A Geometric Perspective on Minimal Peer Prediction†

RAFAEL FRONGILLO, CU Boulder
JENS WITKOWSKI, ETH Zurich

Minimal peer prediction mechanisms truthfully elicit private information (e.g., opinions or experiences) from rational agents without the requirement that ground truth is eventually revealed. In this paper, we use a geometric perspective to prove that minimal peer prediction mechanisms are equivalent to power diagrams, a type of weighted Voronoi diagram. Using this characterization and results from computational geometry, we show that many of the mechanisms in the literature are unique up to affine transformations. We also show that classical peer prediction is "complete" in that every minimal mechanism can be written as a classical peer prediction mechanism for some scoring rule. Finally, we use our geometric characterization to develop a general method for constructing new truthful mechanisms, and we show how to optimize for the mechanisms' effort incentives and robustness.

Categories and Subject Descriptors: J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Economics*; K.4 [**Computers and Society**]: Electronic Commerce

General Terms: Algorithms, Economics, Design

Additional Key Words and Phrases: Peer Prediction, Information Elicitation, Mechanism Design

ACM Reference Format:
Frongillo, Rafael and Witkowski, Jens 2017. A Geometric Perspective on Minimal Peer Prediction *ACM Trans. Econ. Comp.* V, N, Article A (May 2017), 27 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

User-generated content is essential to the effective functioning of many social computing and e-commerce platforms. A prominent example is eliciting information through crowdsourcing platforms, such as Amazon Mechanical Turk, where workers are paid small rewards to do so-called *human computation* tasks, which are easy for humans to solve but difficult for computers. For example, humans easily recognize celebrities, whereas even state-of-the-art computer vision algorithms perform significantly worse.

While statistical techniques can adjust for biases or identify noisy users, they are appropriate only in settings with repeated participation by the same user, and when user inputs are informative in the first place. But what if providing accurate information is costly for users, or if users have incentives to lie? Consider an image annotation task (e.g. for search engine indexing), where workers may wish to save effort by annotating with random words, or words that are too generic (e.g. "animal"). Or consider a public health program that requires participants to report whether they have ever used illegal drugs, and where participants may lie about their drug use due to shame or eligibility concerns.

†This paper is a significantly extended version of "A Geometric Method to Construct Minimal Peer Prediction Mechanisms," in the Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16).

Author's addresses: raf@colorado.edu (Rafael Frongillo) and jensw@inf.ethz.ch (Jens Witkowski).

DOI: http://dx.doi.org/10.1145/0000000.0000000

Peer prediction mechanisms address these incentive problems. They are designed to elicit truthful private information from self-interested participants, such as answers to the question "Have you ever used illegal drugs?" Crucially, peer prediction mechanisms cannot use ground truth. In the public health example this means the program cannot verify whether a participant has or has not ever used illegal drugs; it can only use the participants' voluntary reports.

The classical peer prediction method, introduced by Miller et al. [2005], addresses this challenge by comparing the reported information of a participant (her reported *signal*) with that of another participant, and computing a payment rule which ensures that truth revelation is a strategic equilibrium. The common knowledge assumptions in their analysis, however, are too strict for practical application. In particular, the participants' possible posterior beliefs (their *belief model*) are assumed to be known and the same for all participants. In the public health example, this would mean that every participant using drugs has the same belief that e.g. 40% of participants use drugs, and every participant not using drugs has the same belief that e.g. 20% use drugs. Moreover, these numbers need to be known by the mechanism in order to compute the payment rule. Since the assumption of perfectly known belief models is unrealistic in practice, Jurca and Faltings [2007] apply techniques from robust optimization to the classical peer prediction method and make it robust against small variations in the commonly-held belief model. As we will see in Section 7, the classical peer prediction method already does allow for some deviations from the assumption of perfect knowledge. Moreover, we show how a geometric perspective on peer prediction can be used to compute maximally-robust mechanisms for a generalized notion of robustness, which includes the robustness of Jurca and Faltings as a special case.

Bayesian Truth Serum (BTS) mechanisms relax the common knowledge assumptions of aforementioned mechanisms in that the belief model is still the same for all participants but no longer needs to be known to the mechanism (not even approximately as in Jurca and Faltings [2007]). In addition to the signal that is to be elicited, BTS mechanisms also require participants to report a probability distribution regarding the signal reports in the population. That is, they are not *minimal*. A shortcoming of the original Bayesian Truth Serum (BTS) [Prelec 2004] is that it is truthful only for an infinite number of participants. The Robust Bayesian Truth Serum (RBTS) [Witkowski and Parkes 2012a; Witkowski 2014] achieves truthfulness for small populations (any number of agents $n \geq 2$ with binary signals and any $n \geq 3$ for non-binary signals). The 1/p BTS [Radanovic and Faltings 2013] is truthful for non-binary signals and any number of agents $n \geq 2$ but is not robust towards prediction reports of 0% or 100% and has unbounded ex post payments.[1] It also requires a different belief model constraint than RBTS (see also Section 4).

All BTS mechanisms still assume that all participants share the same belief model. Witkowski and Parkes [2012b] relax this assumption and allow participants to have *subjective* belief models. (In the public health setting, for example, two participants using drugs could have different beliefs about the prevalance of drug users in the population.) However, their mechanism requires the ability to elicit relevant information from an agent both before and after she makes her observation. In particular, their mechanism is not minimal.

[1] Radanovic [2016] suggests a workaround for these shortcomings (p.33, Footnote 5), as follows. If the peer agent assigns 0% to a signal in her prediction report, then the participant is paid a signal score of 1 if and only if the participant and the peer agent both report this signal, and 0 else. However, in this minimal mechanism it is of course a dominant strategy to report the 0% signal independent of the participant's true belief. In contrast, even with 0% prediction reports, the signal scoring component of RBTS with positive delta results in a non-trivial mechanism, which is still strictly truthful for generic belief models.

Multi-task mechanisms are applicable to settings where agents respond to several "similar" questions, i.e. questions for which an agent has the same belief model. This approach allows for the design of minimal peer prediction mechanisms that do not require knowledge of the agents' belief models. Witkowski and Parkes [2013] score participants on a given task using signal distributions that were learned from reports on other tasks. The mechanism due to Dasgupta and Ghosh [2013], as well as its multi-signal extension [Shnayder et al. 2016], rewards agreement with the peer agent on overlapping questions and punish agreement on non-overlapping questions.

In the single-task minimal setting, a line of work introducing the 1/p Mechanism [Jurca and Faltings 2008; Jurca and Faltings 2011] and the Shadowing Method [Witkowski and Parkes 2012a; Witkowski 2014] allows for relaxed common knowledge assumptions when compared to the original work introducing the classical peer prediction method. As we show in this paper, any minimal mechanism, regardless of the knowledge assumptions used to construct it, allows for an analogous relaxation of common knowledge assumptions.

Our Results. In this paper, we provide a complete characterization of the design space of minimal, subjective, single-task peer prediction mechanisms, which includes the classical peer prediction method, Output Agreement, the 1/p Mechanism, and the Shadowing Method as special cases. While it was known that every minimal mechanism requires some constraint on the agents' belief models [Jurca and Faltings 2011], it was unknown which constraints allow for truthful mechanisms and how the constraints of different truthful mechanisms relate to one another. We answer these questions in Section 3 by adapting techniques from the literature on *property elicitation* [Lambert et al. 2008; Lambert and Shoham 2009; Frongillo and Kash 2014]. In particular, we prove the equivalence of minimal peer prediction mechanisms and power diagrams, a type of weighted Voronoi diagram.

In Section 4, we use this equivalence and results from computational geometry to show that all aforementioned mechanisms are the *unique* mechanisms, up to positive-affine transformations, which are truthful with respect to their belief model constraints. One important corollary of this is that maximizing effort incentives for any of these mechanisms reduces to computing the effort-optimal positive-affine transformation. In Section 5, we show how to construct new truthful mechanisms for new conditions, including a detailed example of this construction in Section 5.3.

We address the relative expressiveness of peer prediction mechanisms in Section 6. In particular, we prove that the classical peer prediction method is flexible enough to cover all possible mechanisms, in most cases even when restricting to the quadratic scoring rule. In Section 7, we show how to compute a mechanism that is maximally-robust with respect to deviations between the mechanism's and the agents' belief models. Finally, in Section 8 we give directions for ongoing and future work, including a data-driven non-minimal approach to peer prediction, where prediction reports are used to train classifiers (minimal mechanisms) to score signal reports.

2. PRELIMINARIES

In this section, we introduce the model, and review concepts in peer prediction and computational geometry.

2.1. Model

There is a group of $n \geq 2$ rational, risk-neutral, and self-interested agents. We sometimes use shorthand $[n] := \{1, \ldots, n\}$ to denote the set of agents. When interacting with

the environment, each agent $i \in [n]$ observes a signal S_i,[2] which is a random variable with values $[m] := \{1, \ldots, m\}$ for $m \geq 2$. The signal represents an agent's experience or opinion. The objective in peer prediction is to elicit an agent's signal in an incentive-compatible way, i.e. to compute payments such that agents maximize their expected payment by reporting their signal to the mechanism (center) truthfully.

To achieve this, all peer prediction mechanisms require that agent i's signal observation tells her something about the signal observed by another *peer* agent $j \neq i$. For example, this could be agent $j = i + 1 \,(\text{modulo } n)$, so that the agents form a "ring," where every agent is scored using the "following" agent. (Our results hold for any choice of peer agent.) Let then

$$p_i(s_j|s_i) = \Pr_i(S_j = s_j \mid S_i = s_i) \tag{1}$$

denote agent i's *signal posterior* belief that agent j receives signal s_j given agent i's signal s_i. We refer to $p_i(\cdot|\cdot)$ as agent i's *belief model*. A crucial assumption for the existence of strictly incentive compatible peer prediction mechanisms is that every agent's belief model satisfies *stochastic relevance* [Johnson et al. 1990].

Definition 2.1. Random variable S_i is *stochastically relevant* for random variable S_j if the distribution of S_j conditional on S_i is different for all possible values of S_i.

That is, stochastic relevance holds if and only if $p_i(\cdot|s_i) \neq p_i(\cdot|s_i')$ for all $i \in [n]$ and all $s_i' \neq s_i$. Intuitively, one can think of stochastic relevance as correlation between different agents' signal observations.

2.2. Peer Prediction Mechanisms

We are now ready to define peer prediction mechanisms. For a discussion of possible extensions to more general mechanisms, see Section 8.

Definition 2.2. A (minimal) *peer prediction mechanism* is a function $M : [m] \times [m] \to \mathbb{R}$, where $M(x_i, x_j)$ specifies the payment to agent i when she reports signal x_i and her peer agent j reports signal x_j.

We use ex post subjective equilibrium [Witkowski and Parkes 2012b], which is the most general solution concept for which truthful peer prediction mechanisms are known.

Definition 2.3. Mechanism M is *truthful* if we have

$$s_i = \operatorname*{argmax}_{x_i} \mathop{\mathbf{E}}_{S_j} \Big[M\big(x_i, S_j\big) \;\Big|\; S_i = s_i \Big] ,$$

for all $i \in [n]$ and all $s_i \in [m]$ with the expectation taken using agent i's belief model, i.e. $S_j \sim p_i(\cdot|s_i)$.

The equilibrium is *subjective* because it allows for each agent to have a distinct belief model, and *ex post* because it allows for (but doesn't require) knowledge of other agents' belief models. Agents only need to reason about other agents' signals, not their beliefs. One implication of this equilibrium concept is that $p_i(\cdot|\cdot)$ can be formulated after having seen the belief models of other agents. Note that while it may be the case that repeating this process eventually yields a common prior, ex post subjective equilibrium explicitly allows agents to "agree to disagree." Ex post subjective equilibrium is thus strictly more general than the classic definition of Bayes-Nash equilibrium

[2] We will drop the subscript to denote a generic signal.

(BNE) [Harsanyi 1968] as it coincides with BNE only when all agents share the same belief model, i.e. if $p_i(\cdot|\cdot) = p_j(\cdot|\cdot)$ for all $i, j \in [n]$.[3]

Definition 2.4. A mechanism M' is a *positive-affine transformation* of mechanism M if there exists $f : [m] \to \mathbb{R}$ and $\alpha > 0$ such that for all $x_i, x_j \in [m]$, $M'(x_i, x_j) = \alpha M(x_i, x_j) + f(x_j)$.

The importance of Definition 2.4 lies in the fact that if M is truthful, then M' is truthful as well. We state this as Lemma 2.5, which we prove in Appendix A. As we will see, in certain cases these are the only possible truthful mechanisms.

LEMMA 2.5. *Let M' be a positive-affine transformation of M. Then M' is truthful if and only if M is truthful.*

We conclude with several examples of peer prediction mechanisms from the literature to which we refer throughout the paper. The first are the Output Agreement Mechanism, the 1/p Mechanism and the Shadowing Method, which we now define. We will use the notation Δ_m to refer to the probability simplex, the set of probability distributions over m outcomes (see the following subsection).

Definition 2.6. *Output Agreement* is $M(x_i, x_j) = 1$ if $x_j = x_i$ and 0 otherwise.

Definition 2.7. The *1/p Mechanism* [Jurca and Faltings 2011] is given by $M(x_i, x_j) = \frac{1}{y(x_i)}$ if $x_i = x_j$ and 0 otherwise, for some fixed $\boldsymbol{y} \in \Delta_m$.

Definition 2.8. The *Shadowing Method* [Witkowski and Parkes 2012a; Witkowski 2014] is $M(x_i, x_j) = R_q(\boldsymbol{y}', x_j)$, where $\boldsymbol{y}' = \left(y(1) - \frac{\delta}{m-1}, \ldots, y(x_i) + \delta, \ldots, y(m) - \frac{\delta}{m-1}\right)$ for some fixed $\boldsymbol{y} \in \Delta_m$, $\delta > 0$, and $R_q(\boldsymbol{y}', x_j) = 2y'(x_j) - \sum_{k=1}^{m} y'(k)^2$ is the quadratic scoring rule [Brier 1950].

