

Elicitation for Aggregation

Rafael M. Frongillo

Harvard University
raf@cs.berkeley.edu

Yiling Chen

Harvard University
yiling@seas.harvard.edu

Ian A. Kash

Microsoft Research
iankash@microsoft.com

Abstract

We study the problem of eliciting and aggregating probabilistic information from multiple agents. In order to successfully aggregate the predictions of agents, the principal needs to elicit some notion of *confidence* from agents, capturing how much experience or knowledge led to their predictions. To formalize this, we consider a principal who wishes to elicit predictions about a random variable from a group of Bayesian agents, each of whom have privately observed some independent samples of the random variable, and hopes to aggregate the predictions as if she had directly observed the samples of all agents. Leveraging techniques from Bayesian statistics, we represent confidence as the number of samples an agent has observed, which is quantified by a hyperparameter from a conjugate family of prior distributions. This then allows us to show that if the principal has access to a few samples, she can achieve her aggregation goal by eliciting predictions from agents using proper scoring rules. In particular, if she has access to one sample, she can successfully aggregate the agents' predictions if and only if every posterior predictive distribution corresponds to a unique value of the hyperparameter. Furthermore, this uniqueness holds for many common distributions of interest. When this uniqueness property does not hold, we construct a novel and intuitive mechanism where a principal with two samples can elicit and optimally aggregate the agents' predictions.

1 Introduction

Imagine that a principal, Alice, wishes to estimate the probability of rain tomorrow. She consults two agents, Bob who says 80%, and Carol who says 10%. How should Alice aggregate these two widely disparate predictions? If she knew that Bob happened to have spent the day studying radar imagery, whereas Carol just looked outside for a second, it would seem obvious that Alice should give much higher weight to Bob's prediction than Carol's. In other words, in order to aggregate these predictions, Alice needs to know the agents' *confidence* about their reports.

The aggregation of probabilistic information is an important problem in many domains, from multiagent systems to crowdsourcing. In this paper, we propose a general method of eliciting predictions together with a measure of confidence about those predictions, and show how to use this information to optimally aggregate in many situations.

We consider a Bayesian model where a principal, who can consult a group of risk-neutral agents, wishes to obtain an

informed prediction about a random variable. The random variable follows a parameterized distribution that is generated by some unknown parameters, the prior distribution of which is common knowledge. Each agent privately observes some independent samples of the random variable and forms a belief about it. The principal then elicits the agents' predictions of the random variable, and her goal is to optimally aggregate agents' private beliefs based on these predictions — to compute the distribution of the random variable as if she had observed the samples of all agents.

This paper focuses on designing elicitation mechanisms to achieve this optimal aggregation. We show that when the prior distribution of the unknown parameters comes from a *conjugate prior family* of the distribution of the random variable, the principal can *leverage a few independent samples* that she observes to successfully elicit enough information from the agents to achieve the optimal aggregation. This relies on important properties of the conjugate prior family. Intuitively, we use the hyperparameter of a distribution in the conjugate family to quantify the confidence of an agent's belief as the hyperparameter encodes information about the samples that the agent has observed. Our mechanisms work by eliciting predictions that allow the principal to infer the confidence of the agents and then make use of the confidence to achieve the optimal aggregation.

In particular, we prove that the principal can leverage a single sample to achieve optimal aggregation if and only if each distribution (modulo an equivalence relation) in the conjugate family maps to a unique hyperparameter. With this, we demonstrate how elicitation and optimal aggregation work for many common distributions of the random variable, including the Poisson, Normal, and uniform distributions, among others.

When the unique mapping condition is not satisfied, such as in the rain example above, we show that the hyperparameter of an agent's posterior distribution cannot be inferred with the principal's single sample. Fortunately, in this setting we construct a mechanism where the principal can still achieve the optimal aggregation if she has access to *two* independent samples of the random variable. Our mechanism simply asks each agent for his believed distribution of the first sample, and the likelihood that the two samples are the same. We show that this simple and intuitive approach gives the principal second-order information about agents' beliefs,

which is enough to achieve optimal aggregation.

1.1 Related Work

Our problem simultaneously considers both one-shot elicitation of information from multiple agents and the subsequent aggregation of the information.

In one-shot elicitation, the principal interacts with each agent independently and the agents report their predictions without knowing others' predictions. There is a rich literature on mechanisms for one-shot elicitation. The simplest is the classical proper scoring rules (Brier 1950; Winkler 1969; Savage 1971; Gneiting and Raftery 2007), which incentivize risk-neutral agents to honestly report their predictions. Proper scoring rules are the building blocks for most elicitation mechanisms, including our mechanisms in this paper. To reduce the total payment of the principal, researchers design shared scoring rules (Kilgour and Gerchak 2004; Johnstone 2007) and wagering mechanisms (Lambert et al. 2008; 2014; Chen et al. 2014) that have various desirable theoretical properties. Both shared scoring rules and wagering mechanisms engage agents in a one-shot betting to elicit their information and do not require the principal to subsidize the betting. In contrast to our problem, all these one-shot elicitation mechanisms do not consider the aggregation of the elicited information.

