for an unknown agent. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 5143–5145.

Kong, Y. 2020. Dominantly truthful multi-task peer prediction with a constant number of tasks. In *Proceedings of the Fourteenth annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2398–2411.

Kong, Y. 2024. Dominantly truthful peer prediction mechanisms with a finite number of tasks. *J. ACM 71,* 2 (apr).

Kong, Y. and Schoenebeck, G. 2023. False Consensus, Information Theory, and Prediction Markets. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, Y. Tauman Kalai, Ed. Leibniz International Proceedings in Informatics (LIPIcs), vol. 251. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 81:1–81:23.

Lambert, N. 2018. Elicitation and Evaluation of Statistical Forecasts. *Preprint*.

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

Lambert, N. S., Langford, J., Wortman, J., Chen, Y., Reeves, D., Shoham, Y., and Pennock, D. 2008. Self-financed wagering mechanisms for forecasting. In *Proceedings of the 9th ACM conference on Electronic commerce*. ACM, 170–179.

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*. 129–138.

Levy, G. and Razin, R. 2021. A maximum likelihood approach to combining forecasts. *Theoretical Economics 16,* 1, 49–71.

Levy, G. and Razin, R. 2022. Combining forecasts in the presence of ambiguity over correlation structures. *Journal of Economic Theory 199*, 105075. Symposium Issue on Ambiguity, Robustness, and Model Uncertainty.

Li, Y., Hartline, J. D., Shan, L., and Wu, Y. 2022. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC '22. Association for Computing Machinery, New York, NY, USA, 988–989.

Li, Y. and Libgober, J. 2023. Optimal scoring for dynamic information acquisition.

Lin, T. and Chen, Y. 2023. Sample complexity of forecast aggregation. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. Vol. 36. Curran Associates, Inc., 36053–36093.

Liu, Y., Wang, J., and Chen, Y. 2023. Surrogate scoring rules. *ACM Trans. Econ. Comput. 10,* 3 (feb).

Miller, N., Resnick, P., and Zeckhauser, R. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science 51,* 9, 1359–1373.

Mortazavi, A., Lin, J., and Mehta, N. 2024. On the price of exact truthfulness in incentive-compatible online learning with bandit feedback: a regret lower bound for WSU-UX. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, S. Dasgupta, S. Mandt, and Y. Li, Eds. Proceedings of Machine Learning Research, vol. 238. PMLR, 4681–4689.

Neyman, E., Noarov, G., and Weinberg, S. M. 2021. Binary scoring rules that incentivize precision. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 718–733.

Neyman, E. and Roughgarden, T. 2022. Are you smarter than a random expert? the robust aggregation of substitutable signals. In *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC '22. Association for Computing Machinery, New York, NY, USA, 990–1012.

Osband, K. H. 1985. *Providing Incentives for Better Cost Forecasting*. University of California, Berkeley.

Osokin, A., Bach, F., and Lacoste-Julien, S. 2017. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*. 302–313.

Papireddygari, M. and Waggoner, B. 2022. Contracts with information acquisition, via scoring rules. In *Proceedings of the 2022 ACM Conference on Economics and Computation*. EC. ACM.

Prelec, D. 2004. A Bayesian Truth Serum for Subjective Data. *Science 306,* 5695 (Oct.), 462–466.

Prelec, D. 2021. Bilateral bayesian truth serum: The nxm signals case. *Available at SSRN 3908446*.

Ramaswamy, H. G. and Agarwal, S. 2016. Convex calibration dimension for multiclass loss matrices. *The Journal of Machine Learning Research 17,* 1, 397–441.

Ramaswamy, H. G., Tewari, A., and Agarwal, S. 2018. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics 12,* 1, 530–554.

Roughgarden, T. and Schrijvers, O. 2017. Online prediction with selfish experts. *Advances in Neural Information Processing Systems 30*.

Sadeghi, O. and Fazel, M. 2023. No-regret online prediction with strategic experts. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. Vol. 36. Curran Associates, Inc., 54696–54715.

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

Schoenebeck, G., Yu, C., and Yu, F.-Y. 2021. Timely information from prediction markets. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '21. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1145–1153.

Schoenebeck, G. and Yu, F.-Y. 2023. Two strongly truthful mechanisms for three heterogeneous agents answering one question. *ACM Trans. Econ. Comput. 10,* 4 (feb).

Steinwart, I., Pasin, C., Williamson, R., and Zhang, S. 2014. Elicitation and Identification of Properties. In *Proceedings of The 27th Conference on Learning Theory*. 482–526.

Tewari, A. and Bartlett, P. L. 2007. On the consistency of multiclass classification methods. *The Journal of Machine Learning Research 8*, 1007–1025.

Wang, J., Liu, Y., and Chen, Y. 2021. Forecast aggregation via peer prediction. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 9,* 1 (Oct.), 131–142.

Wang, R. and Wei, Y. 2018. Risk functionals with convex level sets. *Available at SSRN 3292661*.

Wang, R. and Ziegel, J. F. 2015. Elicitable distortion risk measures: A concise proof. *Statistics & Probability Letters 100*, 172–175.

Wang, X., Pennock, D. M., Devanur, N. R., Rothschild, D. M., Tao, B., and Wellman, M. P. 2021. Designing a combinatorial financial options market. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 864–883.

Witkowski, J., Freeman, R., Vaughan, J. W., Pennock, D. M., and Krause, A. 2018. Incentive-compatible forecasting competitions. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. AAAI.

Witkowski, J., Freeman, R., Vaughan, J. W., Pennock, D. M., and Krause, A. 2023. Incentive-compatible forecasting competitions. *Management Science 69,* 3, 1354–1374.

Zhang, Y. and Schoenebeck, G. 2023. High-effort crowds: Limited liability via tournaments. In *Proceedings of the ACM Web Conference 2023*. WWW '23. Association for Computing Machinery, New York, NY, USA, 3467–3477.

Zheng, S., Yu, F.-Y., and Chen, Y. 2021. The limits of multi-task peer prediction. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. EC '21. Association for Computing Machinery, New York, NY, USA, 907–926.