THEOREM 2.9. *The preceding mechanisms are truthful if and only if the following conditions are satisfied for all $s \in [m]$, $s' \neq s$, and for all $i \in [n]$:*

(1) *Output Agreement:* $p_i(s|s) > p_i(s'|s)$
(2) *1/p Mechanism:* $p_i(s|s)/y(s) > p_i(s'|s)/y(s')$
(3) *Shadowing Method:* $p_i(s|s) - y(s) > p_i(s'|s) - y(s')$.

The proofs for each of these statements are in the respective papers cited in each mechanism's definition above. Note that Theorem 2.9 does not say whether or not there are other mechanisms which are truthful under the constraints (1), (2), and (3). We will answer this question in Theorem 4.1.

Finally, we will refer to the Classical Peer Prediction Method in Section 6, which relies on the notion of a *proper scoring rule*, a tool for eliciting probabilistic beliefs from agents given a sample from the ground truth. In short, a proper scoring rule is a function $R : \Delta_m \times [m] \to \mathbb{R}$ such that $\mathbf{E}_p[R(p, X)] \geq \mathbf{E}_p[R(q, X)]$ where X has distribution p, and $q \neq p$ is a non-truthful report. When the inequality is always strict, we say R is *strictly proper*. (One example of a strictly proper scoring rule is the quadratic scoring rule due to Brier [1950], given in Definition 2.8.) As we note in Section 6, all proper scoring rules can be given in terms of convex functions; we refer the reader to that section and to Gneiting and Raftery [2007] for details.

Definition 2.10. Given strictly proper scoring rule R and belief model $p_i(\cdot|\cdot)$, the *Classical Peer Prediction Method* [Miller et al. 2005] is $M(x_i, x_j) = R\big(p_i(\cdot|x_i), x_j\big)$.

[3]More modern definitions of BNE include subjectivity of beliefs, e.g. [Osborne and Rubinstein 1994], and our definition can be seen as a special case where agents have an arbitrary prior over their own signal.

THEOREM 2.11. *The Classical Peer Prediction Method is truthful if and only if* $p_i(\cdot|\cdot)$ *satisfies stochastic relevance for all* $i \in [n]$.

2.3. The Probability Simplex

The intuition for our main results can be provided for $m = 3$ signals already, and so we give such examples throughout the paper. For probability distributions over only 3 signals, there is a convenient graphical representation of the *probability simplex* Δ_m as an equilateral triangle, where the three corners represent the signals (see Figure 1L). The closer a point is to a corner (the distance from the corner's opposing side), the more probability mass of that corner's signal is on that point.[4] The triangular shape ensures that for any point on the triangle the values of the three dimensions sum up to 1. For example, the point $p_i(\cdot|b) = (0.2, 0.75, 0.05)$ in Figure 1L is at height $1/20$ (since the top corner represents signal c), and one fifth away from the right side of the triangle (because the left corner represents signal a). Observe that with three signals, there are only two degrees of freedom, and so fixing the point's position with respect to a and c, the value for signal b is fixed as well. (Confirm that $p_i(\cdot|b)$ is three fourths away from the left side.)

2.4. Power Diagrams

Our results rely on a concept from computational geometry known as a *power diagram*, which is a type of weighted Voronoi diagram [Aurenhammer 1987b].

Definition 2.12. A *power diagram* is a partitioning of Δ_m into sets called *cells*, defined by a collection of points $\{\boldsymbol{v}^s \in \mathbb{R}^m : s \in [m]\}$ called *sites* with associated weights $w(s) \in \mathbb{R}$, given by

$$\text{cell}(\boldsymbol{v}^s) = \left\{ \boldsymbol{u} \in \mathbb{R}^m : \{s\} = \operatorname*{argmin}_{x \in [m]} \{\|\boldsymbol{u} - \boldsymbol{v}^x\|^2 - w(x)\} \right\}. \tag{2}$$

We call $\|\boldsymbol{u} - \boldsymbol{v}^x\|^2 - w(x)$ the *power distance* from $\boldsymbol{u}$ to site $\boldsymbol{v}^x$; thus, for every point $\boldsymbol{u}$ in $\text{cell}(\boldsymbol{v}^s)$, it holds that $\boldsymbol{v}^s$ is closer to $\boldsymbol{u}$ in power distance than any other site $\boldsymbol{v}^x$.

We have defined power diagrams for the special case of the probability simplex, which is the case we need in this paper. The more general definition allows for a different number of sites than dimensions. Also, note that we exclude cell boundaries by ensuring that we have a unique minimizer in eq. (2); in the following we drop the set notation around s and just write $s = \text{argmin}\{\ldots\}$ or $s = \text{argmax}\{\ldots\}$. The usual definition of a Voronoi diagram follows by setting all weights $w(s)$ to 0.

3. MECHANISMS AND POWER DIAGRAMS

As with previous work, we would like to make statements of the form, "As long as the belief models satisfy certain constraints, the mechanism is truthful." For example, the Shadowing Method (Definition 2.8) is truthful if and only if $p_i(s|s) - y(s) > p_i(s'|s) - y(s')$ for all $s, s' \in [m] : s' \neq s$, all $i \in [n]$, and some distribution $\boldsymbol{y}$, which is a parameter of the mechanism. When used directly, and not as a building block for more complex mechanisms, it is often assumed that there is a known, common signal prior, which is then used as $\boldsymbol{y}$. As we will see, both the Shadowing Method and the 1/p mechanism [Jurca and Faltings 2008; Jurca and Faltings 2011] are actually robust in that they are truthful even if there is no common signal prior. All that is required is that the agents' possible posteriors fall into the correct regions. While it has been known that the constraints required by the Shadowing Method and the 1/p mechanism are incomparable,

[4] This is equivalent to the natural embedding into $\mathbb{R}^3$ and viewing in the direction $(-1, -1, -1)$.

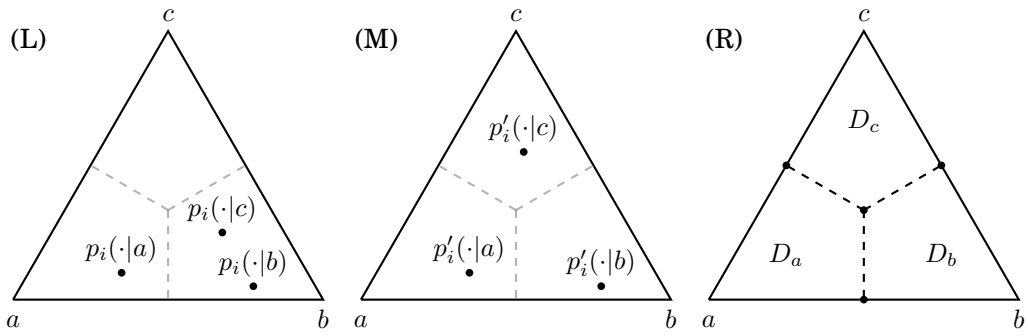

Fig. 1: Belief models for which OA is not truthful (L) and is truthful (M). The belief models for which OA is truthful are characterized by the cells D_a, D_b, D_c, where the posterior following signal s must lie in D_s (R).

i.e. there exist belief models for which the Shadowing Method is truthful but the 1/p mechanism is not, and vice versa [Witkowski 2014], it was not known for which constraints there exist truthful mechanisms. In this section, we answer this question, and characterize all belief model constraints for which truthful mechanisms exist.

3.1. Belief Model Constraints

To begin, consider the Output Agreement mechanism (OA), with two possible belief models $p_i(\cdot|\cdot)$ and $p_i'(\cdot|\cdot)$ as depicted in Figures 1L and 1M, respectively: $p_i(\cdot|a) = p_i'(\cdot|a) = (0.6, 0.3, 0.1)$, $p_i(\cdot|b) = p_i'(\cdot|b) = (0.2, 0.75, 0.05)$, $p_i(\cdot|c) = (0.2, 0.55, 0.25)$, and $p_i'(\cdot|c) = (0.2, 0.25, 0.55)$. For which of the belief models $p_i(\cdot|\cdot), p_i'(\cdot|\cdot)$ is OA truthful? Upon inspection, one can verify that OA is truthful with respect to $p_i'(\cdot|\cdot)$ but not $p_i(\cdot|\cdot)$. The key point is that upon observing signal c, $p_i(b|c) = 0.55 > p_i(c|c)$, meaning under $p_i(\cdot|\cdot)$, the agent thinks that the peer agent is most likely to have seen (and reported) b, so signal b will be the optimal report. Clearly then, OA is not truthful with respect to every belief model, and one may naturally ask, for which belief models is OA truthful? As it turns out, one can describe all such belief models using a very simple *belief model constraint*, of the form $p_i(\cdot|s) \in D_s$ for all signals s, for some sets of distributions D_s. For OA, this constraint is given by the sets $D_s = \{p \,|\, \forall s' \neq s,\ p(s) > p(s')\}$, as depicted in Figure 1R. We will revisit this constraint in Section 3.2.

We now formally define these constraints on belief models that limit which posteriors are possible following which signal.

Definition 3.1. A *belief model constraint* is a collection $\mathcal{D} = \{D_s \subseteq \Delta_m : s \in [m]\}$ of disjoint sets D_s of distributions. If additionally we have $\text{cl}(\cup_s D_s) = \Delta_m$, i.e. if $\mathcal{D}$ partitions the simplex, we say $\mathcal{D}$ is *maximal*.

A belief model constraint $\mathcal{D} = \{D_1, \ldots, D_m\}$ ensures that for each agent i, following signal observation $S_i = s_i$, her belief about her peer agent's signal s_j is restricted to be in D_{s_i}. It is easy to come up with non-maximal belief model constraints, such as "$\forall s\ p(s|s) > 0.6$".[5] Note that under such a constraint, some distributions are not valid posteriors for any signal. In contrast, a maximal constraint covers the simplex, partitioning it into m bordering but non-overlapping regions (e.g. Figure 1R).

We can now talk about mechanisms being truthful with respect to a belief model constraint, which ensures that agents will report honestly as long as their belief model satisfies the constraint.

[5] See Figure 4M for an illustration, as well as the surrounding discussion.

Definition 3.2. A mechanism $M(\cdot,\cdot)$ is truthful with respect to belief model constraint $\mathcal{D}$ if M is truthful whenever $p_i(\cdot|s) \in D_s$ for all agents $i \in [n]$ and all signals $s \in [m]$.

It directly follows from this perspective that all minimal peer prediction mechanisms require a belief model constraint.[6] Consider, for example, a posterior belief $p_i(\cdot|s) = (3/5, 3/20, 1/4)$. Without any constraint on the belief model, it is not clear if this is the posterior following signal 1, 2, or 3. This choice needs to be made since a given posterior belief can only belong to one signal (stochastic relevance, Definition 2.1), and so every truthful minimal peer prediction mechanism requires a belief model constraint.

One very natural constraint to consider is to take an arbitrary mechanism M and restrict to only those belief models under which M is truthful. It turns out that this set can be succinctly described by a belief model constraint, which we call the constraint *induced* by M. Moreover, the regions of this induced constraint must take a particular shape, that of a power diagram, and conversely, *every* power diagram is an induced constraint of some mechanism.

We will now observe that for any mechanism M, there is a belief model constraint $\mathcal{D}^M$, which exactly captures the set of belief models for which M is truthful. In other words, not only is M truthful with respect to $\mathcal{D}^M$, but under any belief model that does not satisfy $\mathcal{D}^M$, M will not be truthful. The construction of $\mathcal{D}^M$ is easy: for each signal s, D_s^M is the set of distributions $p(\cdot|s)$ under which $x_i = s$ is the unique optimal report for M. Note that if D_s^M is empty for any s, then M is not truthful for any belief model.

LEMMA 3.3. *Let $M : [m] \times [m] \to \mathbb{R}$ be an arbitrary mechanism, and let $\mathcal{D}^M$ be the belief model constraint given by*

$$D_s^M = \left\{ p_i(\cdot|s) : s = \operatorname*{argmax}_{x_i} \mathop{\mathbf{E}}_{S_j \sim p_i(\cdot|s)} M(x_i, S_j) \right\} . \tag{3}$$

Then M is truthful with respect to $\mathcal{D}^M$, but not truthful for belief models not satisfying $\mathcal{D}^M$. Moreover, if the rows of M are all distinct, $\mathcal{D}^M$ is maximal.

PROOF. Suppose $p_i(\cdot|s) \in D_s^M$ for all $i \in [n], s \in [m]$. Then by construction of D_s^M, an agent i receiving signal s maximizes expected payoff by reporting s, and hence M is truthful. By definition then, M is truthful with respect to $\mathcal{D}^M$. Now suppose $p_i(\cdot|s) \notin D_s^M$ for some $i \in [n], s \in [m]$. Then $s \neq \operatorname{argmax}_{x_i} \mathbf{E}_{S_j \sim p_i(\cdot|s)} M(x_i, S_j)$, and thus M cannot be truthful. Finally, consider the function $G(p) = \max_x \mathbf{E}_{S_j \sim p} M(x, S_j)$, which is convex as a pointwise maximum of linear functions. By standard results in convex analysis (c.f. [Frongillo and Kash 2014, Theorem 3]) G has subgradient $M(x,\cdot)$ whenever x is in the argmax. As the rows of M are all distinct, multiple elements in the argmax corresponds to multiple subgradients of G, and thus G is nondifferentiable[7] at the set of indifference points $\{p : |\operatorname{argmax}_x \mathbf{E}_{S_j \sim p} M(x, S_j)| \geq 2\}$. As G is convex, it must be differentiable almost everywhere [Aliprantis and Border 2007, Theorem 7.26], these indifference points must have measure 0 in the probability simplex, and thus $\mathcal{D}^M$ is maximal. □

The construction above in Eq. 3 shows even more than the fact that $\mathcal{D}^M$ is maximal: the cells D_s^M must be convex sets. (One can see this directly by taking two posteriors for which s is the argmax, and observing that s must be the argmax for any mixture

[6] Jurca and Faltings [2011] were the first to state that no minimal mechanism can be truthful for all stochastically relevant belief models. However, they did not study the type of constraints that need to be imposed to allow for truthful mechanisms.