Sequential mechanisms have been designed to both elicit and aggregate information from agents. Most well known probably are prediction markets (Berg et al. 2001; Wolfers and Zitzewitz 2004), especially the market scoring rules mechanism (Hanson 2003; 2007), where agents can sequentially interact with the market mechanism for multiple times to reveal their information. Information aggregation happens when agents update their beliefs after observing other agents' activities in the market. However, the dynamic nature of these mechanisms can induce complicated strategic play and obfuscate individual-level information (Hansen, Schmidt, and Strobel 2001; Chen et al. 2010; Gao, Zhang, and Chen 2013). In this paper, we let the principal rather than the agents take the responsibility of aggregating information, and couple aggregation with one-shot elicitation that is incentive compatible for the agents.

To achieve optimal aggregation, the principal in our paper needs to know the confidence of agents' predictions. The work of Fang, Stinchcombe, and Whinston (2007) is the closest to ours in this perspective. They consider the one-shot elicitation of both agents' predictions and the precision of their predictions and then use the elicited precision to optimally aggregate. They use Normal distributions to model both the distribution of the random variable and the prior distribution of the unknown parameters. We consider general parameterized distributions of the random variable and their corresponding conjugate priors, which include the model of Fang, Stinchcombe, and Whinston (2007) as a special case.

2 Model and Background

We introduce our model, which describes how agents form their beliefs, the principal's elicitation mechanism, the principal's aggregation goal, and a family of parameterized prior distributions that we will focus on in this paper.

2.1 Beliefs of Agents

The principal would like to get information from m agents about a random variable with observable outcome space \mathcal{X} . The distribution of the random variable comes from a parameterized family of distributions $\{p(x|\theta)\}_{\theta \in \Theta} \subseteq \Delta_{\mathcal{X}}$, where Θ is the parameter space.¹ There exists a prior distribution $p(\theta)$ over the parameters. Both $\{p(x|\theta)\}_{\theta \in \Theta}$ and $p(\theta)$ are common knowledge to the agents and the principal.

Nature draws the true parameter θ^* , which is unknown to both the agents and the principal, according to the prior $p(\theta)$. Each agent then receives some number of samples from \mathcal{X} which drawn independently according to $p(x|\theta^*)$. In other words, if x_1, \dots, x_N is an enumeration of all samples received by any of the agents, then $p(x_i, x_j|\theta^*) = p(x_i|\theta^*)p(x_j|\theta^*)$ for all i, j and all $\theta^* \in \Theta$.

Agents form their beliefs about the random variable according to the Bayes' rule. If an agent receives samples x_1, \dots, x_N , then we write the agent's belief as

$$p = p(x|x_1, \dots, x_N) = \int_{\Theta} p(x|\theta)p(\theta|x_1, \dots, x_N)d\theta \\ \propto \int_{\Theta} p(\theta)p(x|\theta) \prod_j p(x_j|\theta)d\theta. \quad (1)$$

This distribution is known as the *posterior predictive distribution (PPD)* of x given samples x_1, \dots, x_N , and will be a central object of our analysis.

2.2 Elicitation and Scoring Rules

An important feature of our model is that the principal has access to a sample $x \in \mathcal{X}$ herself, and can leverage this sample using scoring rule techniques to elicit information from the agents. The principal's sample is also independently drawn according to $p(x|\theta^*)$. (In Section 4, we will allow the principal to have two such samples.)

The principal will choose a report space \mathcal{R} and a scoring mechanism $S : \mathcal{R} \times \mathcal{X} \rightarrow \mathbb{R}$, and request a report $r_i \in \mathcal{R}$ from each agent i . Upon receiving her sample x , the principal will give each agent a score of $S(r_i, x)$. We assume that agents seek to maximize their expected score, so that if agent i believes $x \sim p$ for some $p \in \Delta_{\mathcal{X}}$, then he will report $r_i \in \operatorname{argmax}_{r \in \mathcal{R}} \mathbb{E}_{x \sim p}[S(r, x)]$.