[7] Technically, we should restrict to the first $m-1$ coordinates of the distribution, or use the approach of Appendix B, so that G is defined on a full-dimensional subset of $\mathbb{R}^{m-1}$. Neither alters the argument.

of the two.) A connection to property elicitation in the next subsection will reveal even more structure: $\mathcal{D}^M$ must be the cells of a power diagram.

3.2. Relationship to Finite Property Elicitation

Our results rely heavily on a novel connection from minimal peer prediction mechanisms to the literature on property elicitation, where one wishes to extract a particular function, or *property*, of an agent's belief using a scoring rule with access to a single sample from the true distribution (unlike our setting, where no ground truth is ever observed). Formally, a scoring rule $S(\cdot,\cdot)$ elicits a property Γ if for all agent beliefs p, the expected score $\mathbf{E}_{x\sim p}[S(r,x)]$ is maximized by the report $r=\Gamma(p)$. In particular, we leverage results from the *finite property* case, where the reports r are restricted to a finite set [Lambert and Shoham 2009]. For example, the *mode* of a distribution over m possible outcomes, $\Gamma_{\text{mode}}(p)=\operatorname{argmax}_x p(x)$, has m possible values, and is elicited by the scoring rule $S(r,x)=\mathbb{1}\{r=x\}$ where $\mathbb{1}\{\}$ is the indicator function. (This follows by writing out the expected score $\mathbf{E}_{x\sim p}[S(r,x)]=\mathbf{E}_{x\sim p}[\mathbb{1}\{r=x\}]=p(r)$, which is maximized when r is equal to the mode of p.)

In peer prediction, one can view belief model constraints as a kind of finite property in the above sense. Let $\mathcal{D}$ be a maximal belief model constraint, and let us encode it in a property $\Gamma^{\mathcal{D}}:\Delta_m\to[m]$ by taking $\Gamma^{\mathcal{D}}(p)=s$ if $p\in D_s$. (Strictly speaking, this does not describe $\Gamma^{\mathcal{D}}$ everywhere as the cells D_s are open and disjoint; we defer the details to the proofs in Section 4.)

Now let M be a minimal mechanism. If M is truthful for $\mathcal{D}$, then M is truthful with respect to any belief model satisfying $\mathcal{D}$. In particular, if $q(\cdot)$ is agent i's posterior following some signal, then $q\in D_s$ for some s (again, ignoring boundaries); thus by the belief model constraint $\mathcal{D}$, q must be the posterior following s, and by M being truthful for $\mathcal{D}$, agent i will report s. Summarizing, we have just said that whenever an agent has belief q, the report maximizing their expected score is the s such that $q\in D_s$, which is $\Gamma^{\mathcal{D}}(q)$ by definition. Thus, M being truthful for $\mathcal{D}$ implies that M, thought of as a scoring rule, elicits $\Gamma^{\mathcal{D}}$. The converse follows by exactly the same logic, where now M eliciting Γ implies that M is truthful for $\mathcal{D}$ defined by the constraint $\Gamma(p_i(\cdot|s))=s$, which is exactly the induced constraint $\mathcal{D}^M$ as one can verify from Eq. 3.

Now that we know minimal mechanisms elicit finite properties, we can ask which properties are elicited by popular mechanisms in the literature. It follows immediately from the above discussion that Output Agreement, $M(x_i,x_j)=\mathbb{1}\{x_i=x_j\}$, elicits the mode $\Gamma_{\text{mode}}(p)=\operatorname{argmax}_s p(s)$. Similarly, the Shadowing Method elicits a shifted mode, $\Gamma(p)=\operatorname{argmax}_s p(s)-y(s)$, that is, the mode of the "distribution" $q(s)=p(s)-y(s)$. (Note that q may lie outside the simplex Δ_m.) Finally, the 1/p mechanism can be thought of as eliciting a "dampened" mode $\Gamma(p)=\operatorname{argmax} p(s)/y(s)$, which again is the mode of "distribution" $q(s)=p(s)/y(s)$.

3.3. Equivalence to Power Diagrams

Leveraging the connection to finite properties established above, we can now use results in the property elicitation literature, which show an equivalence between elicitable finite properties and power diagrams [Lambert and Shoham 2009; Frongillo and Kash 2014], and extend this equivalence to minimal peer prediction mechanisms. From there, techniques from computational geometry allow us to show further structure, as we explore in Sections 4 and 6. To simplify exposition, we will show the correspondence to power diagrams directly, without reference to finite properties, though properties will appear again in the proofs.

We have seen that every mechanism M induces some belief model constraint $\mathcal{D}^M$, and that M is truthful with respect to $\mathcal{D}^M$. We now show further that $\mathcal{D}^M$ is a power

diagram, and conversely, that every power diagram has a mechanism such that $D_s^M = \text{cell}(\boldsymbol{v}^s)$ for all s.

The concrete mapping is as follows. Given mechanism $M : [m] \times [m] \to \mathbb{R}$, we construct sites and weights by:

$$\boldsymbol{v}^s = M(s, \cdot), \quad w(s) = \|\boldsymbol{v}^s\|^2 = \|M(s, \cdot)\|^2 \,. \tag{4}$$

Conversely, given a power diagram with sites $\boldsymbol{v}^s$ and weights $w(s)$, we construct the mechanism M as follows:

$$M(x_i, x_j) = \boldsymbol{v}^{x_i}(x_j) - \tfrac{1}{2}\|\boldsymbol{v}^{x_i}\|^2 + \tfrac{1}{2}w(x_i) \,, \tag{5}$$

where $\boldsymbol{v}^{x_i}(x_j)$ is the x_jth entry of $\boldsymbol{v}^{x_i}$. We note that these formulas are more explicit versions of those appearing in property elicitation, as mentioned above.

With these conversions in hand, we can now show that they indeed establish a correspondence between minimal peer prediction mechanisms and power diagrams.

THEOREM 3.4. *Given any mechanism $M : [m] \times [m] \to \mathbb{R}$, the induced belief model constraint $\mathcal{D}^M$ is a power diagram. Conversely, for every power diagram given by sites $\boldsymbol{v}^s$ and weights $w(s)$, there is a mechanism M whose induced belief model constraint $\mathcal{D}^M$ satisfies $D_s^M = \text{cell}(\boldsymbol{v}^s)$ for all s.*

PROOF. Observe that if either relation (4) or (5) holds, we have the following for all $x, \boldsymbol{p}$:

$$-2\boldsymbol{p} \cdot \boldsymbol{v}^x + \|\boldsymbol{v}^x\|^2 - w(x) = -2 \mathop{\mathbf{E}}_{S_j \sim \boldsymbol{p}} [M(x, S_j)] \,. \tag{6}$$

To see this, note that $\boldsymbol{p} \cdot M(x, \cdot) = \mathbf{E}_{S_j \sim \boldsymbol{p}}[M(x, S_j)]$. Adding $\|\boldsymbol{p}\|^2$ to both sides of Eq. 6 gives

$$\|\boldsymbol{p} - \boldsymbol{v}^x\|^2 - w(x) = \|\boldsymbol{p}\|^2 - 2 \mathop{\mathbf{E}}_{S_j \sim \boldsymbol{p}} [M(x, S_j)] \,. \tag{7}$$

Now applying Eq. 7 to the definitions of a power diagram and of $\mathcal{D}^M$, we have

$$\begin{aligned}
\boldsymbol{p} \in \text{cell}(\boldsymbol{v}^s) &\iff s = \operatorname*{argmin}_x \{\|\boldsymbol{p} - \boldsymbol{v}^x\|^2 - w(x)\} \\
&\iff s = \operatorname*{argmin}_x \Big\{ \|\boldsymbol{p}\|^2 - 2 \mathop{\mathbf{E}}_{S_j \sim \boldsymbol{p}} [M(x, S_j)] \Big\} \\
&\iff s = \operatorname*{argmax}_x \mathop{\mathbf{E}}_{S_j \sim p(\cdot|s)} M(x, S_j) \\
&\iff \boldsymbol{p} \in D_s^M \,.
\end{aligned}$$

Finally, as Eq. 4 defines a power diagram for any mechanism M, and Eq. 5 defines a mechanism for any power diagram, we have established our equivalence. □

COROLLARY 3.5. *Let $\mathcal{D}$ be a maximal belief model constraint. Then there exists a mechanism that is truthful with respect to $\mathcal{D}$ if and only if $\mathcal{D}$ is a power diagram.*

Corollary 3.5 gives us considerable power in determining whether or not a belief model constraint can yield a truthful mechanism. On the positive side, Figure 2L shows a maximal belief model constraint that is a power diagram, and thus there must be a truthful mechanism (in this case, the Shadowing Method). On the negative side, we can leverage known attributes of power diagrams, such as the fact that their cells are convex, their boundaries are linear/affine, and moreover the boundary between two cells must be perpendicular to the line segment connecting the corresponding sites. Figure 2M shows a maximal constraint in which the D_a cell is nonconvex, and thus cannot be a power diagram; we conclude there can be no mechanism truthful

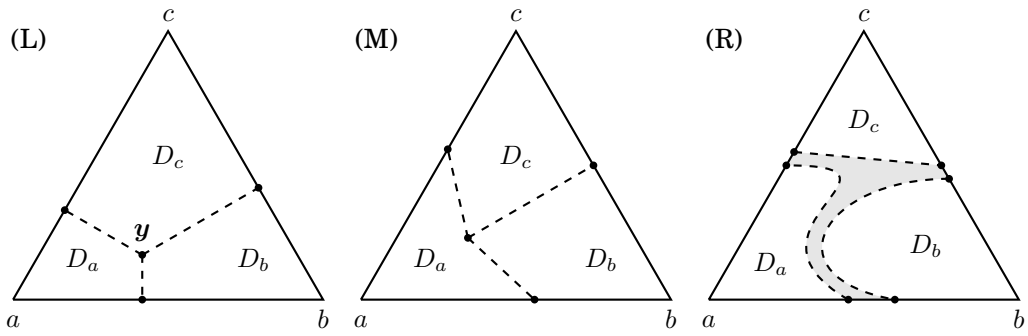

Fig. 2: **(L)** The regions D_s are those for which the Shadowing Method incentivizes the agents to report the respective signal s. For example, for any posterior belief falling into D_c, agent i should report $x_i = c$. To see that the depicted partitioning is indeed coming from this constraint, first observe that there is indifference at the intersection point, i.e. when $p_i(\cdot|s) = y(\cdot)$, and that the constraints are linear, so that the indifference borders are lines. The only remaining piece is then to determine the points for each pair of signals, where the third (left-out) signal's weight is 0, and draw a line between that point and $\boldsymbol{y}$. For example, the indifference point between signals b and c, where a has no weight is $(0, 7/12, 5/12) \Leftrightarrow p_i(c|c) - 1/6 = p_i(b|c) - 1/3; p_i(a|c) = 0$. **(M)** A maximal belief model constraint which is not a power diagram and hence for which there is no truthful mechanism. **(R)** A non-maximal belief model constraint for which there is still no truthful mechanism, because there is no power diagram consistent with the constraints.

with respect to this constraint.[8] Finally, Figure 2R shows a constraint which is not maximal, but which is not consistent with any power diagram (using the linearity of cell boundaries), and hence there still cannot be a truthful mechanism. These examples illustrate the power of this geometric approach in determining whether truthful mechanisms exist, even for non-maximal constraints.

Finally, it is easy to see that the conversion from mechanisms to sites and weights of power diagrams (Eq. 4) and back (Eq. 5) are inverse operations. In the following section, we will leverage this tight connection, and use results from computational geometry to show that several well-known mechanisms are unique in the sense that they are the only mechanisms, up to positive-affine transformations, that are truthful for their respective belief model constraints.

4. UNIQUENESS

Consider again the the Output Agreement mechanism $M(x_i, x_j) = 1$ if $x_j = x_i$ and 0 otherwise. It is easy to see that the mechanism is truthful as long as each agent assigns the highest posterior probability $p_i(\cdot|s_i)$ to their own signal s_i, yielding the constraint "$p(s|s) > p(s'|s)\ \forall s' \neq s$." The type of question we address in this section is: are there any other mechanisms than M that are guaranteed to be truthful as long as posteriors satisfy this condition? We will show that, up to positive-affine transformations, the answer is no: output agreement is unique. Moreover, the Shadowing Method and the 1/p Mechanism are also unique for their respective conditions on posteriors. Note that we identify mechanisms with their matrices, so that uniqueness does not preclude that parameterized mechanisms, such as the 1/p Mechanism and the Shadowing Method, coincide for particular choices of parameters. In fact, for uniform y, both the 1/p Mech-

[8]It is worth noting that even if the D_a was adjusted to be straight (forming a "T" intersection with the c–b boundary), it would still not be a power diagram; even though the D_a cell would now be convex, the diagram would violate the perpendicularity condition (this is not immediately obvious, but becomes clear with experimentation).

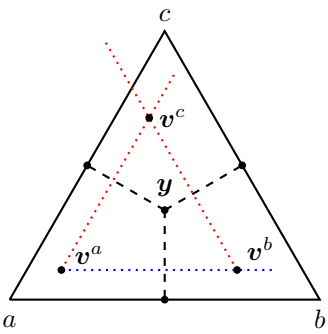

Fig. 3: Constructing a mechanism that is truthful under the same conditions as output agreement.

anism and the Shadowing Method coincide with Output Agreement in the sense that their belief model constraints are identical. (See Section 6.)

To get some intuition for this result, let us see why output agreement is unique for $m = 3$ signals a, b, c. From Theorem 3.4, we know that $\mathcal{D} = \mathcal{D}^M$, the induced belief model constraint, is a power diagram (which is depicted in Figure 3). In general, there may be many sites and weights that lead to the same power diagram, and these may yield different mechanisms via Eq. 5. In fact, it is a general result that any positive scaling of the sites followed by a translation (i.e. some $\alpha > 0$ and $\boldsymbol{u} \in \mathbb{R}^m$ so that $\hat{\boldsymbol{v}}^s = \alpha \boldsymbol{v}^s + \boldsymbol{u}$ for all s) will result in the same power diagram for an appropriate choice of weights [Aurenhammer 1987b]. As it turns out, such scalings and translations exactly correspond to positive-affine transformations when passing to a mechanism through Eq. 5. Thus, we only need to show that the sites for the output agreement power diagram are unique up to scaling and translation. Here, another useful property of sites comes into play: the line between sites of two adjacent cells must be perpendicular to the boundary between the cells. Examining Figure 3, one sees that after fixing $\boldsymbol{v}^a$, the choice of $\boldsymbol{v}^b$ is constrained to be along the blue dotted line, and once $\boldsymbol{v}^a$ and $\boldsymbol{v}^b$ are chosen, $\boldsymbol{v}^c$ is fixed as the intersection of the red dotted lines. Thus, we can specify the sites by choosing $\boldsymbol{v}^a$ (a translation), and how far away from $\boldsymbol{v}^a$ to place $\boldsymbol{v}^b$ (a positive scaling). We can then conclude that output agreement for three signals is unique up to positive-affine transformations. We now give the general result.

THEOREM 4.1. *If there exists a mechanism M that is truthful for some maximal belief model constraint $\mathcal{D}$, and there is some $\boldsymbol{y} \in \Delta_m$ with $y(s) > 0 \ \forall s$ such that $\cap_s \mathrm{cl}(D_s) = \{\boldsymbol{y}\}$, then M is the unique truthful mechanism for $\mathcal{D}$ up to positive-affine transformations.*

PROOF. As M is truthful with respect to $\mathcal{D}$, we have $\mathcal{D} = \mathcal{D}^M$ and thus $\mathcal{D}$ is a power diagram P from Theorem 3.4. By our assumption, we observe that the only vertex (also called a 0-dimensional face) of P must be $\boldsymbol{y}$, the intersection of the $\mathrm{cl}(D_s)$, as there are m cells but Δ_m has dimension $m - 1$. Throughout the proof, as before, we implicitly work in the affine hull $\mathcal{A}$ of Δ_m. We may assume without loss of generality that all sites lie in $\mathcal{A}$, as we may translate any site to $\mathcal{A}$ while adjusting its weight to preserve the diagram and resulting mechanism (see Appendix B). Thus, by definition of a *simple cell complex*, [9] as the vertex $\boldsymbol{y}$ is in the relative interior of Δ_m, we see that the extension $\hat{P}$ of P onto the affine hull of Δ_m must also have $\boldsymbol{y}$ as the only

[9] A cell complex P is *simple* if each of P's vertices is a vertex of exactly m cells of P, the minimum possible [Aurenhammer 1987a, p.50].

vertex. (As a vertex is on the boundary of every cell, and there are m cells in $m-1$ dimensions with disjoint interiors, there can be only one such point.) Now following the proof of Frongillo and Kash [2014, Theorem 4], we note that Aurenhammer [1987b, Lemma 1] and Aurenhammer [1987a, Lemma 4] together imply the following: if $\hat{P}$ is represented by sites $\{\boldsymbol{v}^s\}_{s\in[m]}$ and weights $w(\cdot)$, then any other representation of $\hat{P}$ with sites $\{\hat{\boldsymbol{v}}^s\}_{s\in[m]}$ satisfies $\exists \alpha > 0, \boldsymbol{u} \in \mathbb{R}^m$ s.t. $\hat{\boldsymbol{v}}^s = \alpha \boldsymbol{v}^s + \boldsymbol{u}$ for all $s \in [m]$. In other words, all sites must be a translation and scaling of $\{\boldsymbol{v}^s\}$. To complete the proof, we observe that different choices of $\boldsymbol{u}$ and α (with suitable weights) merely yield an affine transformation of M when passed through Eq. 5, and as any positive-affine transformation preserves truthfulness, the result follows. □

As a Corollary of Theorem 4.1, we can now prove a stronger version of Theorem 2.9. Not only are the three mechanisms truthful with respect to the corresponding constraints, but they are the *only* such mechanisms, up to positive-affine transformations.

COROLLARY 4.2. *The following mechanisms are unique, up to positive-affine transformations, with respect to the corresponding constraints (each with "$\forall s' \neq s$" and "$\forall i \in [n]$" implied):*

(*1*) ***Output Agreement,*** $p_i(s|s) > p_i(s'|s)$
(*2*) ***1/p Mechanism,*** $p_i(s|s)/y(s) > p_i(s'|s)/y(s')$
(*3*) ***Shadowing Method,*** $p_i(s|s) - y(s) > p_i(s'|s) - y(s')$.

PROOF. In all three cases, the given mechanism is known to be truthful for its respective belief model constraint. Moreover, for all three constraints $\mathcal{D}$, one can check that $\cup_s \mathrm{cl}(D_s) = \Delta_m$ and $\cap_s \mathrm{cl}(D_s) = \{\boldsymbol{y}\}$ meaning that $\boldsymbol{y}$ is the unique distribution bordering every set D_s (for Output Agreement, $\boldsymbol{y}$ is the uniform distribution). Hence, the mechanisms are unique up to positive-affine transformations by Theorem 4.1. □

Let us make a few remarks on Theorem 4.1. First, note that the restriction that $\boldsymbol{y}$ must not touch the boundary of the simplex is necessary, as uniqueness will not hold otherwise; see Figure 4L. Similarly, for constraints $\mathcal{D}$ that are not maximal, there may be many more truthful mechanisms; Figure 4M depicts two distinct power diagrams yielding mechanisms that are truthful with respect to the non-maximal constraint "$p(s|s) > 0.6$", and thus they are not merely positive-affine transformations of each other. That said, some non-maximal constraints $\mathcal{D}$ still yield a unique truthful mechanism, as illustrated in Figure 4R. In Section 6, we strengthen the results of this section, showing that under the same conditions as Theorem 4.1, not only is the mechanism unique up to positive-affine transformations, but it can be expressed as a positive-affine transformation of a classical peer prediction mechanism with respect to the quadratic scoring rule. This is done by first showing that in this setting, the maximal constraint $\mathcal{D}$ is not just a power diagram but in fact a Voronoi diagram.

5. CONSTRUCTING MECHANISMS FOR NEW CONDITIONS

In this section, we show how to compute new mechanisms that are truthful with respect to new conditions. Moreover, we find the positive-affine transformation that maximizes effort incentives subject to a budget. It then follows directly from Theorem 4.1 that the final mechanism's effort incentives are globally optimal given this condition. That is, there is no peer prediction mechanism that is truthful with respect to the new condition providing better effort incentives.

5.1. Computing Truthful Mechanisms

To construct mechanisms from belief model constraints, we turn to computational geometry to find sites and weights for the corresponding power diagram. For the class

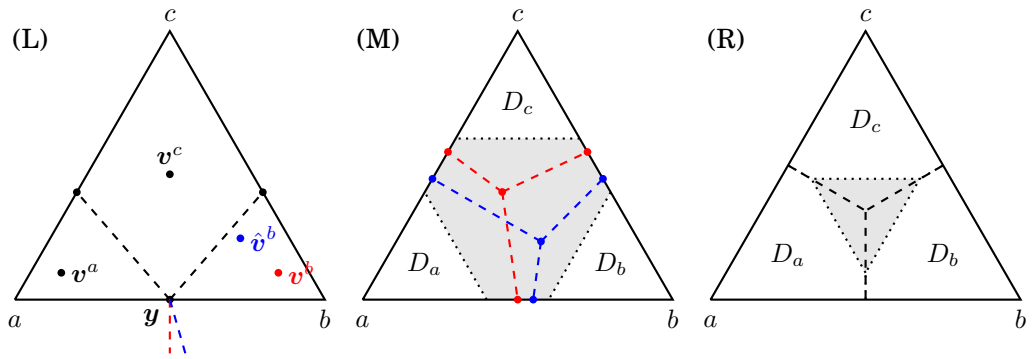

Fig. 4: Illustrations of the conditions of Theorem 4.1. **(L)** An example showing why the condition $\forall s\ \boldsymbol{y}(s) > 0$ is necessary. In this example, there are two sets of sites and weights that represent the same power diagram yet are not translations and scalings of each other. Here, all sites remain the same except $\boldsymbol{v}^b$. **(M)** A non-maximal constraint, with two consistent power diagrams (red and blue). **(R)** An example where the constraint is non-maximal (here the grey region belongs to none of the cells) yet there is a unique truthful mechanism (in this case, Output Agreement).

of constraints satisfying Theorem 4.1, where there is an intersection point $\boldsymbol{y}$ in the interior of the simplex, which is the most common case in peer prediction, one can rely on the $O(m)$ time algorithm given by Aurenhammer [1987c]. [10] We now give a version of this procedure, specialized for our setting.

The first consideration is the form of the input to the procedure. How should we assume the belief model constraint $\mathcal{D}$ is given? One succinct representation is an $m \times m$ matrix representing a truthful mechanism for the constraint, but of course specifying a constraint with a mechanism defeats the whole purpose of the algorithm. To be useful, we need a representation for the belief model constraint that is closer to the partitioning point of view. Here, we assume that one knows the constraint boundary between any pair of signals (s, s'). If $\mathcal{D}$ is a power diagram (a necessary condition for the existence of a truthful mechanism by Theorem 3.4) this boundary must be defined by a hyperplane, which in turn can be represented by its defining normal vector $\boldsymbol{u}^{s,s'}$. All that remains is specifying the orientation of $\boldsymbol{u}^{s,s'}$ so we know which direction is which cell; here we assume $\boldsymbol{u}^{s,s'}$ points in the direction away from D_s, so that we have $\boldsymbol{p} \in D_s \iff \boldsymbol{u}^{s,s'} \cdot \boldsymbol{p} < 0$ for all $s' \neq s$. As an example, the shadowing constraint is given by $\boldsymbol{u}^{s,s'} = \mathbb{1}_{s'} - \mathbb{1}_s - (y(s') - y(s))\mathbb{1}$, where $\mathbb{1}_s$ is the standard unit vector (with 1 in coordinate s and 0 otherwise), and $\mathbb{1}$ is the all-ones vector. Hence, $\boldsymbol{u}^{s,s'} \cdot \boldsymbol{p} < 0 \iff p(s') - p(s) - y(s') + y(s) < 0$, since $\mathbb{1} \cdot \boldsymbol{p} = 1$. We give another example in Section 5.3.

Thus, given any maximal belief model constraint $\mathcal{D}$, we can represent $\mathcal{D}$ by a collection of vectors $\{\boldsymbol{u}^{s,s'} \in \mathbb{R}^m : s, s' \in [m]\}$. With this representation in hand, the following algorithm computes a truthful mechanism for $\mathcal{D}$, under the additional assumption that $\mathcal{D}$ satisfies the conditions of Theorem 4.1.

For the correctness of Algorithm 1, first observe that line i of the algorithm has a unique solution by the assumption of Theorem 4.1. We must also show uniqueness in line 3. This follows by the result of Theorem 4.1, which shows that two different collections of sites representing these constraints must be related by a positive-affine transformation. To see this, observe that once we have fixed $\boldsymbol{v}^s$ and $\boldsymbol{v}^{s'}$, the rest of the

[10] This condition is a special case of the so-called *simple* power diagrams; see footnote 9. For more complicated cases, one would use [Rybnikov 1999].

Algorithm 1 Compute Mechanism from Belief Model Constraint

i. Solve for $\boldsymbol{y}$ as the unique solution to $\boldsymbol{u}^{s,s'} \cdot \boldsymbol{y} = 0 \ \ \forall s, s'$ and $\boldsymbol{y} \cdot \mathbb{1} = 1$.
ii. Let $\hat{\boldsymbol{u}}^{s,s'} = \boldsymbol{u}^{s,s'} - (\frac{1}{m}\boldsymbol{u}^{s,s'} \cdot \mathbb{1})\mathbb{1}$ for all s, s'. # Projects $\boldsymbol{u}^{s,s'}$ onto affine hull of Δ_m
1: Choose $s \in [m]$ and any $\boldsymbol{v}^s$ in the affine hull of Δ_m.
2: Choose $s' \neq s$ and any $\alpha > 0$, and set $\boldsymbol{v}^{s'} = \boldsymbol{v}^s + \alpha\hat{\boldsymbol{u}}^{s,s'}$.
3: For all $s'' \notin \{s, s'\}$, find the unique positive solution (β, γ) to
$\alpha\hat{\boldsymbol{u}}^{s,s'} + \beta\hat{\boldsymbol{u}}^{s',s''} = \gamma\hat{\boldsymbol{u}}^{s,s''}$, and set $\boldsymbol{v}^{s''} = \boldsymbol{v}^s + \gamma\hat{\boldsymbol{u}}^{s,s''}$. # See argument below
4: # The point y must have equal power distance to all sites:
Set $w(s) = 0$ and $w(s'') = \|\boldsymbol{y} - \boldsymbol{v}^{s''}\|^2 - \|\boldsymbol{y} - \boldsymbol{v}^s\|^2$ for all $s'' \neq s$.
5: Compute the mechanism by applying Eq. 5.

sites are uniquely determined.[11] Consider some positive-affine transformation of the sites that keeps $\boldsymbol{v}^s$ and $\boldsymbol{v}^{s'}$. As $\boldsymbol{v}^s$ has not moved, the translation must be 0, and as $\boldsymbol{v}^{s'}$ also remains the same, so does the length of the line segment between $\boldsymbol{v}^s$ and $\boldsymbol{v}^{s''}$, so the scaling must be 1. We have now determined the positive-affine transformation: the identity. Thus all other sites are uniquely determined by the first two. Finally, one can check that the projections $\hat{\boldsymbol{u}}^{s,s'}$ are correctly oriented and remain perpendicular to the boundary of cells D_s and $D_{s'}$.[12]

5.2. Optimizing Effort Incentives

From Section 5.1, we know how to compute a mechanism that is truthful with respect to a given belief model constraint. In this section, we take this one step further and optimize within the space of truthful mechanisms. As explained in Section 1, peer prediction mechanisms are especially useful for incentivizing effort, i.e. the costly acquisition of signals, and we will thus address the following optimization problem:

$$\begin{array}{lll} \max & \text{effort incentives} & e_i(M) \\ s.t. & \text{truthfulness} & \text{with respect to } \mathcal{D} \\ & \text{budget constraint} & M(x_i, x_j) \leq B \\ & \text{non-negative payments} & M(x_i, x_j) \geq 0 \end{array} \tag{8}$$

Effort can be modeled in many different ways. Following Witkowski [2014], we model effort as a binary choice: agents either exert effort or not. In contrast to the rest of the paper, where agent i has observed her signal and reasons about the best report given that signal, here the choice of whether to invest effort also depends on agent i's prior belief about her own signal. In this section, we assume that agents are exchangeable, so that agent i's prior belief about her own signal is the same as her prior belief about agent j's signal. Let then $p_i(s_i) = \Pr_i(S_i = s_i) = p_i(s_j)$ denote agent i's *signal prior*. Note that for an agent's signal prior and belief model, by Bayes' rule it holds that $p_i(s_j) = \sum_{k=1}^{m} p_i(s_j|k) \cdot p_i(k)$.