Strictly proper scoring rules (Brier 1950; Gneiting and Raftery 2007) are the basic tools for designing such scores S that provide good incentive properties. A scoring rule is strictly proper if and only if reporting one's true prediction uniquely maximizes the expected score. Strictly proper scoring rules are most commonly used for eliciting a distribution over a finite outcome space, but also extend naturally to eliciting distributions with continuous support (Matheson and Winkler 1976) and properties of distributions such as moments (Gneiting and Raftery 2007). For example, the logarithmic scoring rule

$$S(p, x) = \log p(x) \quad (2)$$

¹By convention $p(x|\dots)$ often refers to the entire distribution, rather than the density value at a particular x ; the usage should be clear from context.

is a popular strictly proper scoring rule for eliciting a distribution over a finite \mathcal{X} , where $p(x)$ is the reported probability for outcome x . Another popular strictly proper scoring rule, the Brier score (Brier 1950), can be used to elicit the mean of a random variable $\mathbb{E}[x]$, when taking the following form

$$S(r, x) = 2rx - r^2 \quad (3)$$

or the first k moments $(\mathbb{E}[x], \dots, \mathbb{E}[x^k])$, when used as

$$S(r_1, \dots, r_k, x) = \sum_{i=1}^k 2r_i x^i - r_i^2. \quad (4)$$

2.3 Aggregation

The goal of the principal is to aggregate the information of the agents to obtain an accurate distribution of the random variable as if she has access to all of the samples from all agents. Throughout the paper, we will denote by X this multiset² of all observed samples by agents.

Definition 1. Given prior $p(\theta)$ and data X distributed among the agents, the global posterior predictive distribution (global PPD) is the posterior predictive distribution $p(x|X)$.

The goal of this paper is to design mechanisms which truthfully elicit information from agents in such a way that the global PPD $p(x|X)$ can be computed. We capture this desideratum in the following definition.

Definition 2. Let $S : \mathcal{R} \times \mathcal{X} \rightarrow \mathbb{R}$ be given, and let each agent i receive samples X^i , with $X = \uplus_i X^i$ (multi-set addition). Let r_i be the report of agent i , namely $r_i = \operatorname{argmax}_r \mathbb{E}_{p(x|X^i)}[S(r, x)]$. Then S achieves optimal aggregation if there exists some function $g : \mathcal{R}^m \rightarrow \Delta_{\mathcal{X}}$ such that $g(r_1, \dots, r_m) = p(x|X)$.

It is worth noting that the report space \mathcal{R} of the elicitation mechanism is often different from the space of PPD, i.e. $\Delta_{\mathcal{X}}$. In fact, we will design elicitation mechanisms such that the elicited reports help the principal to infer the *confidence* of agents, capturing the amount of samples that the agents have experienced, which then enables the optimal aggregation. This leads to our focus on the conjugate prior family.

As a motivating example, consider the Normal distribution case, with $p(x|\theta) = \mathcal{N}(x, 1)$ and $p(\theta) = \mathcal{N}(\mu, 1)$, where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2 . It is well known that an agent i has posterior distribution $p(\theta|X_i) = \mathcal{N}((\mu + x_1 + \dots + x_{n_i})/(n_i + 1), 1/(n_i + 1))$ after observing samples $X_i = \{x_1, \dots, x_{n_i}\}$. His estimate of the mean of θ is the weighted sum of his sample and the prior mean. The inverse of the variance, $n_i + 1$, is called the *precision*, which encodes the agent's confidence or experience. Hence, if the principal can elicit mean estimate μ_i and precision $n_i + 1$ from each of the m agents, he can calculate the global PPD, which is a Normal distribution with mean $\frac{1}{N+1}(\mu + \sum_i n_i \mu_i)$ and variance $\frac{1}{N+1}$, where $N = \sum_i n_i$. This is the case studied by Fang, Stinchcombe, and Whinston (2007). We will see next that the general notion of conjugate priors will allow us to preserve the important aggregation properties we require elegantly.

²We use multisets, or equivalently unordered lists, as when \mathcal{X} is a finite set it is likely that samples will not be unique.

2.4 Conjugate Priors

In this paper, we focus on prior distributions $p(\theta)$ that come from the conjugate prior family for distributions $\{p(x|\theta)\}_{\theta \in \Theta}$. This ensures that the posterior distribution on θ is in the same family of distributions as the prior $p(\theta)$ and also simplifies the optimal aggregation problem.

While many notions of conjugate priors appear in the literature (Fink 1997; Gelman et al. 2013), we adopt the following definition, which says that the conjugate prior family is parameterized by *hyperparameters* ν and n which are *linearly* updated after observing samples: the new parameters can be written as a linear combination of the old parameters and sufficient statistics for the samples.