Definition 5.1. Given that agent j invests effort and reports truthfully, the *effort incentive* $e_i(M)$ that is implemented for agent i by peer prediction mechanism M is the difference in expected utility of investing effort followed by truthful reporting and not

[11] Note that we are restricting the sites to the affine hull of the simplex, which is simply given by the condition $\boldsymbol{v} \cdot \mathbb{1} = 1$; see Appendix B. Without this, there would be an extra degree of freedom in choosing the sites, but as that section shows, this is irrelevant.
[12] Letting $p, q \in \Delta_m$ be distinct points on the boundary of these cells, and letting $z = p - q$, we have $\hat{\boldsymbol{u}}^{s,s'} \cdot z = \boldsymbol{u}^{s,s'} \cdot z - (\frac{1}{m}\boldsymbol{u}^{s,s'} \cdot \mathbb{1})\mathbb{1} \cdot z = \boldsymbol{u}^{s,s'} \cdot z$ as $p \cdot \mathbb{1} = q \cdot \mathbb{1} = 1$, so $\mathbb{1} \cdot z = 0$.

investing effort, i.e.

$$e_i(M) = \mathop{\mathbf{E}}_{S_i,S_j}\Big[M(S_i,S_j)\Big] - \max_{x_i\in[m]} \mathop{\mathbf{E}}_{S_j}\Big[M(x_i,S_j)\Big],$$

where x_i is agent i's signal report that maximizes her expected utility according to the signal prior, and where the expectation is using agent i's subjective belief model $p_i(\cdot|\cdot)$.

Thus, the effort incentive $e_i(M)$ is agent i's expected gain by exerting effort. Naturally, scaling a mechanism should scale the incentives; in fact, this can be generalized to positive-affine transformations.

LEMMA 5.2. *For any mechanism M, and any positive-affine transformation $M' = \alpha M + f$, we have $e_i(M') = \alpha e_i(M)$.*

PROOF. This follows from a simple computation:

$$\begin{aligned}
e_i(M') &= \mathop{\mathbf{E}}_{S_i,S_j}\Big[M'(S_i,S_j)\Big] - \max_{x_i\in[m]} \mathop{\mathbf{E}}_{S_j}\Big[M'(x_i,S_j)\Big] \\
&= \mathop{\mathbf{E}}_{S_i,S_j}\Big[\alpha M(S_i,S_j) + f(S_j)\Big] - \max_{x_i\in[m]} \mathop{\mathbf{E}}_{S_j}\Big[\alpha M(x_i,S_j) + f(S_j)\Big] \\
&= \alpha \mathop{\mathbf{E}}_{S_i,S_j}\Big[M(S_i,S_j)\Big] + \mathop{\mathbf{E}}_{S_j}\Big[f(S_j)\Big] - \alpha \max_{x_i\in[m]} \mathop{\mathbf{E}}_{S_j}\Big[M(x_i,S_j)\Big] - \mathop{\mathbf{E}}_{S_j}\Big[f(S_j)\Big] \\
&= \alpha e_i(M)\ .
\end{aligned}$$

□

From Section 4 we know that the space to optimize over is restricted to positive-affine transformations of any truthful mechanism once the belief model constraint is fixed and given the conditions of Theorem 4.1. Using this, we can pin down the effort-maximizing mechanism as given by the following theorem.

THEOREM 5.3. *Let mechanism M and belief model constraint $\mathcal{D}$ satisfy the conditions of Theorem 4.1. Let $g(x_j) = \min_{x_i} M(x_i,x_j)$, $G = \max_{x_i,x_j} M(x_i,x_j) - g(x_j)$, and $\alpha = B/G$. Then mechanism $M'(x_i,x_j) = \alpha M(x_i,x_j) - \alpha g(x_j)$ optimizes effort in Eq. 8.*

PROOF. We will show something slightly stronger, giving the full characterization of mechanisms optimizing Eq. 8. By Theorem 4.1, M is the unique truthful mechanism for $\mathcal{D}$ up to affine transformations $M_{\alpha,f}(x_i,x_j) = \alpha M(x_i,x_j) + f(x_j)$, so we need only search for the optimal α, f. By Lemma 5.2, $e_i(M_{\alpha,f}) = \alpha e_i(M)$, which since $e_i(M)$ is a positive constant reduces the optimization to the following:

$$\begin{aligned}
\max \quad & \alpha \\
s.t. \quad & \alpha M(x_i,x_j) + f(x_j) \in [0,B]\ \forall x_i,x_j \\
& f \in \mathbb{R}^m, \alpha \in \mathbb{R}^+\ .
\end{aligned}$$

Let $G = \max_{x_j}\big(\max_{x_i} M(x_i,x_j) - \min_{x_i} M(x_i,x_j)\big)$ denote the maximum payment difference within any column of M. We see that the largest value of α one could hope for is $\alpha = B/G$ since the final mechanism should have values in $[0,B]$, and in particular the difference between the minimum payment and maximum payment of the final mechanism must be at most B. Indeed, taking this α, one sees that any f in the range $f(x_j) \in [-\alpha \min_{x_i} M(x_i,x_j), -\alpha(\max_{x_i} M(x_i,x_j) - \min_{x_i} M(x_i,x_j))] \neq \emptyset$ will satisfy the constraints. One can check that in particular taking $f(x_j) = -\alpha \min_{x_i} M(x_i,x_j)$ gives the mechanism in the theorem statement, which gives the lowest possible payments among the optimal mechanisms. As $f = -\alpha g$, we are done. To confirm that $M'(x_i,x_j) \leq B$ for all $x_i,x_j \in [m]$, note that $M'(x_i,x_j) = B\big(M(x_i,x_j) - g(x_j)\big)/G =$

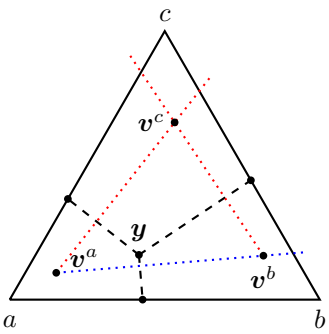

Fig. 5: The truthful mechanism with respect to the "complement 1/p" condition in power diagram form with sites $\boldsymbol{v}^a, \boldsymbol{v}^b$, and $\boldsymbol{v}^c$. Notice that while the intersection point $\boldsymbol{y} = (1/2, 1/3, 1/6)$ is the same as in Figure 2L, the belief model constraint as depicted by the dashed partitioning is now given by the new "complement 1/p" condition. (Compare to Figure 3.)

$B \cdot \frac{M(x_i,x_j) - \min_{x_i'} M(x_i',x_j)}{\max_{x_j'} \left(\max_{x_i'} M(x_i',x_j') - \min_{x_i'} M(x_i',x_j') \right)} \leq B$, as the numerator is a difference of payoff values within a column of M, and G by definition is the largest such difference in any column of M. Similarly, for all $x_i, x_j \in [m]$ we have $M'(x_i, x_j) = \alpha(M(x_i, x_j) - \min_{x_i'} M(x_i, x_j)) \geq 0$ as $\alpha > 0$ and $M(x_i, x_j) \geq \min_{x_i'} M(x_i', x_j)$. □

Note that knowledge of only $\mathcal{D}$ is not sufficient to ensure that a specific effort cost $C > 0$ is implemented (e.g. $C = 1$) since the magnitude of an agent's effort incentive will depend on the particular values of their posteriors $\{p_i(\cdot|s)\}_{s \in [m]}$. In particular, the agents posteriors could all be arbitrarily close to the intersection point, in which case the required scaling α to meet effort cost $C = 1$ would be arbitrarily high. Even knowing only $\mathcal{D}$ and not the agent's actual belief model, however, Theorem 5.3 gives the highest possible scaling within the budget (and thus the highest possible effort incentives) of any mechanism implementing $\mathcal{D}$, no matter the agent's belief model.

To illustrate the power of Theorem 5.3, consider the Shadowing Method (Definition 2.8), which has a parameter δ specifying how much to perturb $\boldsymbol{y}$ in order to obtain the shadow posterior $\boldsymbol{y}'$. A natural question to ask is for which δ the effort is maximized subject to the constraints in Eq. 8. (One can check that the resulting power diagram is the same for any such δ.) A direct analysis is quite tedious, with many lines of algebra, and even with the optimal δ in hand, it is not clear whether one could do better by perhaps adding a score $f(x_j)$ that depends only on the peer agent's signal report x_j, or by allowing for shadow posteriors that aren't valid distributions followed by renormalizing the scoring rule so that the score is again bounded by $[0, B]$. Moreover, the optimal mechanism could have been of a different form entirely. Theorem 5.3 suggests a better approach to this problem: take the Shadowing Method with any $\delta > 0$ as a black box and simply find the maximum scaling allowing a translation that keeps the scores in the interval $[0, B]$.

5.3. Example

We now illustrate Algorithm 1, constructing a new mechanism that is truthful with respect to a new condition. Moreover, we compute the optimal such mechanism with respect to the effort it incentivizes as explained in Section 5.2.

As intuition for the new condition, imagine the mechanism has an estimate of the agents' signal priors $p(a, b, c) = (0.01, 0.04, 0.95)$, which it designates as the intersection point $y(\cdot) = p(\cdot)$ of the belief model constraint. Consider now posterior

$p_i(a, b, c|s) = (0.02, 0.01, 0.97)$, where the 1/p mechanism would pick signal a since its relative increase from prior (as estimated by the mechanism) to posterior is highest (it doubles). However, one could also consider the relative decrease in "error": in a world without noise, the posterior would have $p_i(s|s) = 1$ for every s, and so signal a's relative decrease from $0.99 = 1 - 0.01$ to $0.98 = 1 - 0.02$ is not as "impressive" as signal c's decrease in error from $0.05 = 1 - 0.95$ to $0.03 = 1 - 0.97$ (a reduction of almost one half). Formalizing this intuition yields the "complement 1/p" condition, $\frac{1-y(s)}{1-p_i(s|s)} > \frac{1-y(s')}{1-p_i(s'|s)}$ $\forall\, s' \neq s$ and $\forall\, i \in [n]$. For this condition, one can take $\boldsymbol{u}^{s,s'} = (1 - y(s))\mathbb{1}_{s'} - (1 - y(s'))\mathbb{1}_s + (y(s) - y(s'))\mathbb{1}$.

Theorem 4.1 implies that there is a unique mechanism that is truthful for this new condition, up to positive-affine transformations. We now exemplify the construction of the new "complement 1/p" condition following the steps of Algorithm 1. For that purpose, we return to our running example with $m = 3$ signals and intersection point $\boldsymbol{y} = (1/2, 1/3, 1/6)$ as depicted in Figure 5.

1: Pick any point for $\boldsymbol{v}^a$, say $\boldsymbol{v}^a = (4/5, 1/10, 1/10)$.[13]
2: Pick any $\boldsymbol{v}^b$ on the blue dotted line, ensuring that the line between $\boldsymbol{v}^a$ and $\boldsymbol{v}^b$ is perpendicular to the a, b cell boundary. Here we choose $\boldsymbol{v}^b = (1/10, 81/110, 9/55)$.
3: For all other signals s, $\boldsymbol{v}^s$ is now uniquely determined by $\boldsymbol{v}^a$ and $\boldsymbol{v}^b$ as the lines between any two sites must be perpendicular to their cell boundary. Here we only have one other signal, c, so we take $\boldsymbol{v}^c$ to be the unique point at the intersection of the red dotted lines, which is $\boldsymbol{v}^c = (38/275, 111/550, 33/50)$.
4: Calculate the weights by observing that $\boldsymbol{y}$ must be equidistant (in the power distance) to all sites simultaneously: $w(a) = 0$, $w(b) = 23/100$, $w(c) = 548/1875$.
5: We obtain the resulting mechanism by applying Eq. 5:

$$M(\cdot, \cdot) = \frac{1}{1100}\begin{bmatrix} 517 & -253 & -253 \\ -113 & 587 & -43 \\ 13 & 83 & 587 \end{bmatrix}.$$

6: From Theorem 5.3, it then follows that, among all positive-affine transformations of M, the mechanism M^* optimizing effort incentives given a budget $B = 1$ is:

$$M^*(\cdot, \cdot) = \frac{1}{20}\begin{bmatrix} 15 & 0 & 0 \\ 0 & 20 & 5 \\ 3 & 8 & 20 \end{bmatrix}.$$

This step can be computed as follows: subtract the largest amount from each column that keeps payments nonnegative, and then scale so the largest entry is 1.

We conclude by noting that this example condition admits a closed form solution: $M(s, s') = 0$ if $s = s'$, and $\frac{-1}{1-y(s)}$ otherwise.[14] One can check that adding constants to each column to give non-negative payments recovers M^*. Of course, our construction applies even when no convenient closed-form solution exists; we discuss such examples in Section 8.

6. EXPRESSIVENESS OF CLASSICAL PEER PREDICTION

We now turn to the general question of whether one minimal peer prediction method is more or less expressive than another. For example, it is well known that Output

[13] While we chose $\boldsymbol{v}^a \in D_a$, it could be in any other cell, e.g. D_c, or even outside the simplex.

[14] The expected payoff of reporting s' when $s_i = s$ is $-\frac{\Pr[s_j \neq s']}{1-y(s')} = -\frac{1-p(s'|s)}{1-y(s')} < -\frac{1-p(s|s)}{1-y(s)}$ by the condition.

Agreement (OA) is a special case of both Shadowing and 1/p when $\boldsymbol{y}$ is the uniform distribution, and therefore the latter are both "more expressive" than OA. We can formalize this notion of expressiveness by thinking of a method as a set of minimal mechanisms parameterized by some inputs, such as the distribution $\boldsymbol{y}$ for Shadowing and 1/p. We then say method A is at least as expressive as method B, denoted $A \succeq B$, if the corresponding set of minimal mechanisms for A entirely contains the set for B, meaning every B-mechanism is also an A-mechanism (for an appropriate choice of parameters).[15] Thus, we have Shadowing $\succeq$ OA and 1/p $\succeq$ OA; in fact these are strict relations ($\succ$). Moreover, 1/p and Shadowing are incomparable [Witkowski 2014], in the sense that there are some choices of $\boldsymbol{y}$ for which the resulting 1/p mechanism cannot be written as a Shadowing mechanism for any $\boldsymbol{y}'$, and vice versa, even up to positive-affine transformations. (In fact, this is true for any nonuniform $\boldsymbol{y}$.)

In this section, we explore the expressiveness of the classical peer prediction method (Definition 2.10). We have seen from Theorem 3.4 that, as a result of viewing a minimal mechanism as eliciting a "property" of an agent's posterior, we can establish a strong correspondence between minimal peer prediction mechanisms and power diagrams. Here we delve further and find that, perhaps surprisingly, every minimal mechanism can in fact be written as a classical peer prediction mechanism for some choice of distributions and proper scoring rule; in the notation above, classical PP $\succeq$ minimal PP, which of course implies classical PP $\simeq$ minimal PP. (Since Shadowing and 1/p are incomparable, classical PP $\simeq$ minimal PP in turn shows that classical PP $\succ$ Shadowing and classical PP $\succ$ 1/p.) Moreover, provided the belief model constraint has an intersection point in the interior of the simplex, as needed in our uniqueness results from Section 4, we can take this scoring rule to be the quadratic score without loss of generality, thus implying that "nondegenerate" minimal mechanisms in fact correspond to Voronoi diagrams. These results tell us that the classical peer prediction method is "peer prediction complete," and for most cases remains so even when restricting to the quadratic score.

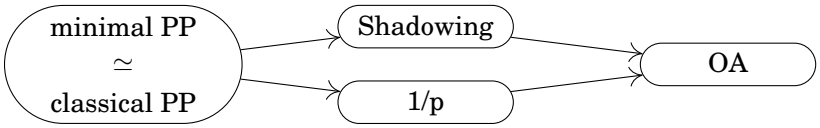

Fig. 6: Relative expressiveness of minimal peer prediction methods. Here the directed arrows signify "$\succ$", meaning "strictly more expressive than."

The key insight we will need has to do with a class of diagrams known as *Bregman Voronoi diagrams*. These are Voronoi diagrams where instead of Euclidean distance, the distance is given by a (typically asymmetric) *Bregman divergence*,

$$D_G(\boldsymbol{u}, \boldsymbol{v}) = G(\boldsymbol{u}) - G(\boldsymbol{v}) - \nabla G(\boldsymbol{v}) \cdot (\boldsymbol{u} - \boldsymbol{v}) \,, \tag{9}$$

where G is any convex function [Nielsen et al. 2007]. Specifically, the distance between a site $\boldsymbol{v}^s$ and a point $\boldsymbol{u}$ is given by $D_G(\boldsymbol{u}, \boldsymbol{v}^s)$.[16] Bregman divergences are known to be equivalent to proper scoring rules when $\boldsymbol{u}, \boldsymbol{v} \in \Delta_m$ [Gneiting and Raftery 2007], making Bregman Voronoi diagrams precisely those which arise from classical peer

[15] As before, we will only distinguish minimal mechanisms up to positive-affine transformations, as these transformations do not alter the belief model constraints. In principle, one could also refine the notion of expressiveness to distinguish minimal mechanisms with the same belief model constraints.

[16] Nielsen et al. dub this the first-type Bregman Voronoi diagram; the second reverses the arguments.

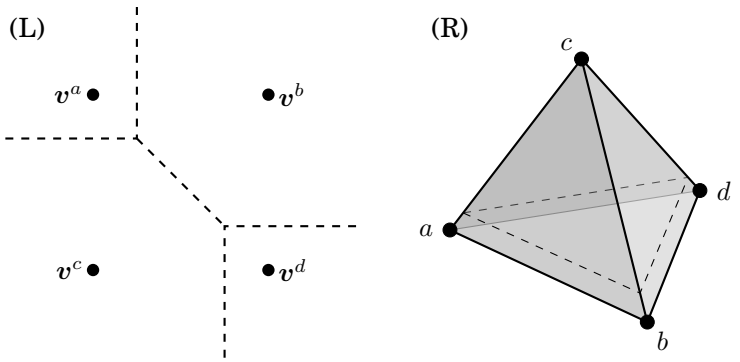

Fig. 7: **(L)** A power diagram in the plane with 4 sites that is not a Voronoi diagram. If the sites are the vertices of the unit square, the weights are $w(a) = w(d) = 0$ and $w(b) = w(c) = 1/2$. To see that the diagram cannot be Voronoi, note that because the diagram is simple (see footnote 9) the sites drawn are unique up to positive-affine transformations; due to the angles of the cell boundaries, they must always form a square with the correct orientation. But clearly the Voronoi diagram for such sites would always be "+" shaped, so this diagram cannot be Voronoi. Note that with 3 sites, every power diagram is actually Voronoi: either the cell boundaries are parallel or they all intersect at some point $\boldsymbol{y}$, the former of which is trivially Voronoi and the latter of which is Voronoi by Theorem 6.2.
(R) For an example in our peer prediction setting, we must consider $m = 4$ signals to have four sites, where the simplex Δ_4 can be visualized as a tetrahedron. Here we simply embed the sites on a plane within the tetrahedron in the same way; for example, one could take sites $\boldsymbol{v}^a = (0.1, 0.1, 0.1, 0.7)$, $\boldsymbol{v}^b = (0.4, 0.1, 0.1, 0.3)$, $\boldsymbol{v}^c = (0.1, 0.4, 0.1, 0.3)$, $\boldsymbol{v}^d = (0.4, 0.4, 0.1, 0.1)$, where the first two coordinates are tracing out a square on the plane $\{\boldsymbol{v} : \boldsymbol{v}_3 = 0.1\}$ shown as dashed lines. Using the same weights as above, we get a similar diagram, which is not Voronoi for the same reason.

prediction mechanisms. Note that here $\boldsymbol{v}$ is the agent's prediction, so we swap the order of the arguments: $R(p, q) = D_G(q, p)$.

Somewhat surprisingly, Nielsen et al. [2007] show that Bregman Voronoi diagrams are equivalent to power diagrams. The intuition for this is as follows. Examining the form of $D_G(\boldsymbol{u}, \boldsymbol{v}^s) = G(\boldsymbol{u}) - G(\boldsymbol{v}^s) - \nabla G(\boldsymbol{v}^s) \cdot \boldsymbol{u} + \nabla G(\boldsymbol{v}^s) \cdot \boldsymbol{v}^s$, we see that the cell boundaries are given by only the last three terms, as for a given $\boldsymbol{u}$, the $G(\boldsymbol{u})$ term will be the same for all sites. Moreover, the interaction between $\nabla G(\boldsymbol{v}^s)$ and $\boldsymbol{u}$ is linear, so we can take $\nabla G(\boldsymbol{v}^s)$ to be the site of the power diagram, and adjust the weight so that when we "complete the square", we will have a squared Euclidean distance instead. Specifically, we can rewrite the last three terms as $\|\hat{\boldsymbol{v}}^s\|^s - \hat{\boldsymbol{v}}^s \cdot \boldsymbol{u} - w(\hat{\boldsymbol{v}}^s)$ for $\hat{\boldsymbol{v}}^s = \nabla G(\boldsymbol{v}^s)$ and $w(\hat{\boldsymbol{v}}^s) = \|\hat{\boldsymbol{v}}^s\|^s + \hat{\boldsymbol{v}}^s \cdot \boldsymbol{v}^s - G(\boldsymbol{v}^s)$. Adding the term $\|\boldsymbol{u}\|^s$ which does not depend on the site, we arrive at Eq. 2. (See Frongillo and Kash [2014] for a more detailed analysis.)

Putting the above together with Theorem 3.4, we have minimal PP $\iff$ power diagram $\iff$ Bregman Voronoi $\iff$ classical PP, or in other words, every minimal mechanism can be written as a classical peer prediction mechanism with respect to some proper scoring rule. We summarize this discussion as a corollary of Theorem 3.4.

COROLLARY 6.1. *Every minimal peer prediction mechanism M can be written as a classical peer prediction mechanism $M(x_i, x_j) = R(\boldsymbol{p}^{x_i}, x_j)$ for some choice of proper scoring rule R and sites $\{\boldsymbol{p}^s\}_{s\in[m]}$ in the affine hull of Δ_m. (Here $\boldsymbol{p}^{x_i}$ is the vector form of $p_i(\cdot|x_i)$ from Definition 2.10.)*

Note that the sites in Corollary 6.1 need not be in the simplex. Also, while it is easy to construct such a proper scoring rule R given M by expressing it as the Bregman divergence with respect to the convex function $G(\boldsymbol{p}) = \operatorname{argmax}_{x_i} \mathbf{E}_{x_j\sim\boldsymbol{p}}[M(x_i, x_j)]$, the resulting score R will not be *strictly* proper, as G is not strictly convex. It seems

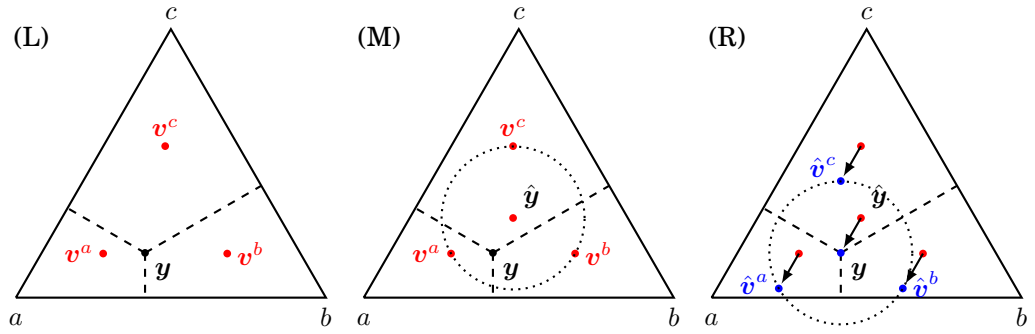

Fig. 8: Converting a power diagram to a Voronoi diagram.

intuitively clear that some other R can be chosen to be strictly proper, yet an explicit construction remains an open question. Such a construction would involve finding a strictly convex function G and points $\boldsymbol{p}^s$ satisfying $\boldsymbol{v}^s = \frac{1}{2}\nabla G(\boldsymbol{p}^s)$ and $w(s) = \frac{1}{4}\|\nabla G(\boldsymbol{p}^s)\|^2 + G(\boldsymbol{p}^s) - \boldsymbol{p}^s \cdot \nabla G(\boldsymbol{p}^s)$ [Frongillo and Kash 2014, Appendix B].

Now that we have seen that classical peer prediction mechanisms cover all possible minimal mechanisms, let us impose a further restriction on the mechanism: the belief model constraint it induces must satify the conditions of Theorem 4.1, namely that the constraint has a point at the intersection of every cell. Under this condition, we will show that not only is the mechanism unique up to positive-affine transformations, but it can be expressed as a positive-affine transformation of a classical peer prediction mechanism with respect to the quadratic scoring rule. In other words, the induced belief model constraint must in fact be a Voronoi diagram.[17] For comparison, see Figure 7 for an example of a power diagram that is not a Voronoi diagram.

The intuition for this result is as follows. Because all pairs of cells have a nonempty border, the sites must be in general position, meaning that they are affinely independent within the affine hull of the simplex.[18] Leveraging a standard geometric construction, we can compute the center $\hat{\boldsymbol{p}}$ of an $(m-1)$-sphere such that the sites all lie on the surface of the sphere (Figure 8M). This implies that the sites are all equidistant (in Euclidean distance) to center $\hat{\boldsymbol{p}}$, so we merely translate the sites so that $\hat{\boldsymbol{p}}$ coincides with the cell boundary intersection point $\boldsymbol{p}$ (Figure 8R). With appropriate weights, we know that this translation preserves the power diagram, yet now all sites are equidistant to $\boldsymbol{p}$ in Euclidean distance, so in fact we can set $w(s) = 0$ for all s and the diagram is Voronoi.

THEOREM 6.2. *Let M, $\mathcal{D}$, and $\boldsymbol{p}$ satisfy the assumption of Theorem 4.1. Then there exist points $\boldsymbol{p}^1, \ldots, \boldsymbol{p}^m$ in the affine hull of Δ_m such that M' is a truthful mechanism for $\mathcal{D}$ if and only if*

$$M'(x_i, x_j) = \alpha \boldsymbol{p}^{x_i}(x_j) - \frac{\alpha}{2}\|\boldsymbol{p}^{x_i}\|^2 + f(x_j)\,, \tag{10}$$

for some $f : [m] \to \mathbb{R}$ and $\alpha > 0$. In particular, $\mathcal{D}$ is Voronoi.