Definition 3. Let $P = \{p(x|\theta) : \theta \in \Theta\} \subseteq \Delta_{\mathcal{X}}$ be given. A family of distributions $\{p(\theta|\nu, n) : \nu \in \mathbb{R}^k, n \in \mathbb{R}_+\} \subseteq \Delta_{\Theta}$ is a conjugate prior family for P if there exists a statistic $\phi : \mathcal{X} \rightarrow \mathbb{R}^k$ such that, given the prior distribution $p(\theta|\nu_0, n_0)$, the posterior distribution on θ after observing x ,

$$p(\theta|\nu_0, n_0, x) = \frac{p(\theta|\nu_0, n_0)p(x|\theta)}{\int_{\Theta} p(\theta'|\nu_0, n_0)p(x|\theta')d\theta'}, \quad (5)$$

is equal to $p(\theta|\nu_0 + \phi(x), n_0 + 1)$ for all ν_0 and n_0 .

Using conjugate priors, the optimal aggregation problem simplifies considerably. Given prior $p(\theta|\nu_0, n_0)$ and data $X = \{x_1, \dots, x_N\}$ distributed among the agents, the global PPD can be written succinctly as

$$p(x|\nu_0, n_0, X) = p\left(x \mid \nu_0 + \sum_{i=1}^N \phi(x_i), n_0 + N\right). \quad (6)$$

We can see that as we require n to update by 1 for each additional sample, $n - n_0$ exactly corresponds to the number of samples seen in total. This is precisely the notion of confidence we wish to quantify — the amount of data or experience that led to a prediction. In particular, if we could obtain the hyperparameters (ν_i, n_i) for an agent's report, we could directly compute the number of samples $N_i = n_i - n_0$ they observed, as well as the sum of the sufficient statistics of their samples, $\sum_{x \in X^i} \phi(x)$. If the principal can gather these two quantities from each agent i , then using the identities $\sum_{x \in X^i} \phi(x) = \nu_i - \nu_0$ and $N_i = n_i - n_0$, the principal can aggregate these parameters by the observation that

$$\begin{aligned} \sum_{i=1}^N \phi(x_i) &= \sum_{i=1}^m \sum_{x \in X^i} \phi(x) = \sum_{i=1}^m (\nu_i - \nu_0) \quad (7) \\ N &= \sum_{i=1}^m N_i = \sum_{i=1}^m (n_i - n_0). \quad (8) \end{aligned}$$

From here, the principal simply plugs these values into eq. (6) to obtain the global PPD.

3 Unique Predictive Distributions

In this section, we show how the principal can leverage a single sample $x \in \mathcal{X}$ to elicit the hyperparameters of the posterior distributions of the agents, provided that the mapping from hyperparameters to predictive posterior distributions is unique. Note that this statement contains two different types of posterior distributions, and as the distinction is

important we take a moment to recall their differences. After making his observations, an agent will have updated his hyperparameters to (ν, n) . This gives him a *posterior* distribution $p(\theta|\nu, n)$ over the parameter of the random variable and a *predictive posterior* distribution (PPD) $p(x|\nu, n)$ of the random variable itself.

We begin with two simple but important results. The first is an analog of the revelation principle from economic theory, showing that the most a principal with a single sample $x \in \mathcal{X}$ can get from an agent is the agent's private belief $p \in \Delta_{\mathcal{X}}$ about x .

Lemma 1. *Given a sample $x \in \mathcal{X}$ which an agent believes to be drawn from $p \in \Delta_{\mathcal{X}}$, any information obtained with a mechanism $S : \mathcal{R} \times \mathcal{X} \rightarrow \mathbb{R}$, from an agent maximizing his expected score, can be written as a function of p .*

Proof. We need only find a function $f : \Delta_{\mathcal{X}} \rightarrow \mathcal{R}$ such that $f(p) \in \operatorname{argmax}_{r \in \mathcal{R}} \mathbb{E}_{x \sim p}[S(r, x)]$ whenever the argmax exists. Let $r_0 \in \mathcal{R}$ be arbitrary. For all $p \in \Delta_{\mathcal{X}}$, simply select $r_p \in \operatorname{argmax}_{r \in \mathcal{R}} \mathbb{E}_{x \sim p}[S(r, x)]$, or $r_p = r_0$ if the argmax is not defined, and let $f(p) = r_p$. \square

While intuitive and almost obvious, Lemma 1 is quite useful when thinking about elicitation problems. For example, in our setting it is certainly clear that the principal can take $\mathcal{R} = \Delta_{\mathcal{X}}$ and use any strictly proper scoring rule to get the agent's PPD $p(x|\nu, n)$. One might be tempted, however, to try to get more information: if one could simply elicit the posterior $p(\theta|\nu, n)$, then the hyperparameters (ν, n) would be readily available for aggregation. One tantalizing scheme would be to compute the distribution $p(\theta|x)$ and draw a sample $\hat{\theta} \sim p(\theta|x)$, and then use this $\hat{\theta}$ to elicit $p(\theta|\nu, n)$ using the log scoring rule (2). Lemma 1 says that, while this may succeed, it will only succeed when the principal could have simply computed $p(\theta|\nu, n)$ from the PPD $p(x|\nu, n)$ to begin with.

For precisely this reason, we will see that being able to map the PPD to the posterior distribution is crucial to being able to optimally aggregate. Before proving this, we need to introduce some more precise notation to describe the relationship between the hyperparameters and the PPD.

Definition 4. *Given hyperparameters (ν_0, n_0) , we say (ν, n) is reachable from (ν_0, n_0) if there exists a multiset X of \mathcal{X} such that $\nu = \nu_0 + \sum_{x \in X} \phi(x)$ and $n = n_0 + |X|$. Additionally, we define the relation $(\nu, n) \equiv (\nu', n')$ if for all such X , including \emptyset , we have $p(x|\nu, n, X) = p(x|\nu', n', X)$.*

Theorem 2. *Given a family of distributions $\{p(x|\theta)\}$ and conjugate prior $p(\theta|\nu_0, n_0)$, there exists a mechanism S achieving optimal aggregation if and only if for all (ν, n) and (ν', n') reachable from (ν_0, n_0) we have that $p(x|\nu, n) = p(x|\nu', n')$ implies $(\nu, n) \equiv (\nu', n')$.*

Proof. We first prove the if direction. Let S be the log scoring rule (2); then by propriety, the principal elicits $p_i = p(x|\nu_0, n_0, X_i) = p(x|\nu_i, n_i)$ for all i . From p_i the principal cannot necessarily compute (ν_i, n_i) , but she can choose some (ν'_i, n'_i) reachable from (ν_0, n_0) such that $p_i = p(x|\nu'_i, n'_i)$. We will show that since $(\nu_i, n_i) \equiv (\nu'_i, n'_i)$, this is enough to optimally aggregate. We will restrict to the

case of two agents; the rest then follows by induction. Let $\phi(X) = \sum_{x \in X} \phi(x)$; by reachability, we have X'_1, X'_2 such that $\nu'_i = \nu_0 + \phi(X'_i)$ and $n'_i = n_0 + |X'_i|$. Thus,

$$\begin{aligned} & p(x|\nu_0 + \sum_i (\nu'_i - \nu_0), n_0 + \sum_i (n'_i - n_0)) \\ &= p(x|\nu'_2 + (\nu'_1 - \nu_0), n'_2 + (n'_1 - n_0)) \\ &= p(x|\nu'_2 + \phi(X'_1), n'_2 + |X'_1|) \\ &\stackrel{*}{=} p(x|\nu_2 + \phi(X'_1), n_2 + |X'_1|) \\ &= p(x|\nu_2 + (\nu'_1 - \nu_0), n_2 + (n'_1 - n_0)) \\ &= p(x|\nu'_1 + (\nu_2 - \nu_0), n'_1 + (n_2 - n_0)) \\ &= p(x|\nu'_1 + \phi(X_2), n'_1 + |X_2|) \\ &\stackrel{*}{=} p(x|\nu_1 + \phi(X_2), n_1 + |X_2|) \\ &= p(x|\nu_1 + (\nu_2 - \nu_0), n_1 + (n_2 - n_0)) \\ &= p(x|\nu_0 + \sum_i (\nu_i - \nu_0), n_0 + \sum_i (n_i - n_0)), \end{aligned}$$

which is the global PPD. The starred equations used the fact that $(\nu_i, n_i) \equiv (\nu'_i, n'_i)$.

For the only-if direction, assume that there are X, X' such that for $\nu = \nu_0 + \phi(X)$ and $\nu' = \nu_0 + \phi(X')$, we have $p(x|\nu, n) = p(x|\nu', n')$ but $(\nu, n) \not\equiv (\nu', n')$. Then we have some multiset X_1 of \mathcal{X} such that $p(x|\nu, n, X_1) \neq p(x|\nu', n', X_1)$. Now let agent 1 receive X_1 , and consider two worlds, one in which $X_2 = X$ and the other in which $X_2 = X'$. By Lemma 1, without loss of generality, the principal uses S to elicit the PPD from both agents. However, she cannot distinguish between these two worlds, as by assumption agent 2's PPD is the same in both. Unfortunately, the global PPDs in these two situations are different:

$$\begin{aligned} p(x|\nu_0, n_0, X_1 \uplus X) &= p(x|\nu, n, X_1) \\ &\neq p(x|\nu', n', X_1) \\ &= p(x|\nu_0, n_0, X_1 \uplus X'). \end{aligned}$$

Hence, the principal is unable to optimally aggregate. \square

An important corollary of Theorem 2, which we will make extensive use of below, is that the principal can always optimally aggregate if the PPD gives her full information about the hyperparameters.