PROOF. Note that from Theorem 4.1 we already know that $\mathcal{D}$ is a power diagram with sites $\boldsymbol{v}^s \in \mathbb{R}^m$ and weights $w(s) \in \mathbb{R}$ for $s \in [m]$. Following [Eberly 2008], we will

[17] We can see from Appendix B, as is well-known, every power diagram on an affine subspace of lower dimension (in our case, $\{\boldsymbol{v} : \boldsymbol{v} \cdot \mathbb{1} = 1\}$) can be expressed as a Voronoi diagram in the full space by moving the sites perpendicularly to the affine subspace. Here we are saying something stronger: these diagrams are Voronoi for sites *within* the affine subspace.

[18] Geometrically, no hyperplane in $\mathbb{R}^m$ can contain all m sites.

repeatedly refer to the identity

$$\|\boldsymbol{a}+\boldsymbol{c}\|^2 - \|\boldsymbol{b}+\boldsymbol{c}\|^2 = \|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2 + 2\boldsymbol{c}\cdot(\boldsymbol{a}-\boldsymbol{b})\ . \tag{11}$$

Let $d = m-1$, and as outlined in Appendix B, we restrict attention to $\mathbb{R}^d$ via isometry to the affine hull of Δ_m. Thus, we can assume without loss of generality that $\boldsymbol{v}^1, \dots, \boldsymbol{v}^m, \boldsymbol{y} \in \mathbb{R}^d$. Let A be the $d \times d$ matrix whose ith row is given by $\boldsymbol{v}^i - \boldsymbol{v}^m$.

Note that $\boldsymbol{y}$ is the unique point satisfying $\boldsymbol{y} \in \mathrm{cl}(\mathrm{cell}(\boldsymbol{v}^s))$ for all s, meaning that for some power distance $c \in \mathbb{R}$ we have

$$\forall s \in [m], \quad \|\boldsymbol{v}^s - \boldsymbol{y}\|^2 - w(s) = c\ . \tag{12}$$

Let $w(m) = 0$ without loss of generality. Subtracting the equation for $s = m$ we have,

$$\forall s \in [d], \quad 0 = \|\boldsymbol{v}^s - \boldsymbol{y}\|^2 - \|\boldsymbol{v}^m - \boldsymbol{y}\|^2 - w(s) \tag{13}$$

$$= \|\boldsymbol{v}^s\|^2 - \|\boldsymbol{v}^m\|^2 - 2\boldsymbol{y}\cdot(\boldsymbol{v}^s - \boldsymbol{v}^m) - w(s)\ , \tag{14}$$

where we have used the identity in Eq. 11. Letting $\boldsymbol{u} \in \mathbb{R}^d$ be the vector with $\boldsymbol{u}(s) = \frac{1}{2}(\|\boldsymbol{v}^s\|^2 - \|\boldsymbol{v}^m\|^2 - w(s))$, and recalling the definition of $A \in \mathbb{R}^{d\times d}$, we see that $\boldsymbol{y}$ is the unique solution to $A\boldsymbol{y} = \boldsymbol{u}$, meaning that A is invertible.

Let $\hat{\boldsymbol{u}}$ be defined as $\boldsymbol{u}$ but assuming $w(s) = 0$ for all s, i.e., with $\hat{\boldsymbol{u}}(s) = \frac{1}{2}(\|\boldsymbol{v}^s\|^2 - \|\boldsymbol{v}^m\|^2)$. Let $\hat{\boldsymbol{y}} = A^{-1}\hat{\boldsymbol{u}}$, the unique point satisfying $\|\boldsymbol{v}^s - \hat{\boldsymbol{y}}\|^2 = \|\boldsymbol{v}^m - \hat{\boldsymbol{y}}\|^2$ for all $s \in [d]$. (One can verify this by unfolding $A\hat{\boldsymbol{y}} = \hat{\boldsymbol{u}}$ and applying Eq. 14.)

Geometrically, it is now clear that sites $\hat{\boldsymbol{v}}^s = \boldsymbol{v}^s + \hat{\boldsymbol{y}} - \boldsymbol{y}$ and weights $\hat{w}(s) = 0$ represent $\mathcal{D}$, simply because $\boldsymbol{y}$ is equidistant from $\boldsymbol{v} + \hat{\boldsymbol{y}} - \boldsymbol{y}$ for all s, and as we only translated the sites and changed the weights, we can only translate the hyperplanes separating cells (i.e. we have not rotated any cell boundaries). To verify this claim algebraically, we will show that for any $\boldsymbol{q} \in \mathbb{R}^d$, and any cells s, s', the difference in power distances to $\boldsymbol{q}$ remains the same:

$$\begin{aligned}
&\|\hat{\boldsymbol{v}}^s - \boldsymbol{q}\|^2 - \|\hat{\boldsymbol{v}}^{s'} - \boldsymbol{q}\|^2 \\
&= \|\boldsymbol{v}^s + \boldsymbol{y} - \hat{\boldsymbol{y}} - \boldsymbol{q}\|^2 - \|\boldsymbol{v}^{s'} + \boldsymbol{y} - \hat{\boldsymbol{y}} - \boldsymbol{q}\|^2 \\
&= \|\boldsymbol{v}^s - \hat{\boldsymbol{y}}\|^2 - \|\boldsymbol{v}^{s'} - \hat{\boldsymbol{y}}\|^2 + 2(\boldsymbol{y}-\boldsymbol{q})\cdot(\boldsymbol{v}^s - \boldsymbol{v}^{s'}) \\
&= 2\boldsymbol{y}\cdot(\boldsymbol{v}^s - \boldsymbol{v}^{s'}) - 2\boldsymbol{q}\cdot(\boldsymbol{v}^s - \boldsymbol{v}^{s'}) \\
&= \|\boldsymbol{v}^s - \boldsymbol{q}\|^2 - \|\boldsymbol{v}^s - \boldsymbol{y}\|^2 - \|\boldsymbol{v}^{s'} - \boldsymbol{q}\|^2 + \|\boldsymbol{v}^{s'} - \boldsymbol{y}\|^2 \\
&= \|\boldsymbol{v}^s - \boldsymbol{q}\|^2 - w(s) - \|\boldsymbol{v}^{s'} - \boldsymbol{q}\|^2 + w(s')\ ,
\end{aligned}$$

where we have applied the identity of Eq. 11 three times, and in the final step we use Eq. 13. Thus, $\mathcal{D}$ is a Voronoi diagram with sites $\hat{\boldsymbol{v}}^s$. To complete the proof, we simply convert to a mechanism via Theorem 3.4 and then apply Theorem 4.1. □

7. MAXIMALLY-ROBUST MECHANISMS

In the classical peer prediction method [Miller et al. 2005], the mechanism designer is assumed to have full knowledge of the agents' belief models. Recent work relaxes the method's knowledge requirements, e.g. using additional reports [Prelec 2004; Witkowski and Parkes 2012b; Witkowski and Parkes 2012a] or using reports on several items [Dasgupta and Ghosh 2013; Witkowski and Parkes 2013]. An approach closer to the classical method has been suggested by Jurca and Faltings [2007], who compute a minimal mechanism as the solution of a conic optimization problem that ensures truthfulness as long as the agents' belief models are close to the mechanism designer's, with respect to Euclidean distance. This restriction, a form of *robustness*, is defined as follows.

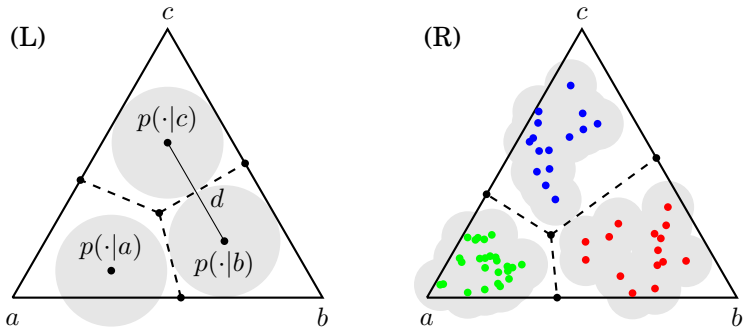

Fig. 9: **(L)** Maximally-robust mechanism with respect to deviations from the mechanism designer's belief model. The sites of the power diagram are $\boldsymbol{v}^s = p(\cdot|s)$, and all weights are 0. The robustness areas, pictured in grey, are circles of radius $d/2$, where d is the minimum distance between two posteriors. **(R)** A mechanism with maximal robustness with respect to labeled posterior data.

Definition 7.1. [Jurca and Faltings 2007] A mechanism M is ϵ-robust with respect to belief model $p(\cdot|\cdot)$ if M is truthful for $p^*(\cdot|\cdot)$ whenever the following holds for all $s_i \in [m]$,

$$\sum_{s_j \in [m]} (p(s_j|s_i) - p^*(s_j|s_i))^2 \leq \epsilon^2 \ . \tag{15}$$

While Jurca and Faltings fix the robustness ϵ as a hard constraint, one may also seek the mechanism that maximizes this robustness. The achievable robustness is of course limited by the mechanism designer's belief model $p(\cdot|\cdot)$; in particular, the "robustness areas" around the posteriors cannot overlap; see Figure 9L. Viewing robustness in geometric terms, we obtain a closed-form solution.[19] Note that here the sites of the power (actually Voronoi) diagram are the mechanism designer's model posteriors.

THEOREM 7.2. *Let $p(\cdot|\cdot)$ be the mechanism designer's belief model in classical peer prediction. Then the following mechanism is maximally robust:*

$$M(x_i, x_j) = p(x_j|x_i) - \frac{1}{2}\sum_{s=1}^{m} p(s|x_i)^2. \tag{16}$$

PROOF. In light of Theorem 3.4, we may focus instead on power diagrams. From Eq. 17, for all s we must have $B_\epsilon(p(\cdot|s)) \subseteq \text{cell}(\boldsymbol{v}^s)$, where $B_\epsilon(\boldsymbol{u})$ is the Euclidean ball of radius ϵ about $\boldsymbol{u}$ (restricted to the probability simplex). Letting $d = \min_{s,s' \in [m]} \|p(\cdot|s) - p(\cdot|s')\|$ be the minimum Euclidean distance between any two posteriors, it becomes clear that robustness of $d/2$ or greater cannot be achieved, as $\frac{1}{2}p(\cdot|s) + \frac{1}{2}p(\cdot|s') \in B_{d/2}(p(\cdot|s)) \cap B_{d/2}(p(\cdot|s'))$. (See Figure 9L.) Robustness of any $\epsilon < d/2$ can be achieved, however, by taking a Voronoi diagram with sites $\boldsymbol{v}^s = p(\cdot|s)$; the definition of d ensures that $B_\epsilon(p(\cdot|s)) = B_\epsilon(\boldsymbol{v}^s) \subseteq \text{cell}(\boldsymbol{v}^s)$ for all s. One then recovers Eq. 16 via Eq. 5 with $\boldsymbol{v}^s = p(\cdot|s)$ and $w(s) = 0$ for all $s \in [m]$. □

[19] Note that the maximal robustness is determined by the two posteriors that are closest to each other. This gives some flexibility as to where the cell boundaries shall be between posteriors that are further away than twice the maximal robustness. Our solution always selects the cell boundary that is equidistant from the respective posteriors. See Figure 9L, where the maximal robustness is determined through $p(\cdot|a)$ and $p(\cdot|b)$, and where the cell boundary between D_a and D_c could be moved slightly towards either $p(\cdot|a)$ or $p(\cdot|c)$ without breaking into the respective robustness areas.

COROLLARY 7.3. *The classical peer prediction method with the quadratic scoring rule is maximally robust.*

In Appendix C, we adapt the above to design maximally robust mechanisms with respect to non-Euclidean distances as well, so long as that distance can be expressed as a Bregman divergence (Eq. 9). Each such divergence has a corresponding scoring rule which one simply uses in the place of the quadratic score [Frongillo and Kash 2014, Appendix F].

8. DISCUSSION AND CONCLUSION

We have presented a new geometric perspective on minimal peer prediction mechanisms, and proved that it is without loss of generality to think of a minimal peer prediction mechanism as a power diagram. This perspective then allowed us to prove uniqueness of several well-known mechanisms up to positive-affine transformations, to construct novel peer prediction mechanisms for new conditions, to optimize for effort incentives within this space, to reason about the relative expressiveness of peer prediction methods, and to compute the mechanism that is maximally robust with respect to the agents' subjective belief models deviating from the mechanism designer's.

Several extensions of our model and results are straightforward, and are discussed further in Appendix D. For example, mechanisms that score an agent with the reports of multiple reference agents are still equivalent to power diagrams, but on a higher-dimensional belief space (all distributions on tuples of signals). Additionally, if agents are each given separate mechanisms M_i, essentially all of our main results go through, including the uniqueness of each M_i with respect to some belief model constraint $\mathcal{D}_i$.

We believe the most exciting direction for future work is to construct mechanisms from real-world data. One way to do this (aside, of course, from using gold standard data), is to use the geometric framework to learn Bayesian Truth Serum mechanisms from the agents' reports. In addition to the signal report, these mechanisms also elicit posterior reports, and with these (signal, posterior) pairs in hand, the mechanism designer can then train a classifier within the class of power diagrams that predicts the signal associated with a new posterior. For example, some formulations of multi-class Support Vector Machines naturally produce power diagrams [e.g. Borgwardt 2015], which correspond to a max-margin criterion as depicted in Figure 9R. Using Eq. 5, this power diagram can then be converted to a mechanism which is maximally robust with respect to the training set. (When the data are not separable, a soft-margin solution may be appropriate.) An alternative to the max-margin criterion may be to optimize effort incentives, which involves searching over the space of power diagrams consistent with the data and then positive-affine transformations for each. We explore these approaches in ongoing work.

ACKNOWLEDGMENTS

We thank anonymous reviewers for very useful comments and feedback.

REFERENCES

Charalambos D. Aliprantis and Kim C. Border. 2007. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer.