Corollary 3. *If the map $\varphi : (\nu, n) \mapsto p(x|\nu, n)$ is injective, the principal can optimally aggregate.*

Proof. By injectivity, $p(x|\nu, n) = p(x|\nu', n')$ implies $(\nu, n) = (\nu', n')$, and \equiv is an equivalence relation. Moreover, any strictly proper scoring rule S suffices as the mechanism, as this will elicit the PPD p , and then the principal can compute $(\nu, n) = \varphi^{-1}(p)$. \square

In the following, we provide several examples illustrating the utility of Theorem 2, and Corollary 3 in particular. Before continuing, however, we would like to remark on some practical considerations. Strictly speaking, the mechanism given by Corollary 3, which elicits the PPD and inverts the map φ , suffices when the modeling assumptions are all correct. However, in the case where the model is slightly off, be it in our conditional independence assumption, the core family $p(x|\theta)$, or even the particular choice of prior,

this approach appears to provide no guarantees. In the examples that follow, we seek not only to elicit the hyperparameters of the PPD, but to do so using scoring rules which provide meaningful information about the PPD *regardless of its form*. For example, we show below how to elicit the PPD for the Poisson distribution with a Gamma prior using a scoring rule for the first and second moment (or equivalently, the mean and variance). This scoring rule has the property that it will elicit the correct moments of *any* distribution, and thus if the agents' PPD does not have the assumed form, a practitioner would still have meaningful information about the agent's belief for a variety of approximate aggregation techniques.

Poisson Imagine that a citizen science project such as eBird (Sullivan et al. 2009) wishes to collect observations about sightings of various birds to deduce bird migration patterns. Such a project may wish users to report the number of birds of a particular species seen per minute. Of course, to combine such estimates, eBird would like to know not only the observed rate, but how long the user spend bird watching, so that it may weigh more highly reports from longer time intervals; this is precisely what our approach offers.

For situations such as this one which involve counting events in a specified time interval, the Poisson distribution is a common choice. The parameter of the Poisson distribution is $\lambda \in \mathbb{R}$, the *rate* parameter, and the probability of observing $x \in \{0, 1, 2, \dots\}$ events in a unit time interval is given by $p(x|\lambda) = \lambda^x e^{-\lambda}/x!$. The canonical conjugate prior for the Poisson distribution is the Gamma distribution, given by $p(\lambda|\nu, n) = \frac{n^\nu}{\Gamma(\nu)} \lambda^{\nu-1} e^{-n\lambda}$, and the statistic is $\phi(x) = x$. The form of the PPD $p(x|\nu, n)$ is also a familiar distribution, in the negative binomial family (Gelman et al. 2013, p.44).

As mentioned above, we will show how to compute the hyperparameters ν and n of the PPD from its first two moments μ_1 and μ_2 . As the form of the PPD is known to be negative binomial, one can easily calculate or look up what these moments are in terms of the hyperparameters: $\mu_1 = \nu/n$ and $\mu_2 = \nu(\nu + n + 1)/n^2$. Fortunately, given these equations, we can simply solve for the hyperparameters in terms of the moments, which we can elicit robustly: $n = \mu_1/(\mu_2 + \mu_1^2 + \mu_1)$ and $\nu = n\mu_1$. This already verifies the injectivity condition of Corollary 3, so we know that optimal aggregation is possible.

For concreteness, let us return to the bird watching example to show how eBird might reward users in such a way as to truthfully obtain predictions and then compute their optimal aggregation. The protocol would be for eBird to announce that a representative will be sent tomorrow to count the number x of birds seen in a minute, and to ask each user i for a prediction $r_{i,1}$ about $\mathbb{E}[x]$ and $r_{i,2}$ about $\mathbb{E}[x^2]$, with the understanding that after the count x is revealed, agent i will receive a reward (cf. (4)) of

$$S(r_{i,1}, r_{i,2}, x) = 2r_{i,1}x - r_{i,1}^2 + 2r_{i,2}x^2 - r_{i,2}^2. \quad (9)$$

With the reports in hand, eBird can compute $n_i = r_{i,1}/(r_{i,2} + r_{i,1}^2 + r_{i,1})$ and $\nu_i = n_i r_{i,1}$. Assuming the common prior parameters (ν_0, n_0) are known, eBird simply aggregates these reports to $n = n_0 + \sum_{i=1}^m (n_i - n_0)$

and $\nu = \nu_0 + \sum_{i=1}^m (\nu_i - \nu_0)$, arriving at the global PPD $p(x|X) = p(x|\nu, n)$.

Normal As we saw in Section 2, the Normal distribution with known variance but unknown mean allows for optimal aggregation. This follows easily from Corollary 3 as well, since $N(\mu, \sigma^2)$ is a different distribution for each setting of μ, σ .

Uniform Perhaps the most natural of distributions is the uniform distribution on $[0, \theta]$, where $p(x|\theta) = 1/\theta$ in that interval. As a simple application, consider the problem of determining the number of raffle tickets sold at a fair by asking random people what their ticket number is. It is well-known that the Pareto distribution is a conjugate prior for this case, and the hyperparameter update is $\nu = \max(\nu_0, x)$ and $n = n_0 + 1$. Observe that the hyperparameter update is not linear, so we cannot simply apply Corollary 3. However, it is easy to see that the conclusion still holds here, as the principal can easily aggregate $\{(\nu_i, n_i)\}_i^m$ by taking $\nu = \max\{\nu_i\}_{i=0}^m$ and $n = n_0 + \sum_i (n_i - n_0)$ as usual.

By a simple calculation, one can show that the PPD in this case is a mixture of a uniform distribution and a Pareto distribution, from which one can compute the moments $\mu_1 = n\nu/2(n-1)$ and $\mu_2 = n\nu^2/3(n-2)$. Cancelling ν , these equations give a quadratic equation with a unique root n satisfying $n > 2$ (a requirement of the prior), from which ν can also be calculated. Thus, the principal can achieve optimal aggregation in this case as well.

4 The Non-Unique Case

Imagine a setting where the principal wants to aggregate information from agents to estimate the bias of a coin. The principal asks agents Bob and Carol, who each see some unknown number of coin flips, after which Bob reports that the coin is unbiased, whereas Carol reports that it is biased 10-to-1 toward Heads. With only this information, which corresponds to the full PPDs of both agents, it is easily seen to be *impossible* to optimally aggregate these reports, as it is unclear how many flips each agent saw. Even if the principal knows that Carol saw 20 flips, she cannot tell whether Bob saw none and just reported the prior, or whether he saw 1000 and is practically certain of the bias of the coin. (Formally, we can explain this by noting that the conjugate prior is the Beta distribution, which does not satisfy Theorem 2.) How can the principal circumvent this impossibility to still achieve optimal aggregation in this setting?

In this section we will consider a more general version of the coin flip example, using the *categorical* family of distributions, i.e., the whole of $\Delta_{\mathcal{X}}$ for $\mathcal{X} = [K] = \{1, 2, \dots, K\}$. Here the common conjugate prior is the Dirichlet distribution $p(\theta|\alpha)$, whose hyperparameters $\alpha \in \mathbb{R}^K$ encode *pseudo-counts*, so that α_i corresponds to the number of occurrences of outcome i an agent has seen. More formally, we take $\Theta = \Delta_{\mathcal{X}} = \Delta_K$, and for $\alpha \in \mathbb{R}^K$ we let

$$p(i|\theta) = \theta_i, \quad p(\theta|\alpha) = \frac{\Gamma(n)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i-1}, \quad (10)$$

where $n = \sum_{i=1}^K \alpha_i$ corresponds to the total number of (pseudo-) samples observed, and Γ is the Gamma distribution.³ It is well-known that the mean of the Dirichlet distribution is $\mathbb{E}[\theta|\alpha] = \alpha/n$, which is just a normalized version of the pseudo-counts. Taken as an element of $\Delta_{\mathcal{X}}$, this is also the PPD: if an agent sees $x = 1$ and $x = 2$ each eight times and $x = 3$ four times, then $\alpha = (8, 8, 4)$ and his PPD will be $(2/5, 2/5, 1/5)$. We can see now why Theorem 2 tells us that optimal aggregation is impossible: scaling α by any positive amount yields the same PPD, just as with the coin flip example above, but when aggregating α 's from multiple agents, different relative scales yield different global PPDs.

Fortunately, despite this impossibility, we now show that if the principal can simply obtain *two* of her own samples, she can use them both to glean second-order information from the agents, and then optimally aggregate. The idea behind the mechanism is extremely simple: ask the agent for the distribution p of the first sample, and the probability b that the two samples are the same. As discussed above, the reported p gives α/n , and it turns out that the scaling factor n , which corresponds to the confidence of the agent, can be expressed as a simple formula of p and b .

Theorem 4. *Let $\mathcal{X} = [K]$, and let $\{p(i|\theta)\}$ and $\{p(\theta|\alpha)\}$ be the categorical and Dirichlet families from eq. (10). Then given two independent samples $x_1, x_2 \in \mathcal{X}$, the mechanism $S : \Delta_{\mathcal{X}} \times [0, 1] \times \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ defined by*

$$S(p, b, x_1, x_2) = \log p(x_1) + 2b \cdot \mathbb{1}\{x_1 = x_2\} - b^2 \quad (11)$$

achieves optimal aggregation.