Franz Aurenhammer. 1987a. A criterion for the affine equivalence of cell complexes in R d and convex polyhedra in R d+1. *Discrete & Computational Geometry* 2, 1 (Dec. 1987), 49–64. DOI:http://dx.doi.org/10.1007/BF02187870

Franz Aurenhammer. 1987b. Power diagrams: properties, algorithms and applications. *SIAM J. Comput.* 16, 1 (1987), 78–96. http://epubs.siam.org/doi/pdf/10.1137/0216006

Franz Aurenhammer. 1987c. Recognising polytopical cell complexes and constructing projection polyhedra. *Journal of Symbolic Computation* 3, 3 (June 1987), 249–255. DOI:http://dx.doi.org/10.1016/S0747-7171(87)80003-2

Steffen Borgwardt. 2015. On Soft Power Diagrams. *Journal of Mathematical Modelling and Algorithms in Operations Research* 14, 2 (2015), 173–196.

Glenn W. Brier. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 78, 1 (1950), 1–3.

Anirban Dasgupta and Arpita Ghosh. 2013. Crowdsourced Judgement Elicitation with Endogenous Proficiency. In *Proceedings of the 22nd ACM International World Wide Web Conference (WWW'13)*. 319–330.

David Eberly. 2008. Centers of a Simplex. (2008).

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

Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378.

John C. Harsanyi. 1967–1968. Games with incomplete information played by Bayesian players. *Management Science* 16 (1967–1968), 159–182, 320–334, 486–502.

Scott Johnson, John W. Pratt, and Richard J. Zeckhauser. 1990. Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case. *Econometrica* 58, 4 (1990), 873–900.

Radu Jurca and Boi Faltings. 2007. Robust Incentive-Compatible Feedback Payments. In *Trust, Reputation and Security: Theories and Practice*. LNAI, Vol. 4452. Springer-Verlag, 204–218.

Radu Jurca and Boi Faltings. 2008. Incentives for Expressing Opinions in Online Polls. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC'08)*. 119–128.

Radu Jurca and Boi Faltings. 2011. Incentives for Answering Hypothetical Questions. In *Proceedings of the 1st Workshop on Social Computing and User Generated Content (SC'11)*.

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

Nicolas S. Lambert and Yoav Shoham. 2009. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the 10th ACM conference on Electronic commerce*. 109–118.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science* 51, 9 (2005), 1359–1373.

Frank Nielsen, Jean-Daniel Boissonnat, and Richard Nock. 2007. On bregman voronoi diagrams. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 746–755. http://dl.acm.org/citation.cfm?id=1283463

Martin J. Osborne and Ariel Rubinstein. 1994. *A Course in Game Theory* (seventh ed.). MIT Press.

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

Goran Radanovic. 2016. *Elicitation and Aggregation of Crowd Information*. Ph.D. Dissertation. Ecole Polytechnique Fédérale de Lausanne (EPFL).

Goran Radanovic and Boi Faltings. 2013. A Robust Bayesian Truth Serum for Non-binary Signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI'13)*. 833–839.

Konstantin Rybnikov. 1999. Stresses and Liftings of Cell-Complexes. *Discrete & Computational Geometry* 21, 4 (June 1999), 481–517. DOI:http://dx.doi.org/10.1007/PL00009434

Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C. Parkes. 2016. Informed Truthfulness in Multi-Task Peer Prediction. In *Proceedings of the 17th ACM Conference on Economics and Computation (EC'16)*. 179–196.

Jens Witkowski. 2014. *Robust Peer Prediction Mechanisms*. Ph.D. Dissertation. Department of Computer Science, Albert-Ludwigs-Universität Freiburg.

Jens Witkowski and David C. Parkes. 2012a. A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*. 1492–1498.

Jens Witkowski and David C. Parkes. 2012b. Peer Prediction Without a Common Prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12)*. 964–981.

Jens Witkowski and David C. Parkes. 2013. Learning the Prior in Minimal Peer Prediction. In *Proceedings of the 3rd Workshop on Social Computing and User Generated Content (SC'13)*.

A. PROOF OF LEMMA 2.5

PROOF. The result follows directly from the fact that the argmax from Definition 2.3 is unchanged by positive-affine transformations. That is, for all $i, j \in [n]$ and all $s_i, x_j \in [m]$, we have:

$$\begin{aligned} s_i &= \operatorname*{argmax}_{x_i} \mathop{\mathbf{E}}_{S_j}\Big[M\big(x_i, S_j\big) \;\Big|\; S_i = s_i\Big] \\ &= \operatorname*{argmax}_{x_i} \mathop{\mathbf{E}}_{S_j}\Big[\alpha M\big(x_i, S_j\big) \;+ f(x_j)\Big|\; S_i = s_i\Big] \\ &= \operatorname*{argmax}_{x_i} \mathop{\mathbf{E}}_{S_j}\Big[M'\big(x_i, S_j\big) \;\Big|\; S_i = s_i\Big], \end{aligned}$$

where the second line follows from the first because $f(x_j)$ does not depend on x_i. □

B. WORKING IN THE AFFINE HULL OF THE SIMPLEX

In this section we will justify the step in Sections 3 and 4 where we assume without loss of generality that we may work with power diagrams in $d = m - 1$ dimensions, where as always m is the number of signal values. Specifically, we show that we may assume WLOG that all sites $\boldsymbol{v}^s$ are in the affine hull $\mathcal{A} = \{\boldsymbol{u} \in \mathbb{R}^m : \sum_s \boldsymbol{u}(s) = 1\}$ of Δ_m. Then, as this affine subspace has dimension d, we can apply an isometry to restate the problem in $\mathbb{R}^d$ (which exists as they are both flats of the same dimension), so that we can work with $\hat{\boldsymbol{v}}^1, \dots, \hat{\boldsymbol{v}}^m, \hat{\boldsymbol{p}} \in \mathbb{R}^d$.

To begin, we show that assuming $\boldsymbol{v}^s \in \mathcal{A}$ is WLOG. Let $\boldsymbol{p} \in \Delta_m$ and $\boldsymbol{v}^s \in \mathbb{R}^m$, and define $\hat{\boldsymbol{v}}^s = \boldsymbol{v}^s + c(s)\mathbb{1}$ for $c(s) \in \mathbb{R}$ where $\mathbb{1} \in \mathbb{R}^m$ is the all-ones vector. Then we have

$$\begin{aligned} \|\hat{\boldsymbol{v}}^s - \boldsymbol{p}\|^2 &= \|\boldsymbol{v}^s + c(s)\mathbb{1}\|^2 - 2(\boldsymbol{v}^s + c(s)\mathbb{1}) \cdot \boldsymbol{p} + \|\boldsymbol{p}\|^2 \\ &= \|\boldsymbol{v}^s\|^2 + 2\boldsymbol{v}^s \cdot c(s)\mathbb{1} + \|c(s)\mathbb{1}\|^2 - 2\boldsymbol{v}^s \cdot \boldsymbol{p} - 2c(s) + \|\boldsymbol{p}\|^2 \\ &= \|\boldsymbol{v}^s - \boldsymbol{p}\|^2 + c(s)^2 m + 2c(s)(\boldsymbol{v}^s \cdot \mathbb{1} - 1)\ . \end{aligned}$$

Taking $c(s) = (1 - \boldsymbol{v}^s \cdot \mathbb{1})/m$, we have $\hat{\boldsymbol{v}}^s \cdot \mathbb{1} = 1$ so $\hat{\boldsymbol{v}}^s \in \mathcal{A}$. We now check that taking $\hat{w}(s) = w(s) - (1 - \boldsymbol{v}^s \cdot \mathbb{1})^2/m$ preserves power distances. From the above, we see that

$$\begin{aligned} \|\hat{\boldsymbol{v}}^s - \boldsymbol{p}\|^2 - \hat{w}(s) &= \|\boldsymbol{v}^s - \boldsymbol{p}\|^2 - w(s) + \tfrac{(1-\boldsymbol{v}^s\cdot\mathbb{1})^2}{m} + \left(\tfrac{1-\boldsymbol{v}^s\cdot\mathbb{1}}{m}\right)^2 m \\ &\qquad + 2\left(\tfrac{1-\boldsymbol{v}^s\cdot\mathbb{1}}{m}\right)(\boldsymbol{v}^s \cdot \mathbb{1} - 1) \\ &= \|\boldsymbol{v}^s - \boldsymbol{p}\|^2 - w(s)\ . \end{aligned}$$

Thus, we can move the sites to $\mathcal{A}$ and simply modify the weights to preserve the original power diagram on all of $\mathcal{A}$.

C. ROBUSTNESS FOR GENERAL (BREGMAN) DISTANCES

In this section, we generalize our definition of robustness to allow for any distance, and show that classical peer prediction mechanisms are maximally robust for the distance metric given by the Bregman divergence corresponding to the scoring rule used (see Section 6 and Eq. 9). As before, robustness will mean that the mechanism must be truthful as long as posteriors are "close" to the mechanism's model, but now distance will be measured via a general function.

Definition C.1. Given $\epsilon > 0$ and a nonnegative function $D(\boldsymbol{u}, \boldsymbol{v})$, a mechanism M is ϵ-D-robust with respect to belief model $p(\cdot|\cdot)$ if M is truthful for $p^*(\cdot|\cdot)$ whenever the following holds for all $s_i \in [m]$,

$$D(p^*(\cdot|s_i), p(\cdot|s_i)) < \epsilon\ . \tag{17}$$

Note in contrast to Definition 7.1, we drop the square for ϵ and adopt a strict inequality for convenience. Naturally, when D is a Bregman divergence, the mechanism which is maximally robust is the one generated from the same Bregman divergence.

THEOREM C.2. *Let $p(\cdot|\cdot)$ be the mechanism's belief model in classical peer prediction, and let D_G be a Bregman divergence defined on the simplex Δ_m. Then the classical peer prediction mechanism M using scoring rule $S(p,q) = D_G(q,p)$ is maximally robust with respect to distance $D = D_G$.*

PROOF. For brevity write $\boldsymbol{p}^s := p(\cdot|s)$. By Theorem 3.4, we may instead work with a power diagram with sites $\boldsymbol{v}^s$. Let d be the minimum value of $D(\boldsymbol{u}, \boldsymbol{p}^s)$ for any point $\boldsymbol{u}$ on some cell boundary and any $\boldsymbol{p}^s$. More formally, $d = \min_{s,s'} \min_{\boldsymbol{u} \in \Delta_m : D_G(\boldsymbol{u},\boldsymbol{p}^s) = D_G(\boldsymbol{u},\boldsymbol{p}^{s'})} D_G(\boldsymbol{u}, \boldsymbol{p}^s)$. Taking $\epsilon = d$, we now see that M is ϵ-D_G-robust, as $D_G(p^*(\cdot|s), \boldsymbol{p}^s) < \epsilon = d$ implies that $p^*(\cdot|s) \in \text{cell}(\boldsymbol{v}^s)$, since d is by definition is the smallest Bregman distance from any $\boldsymbol{p}^s$ to its cell boundary, so any smaller distance must stay within the cell. As this holds for all signals s, each posterior is in its correct cell and M is truthful.

To show maximality, let $s, s', \boldsymbol{u}$ be the signals and point achieving this minimum in the definition of d. For any $\epsilon > d$, we in particular have $D_G(\boldsymbol{u}, \boldsymbol{p}^s) = D_G(\boldsymbol{u}, \boldsymbol{p}^{s'}) < \epsilon$. Now take $p^*(\cdot|s) = p^*(\cdot|s') = \boldsymbol{u}$, and let $p^*(\cdot|s'') = p(\cdot|s'')$ for all other signals s''; while this choice of p^* clearly satisfies the ϵ-D_G-robustness condition (note that for any Bregman divergence, $D_G(\boldsymbol{v}, \boldsymbol{v}) = 0$ for all $\boldsymbol{v}$), no mechanism can be (strictly) truthful for p^*.[20] □

D. EXTENSIONS

In Section 8, we discuss two possible extensions of our model and results. The first is to mechanisms that score an agent with the reports of multiple reference agents. If the mechanism uses k such reference agents, the mechanism will no longer be a matrix, but a "tensor" of the form $M(s_i, s_{j_1}, s_{j_2}, \ldots, s_{j_k})$. To reason about their expected utility then, agents will form posterior beliefs of the form $p_i(s_{j_1}, s_{j_2}, \ldots, s_{j_k}|s)$, which is a distribution over m^k possible values for the k reference agent reports/signals. For example, with three signals $\{a, b, c\}$, and two reference agents, we would have $p_i(\cdot|s) \in \Delta_9$, a distribution over values $\{aa, ab, ac, ba, bb, bc, ca, cb, cc\}$ of the reports of the two peer agents.

As the agent's report space is still $[m]$, mechanisms are still equivalent to power diagrams, just on the higher-dimensional belief space Δ_{m^k} corresponding to all distributions on tuples of signals. Uniqueness would be "easier" to satisfy in the sense that the ambient dimension, in this case the $m^k - 1$ degrees of freedom, is now going to be higher than the number of sites of the power diagram, which is still m. Thus, the sites will "probably" be in general position; for example, with three signals and two peer agents, any mechanism corresponds to a power diagram with $m = 3$ sites in Δ_9, an 8-dimensional space, and uniqueness holds as long as the sites are not collinear. (Even if one requires symmetry in agent beliefs, i.e. that $p_i(ab|s) = p_i(ba|s)$, one still has a distribution on outcomes $\{aa, ab, ac, bb, bc, cc\}$, a 5-dimensional space.)

The other extension considered is for agents to face separate mechanisms $M_i(x_i, x_j)$. Here one could have a different belief model constraint $\mathcal{D}_i$ for each agent i, but essentially all of our main results go through, including the uniqueness of each M_i with respect to $\mathcal{D}_i$.

[20] This proof could be extended to show that for any $\epsilon > d$, there in fact are posteriors satisfying the robustness condition for which no mechanism can even be "weakly" (non strictly) truthful, as the overlap between the robustness regions will be of positive measure.