Proof. Focusing first on a single agent, by propriety of the log scoring rule, the agent will report $p = p(\cdot|\alpha) = \alpha/n$, where once again $n = \sum_{i=1}^K \alpha_i$. Similarly, by propriety of the Brier score, the agent will report his belief about the probability that $x_1 = x_2$. We can calculate this easily:

$$\begin{aligned} b &= \Pr[x_1 = x_2] \\ &= \mathbb{E}_{\theta \sim p(\theta|\alpha)} \left[\sum_{i=1}^K p(x_1 = i, x_2 = i | \theta) \right] \\ &= \mathbb{E}_{\theta \sim p(\theta|\alpha)} \left[\sum_{i=1}^K p(x_1 = i | \theta) p(x_2 = i | \theta) \right] \\ &= \mathbb{E}_{\theta \sim p(\theta|\alpha)} \left[\sum_{i=1}^K \theta_i \theta_i \right] = \sum_{i=1}^K \text{Var}[\theta_i|\alpha] + \mathbb{E}[\theta_i|\alpha]^2. \end{aligned}$$

It is known that $\text{Var}[\theta_i|\alpha] = \frac{\alpha_i(n-\alpha_i)}{n^2(n+1)}$, so the first term becomes

$$\sum_i \text{Var}[\theta_i|\alpha] = \frac{(\sum_i \alpha_i)n - \sum_i \alpha_i^2}{n^2(n+1)} = \frac{1 - \|p\|^2}{n+1},$$

as we also have $\sum_i \mathbb{E}[\theta_i|\alpha]^2 = \|p\|^2 = \|\alpha\|^2/n^2$. Putting this together, we have $b = \frac{1 - \|p\|^2}{n+1} + \|p\|^2$, so $n = \frac{1-b}{b - \|p\|^2}$ and finally $\alpha = np$. Finally, turning to the aggregation of multiple predictions, the result follows by the same argument as in Theorem 2: we simply discount the prior from each agent's report and sum. \square

³Note that we have departed from our (ν, n) notation to match the convention for the Dirichlet distribution; otherwise we could take ν to be the first $K - 1$ coordinates of α , and keep n the same.

Returning to the coin flip example, we can now see how the principal can resolve the dilemma from before. Instead of simply asking the probability that a single flip is Heads, the principal should obtain two independent flips and then ask the agents for the probability that the first is Heads, and the probability that the two flips are the same. By Theorem 4, the answers to these two intuitive questions give the principal enough information to optimally aggregate.

5 Future Work

A well known and broad class of distributions with conjugate priors are the exponential families (see Appendix A for a primer). Many of the examples discussed in this paper fall into the exponential families, and thus it is a natural question to ask whether our results can be shown to hold for all such distributions. In particular, our study opens two interesting questions, which under the surface would imply some interesting structure of exponential families.

The first follows naturally from Theorem 2 and the examples in Section 3, several of which are exponential families, and all of which admit optimal aggregation. We conjecture that for exponential families, the success of a single-sample mechanism depends only on the dimension k of the statistic ϕ .

Conjecture 1. *Optimal aggregation with a single sample is possible for an exponential family if and only if $|\mathcal{X}| > \dim \phi + 1$.*

The second open question is similar: does the two-sample technique from Section 4 succeed for all exponential families? Again, we conjecture positively.

Conjecture 2. *Given an exponential family with statistic ϕ , the mechanism which elicits the expected values of $\phi(x_1)$ and $\phi(x_1)\phi(x_2)^\top$ can optimally aggregate.*

The intuition behind these conjectures, which we outline in Appendix B, lies in concentration properties in the posterior distribution $p(x|\nu, n)$ as n increases to infinity. Because of the simple form of exponential families, and the exponential decay inherent in their definition, we believe that these results can be obtained.

Finally, we would like to mention a possible extension. While our model assumes that the principal wishes to aggregate *all* information, in reality, agents may have different costs to gather their samples, and the principal may therefore desire to aggregate a more efficient amount of information given this cost. Fang, Stinchcombe, and Whinston (2007) show that this can be done in a restricted setting with Normal distributions. Can this still be done in our more general setting? What if agents can acquire different amounts of information at different costs, for example, if a convex function specifies their cost to acquire any number of samples? We hope to address these and related questions in future work.

References

- Berg, J. E.; Forsythe, R.; Nelson, F. D.; and Rietz, T. A. 2001. Results from a dozen years of election futures markets research. In Plott, C. A., and Smith, V., eds., *Handbook of Experimental Economic Results*.

- Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78(1):1–3.
- Chen, Y.; Dimitrov, S.; Sami, R.; Reeves, D. M.; Pennock, D. M.; Hanson, R. D.; Fortnow, L.; and Gonen, R. 2010. Gaming prediction markets: Equilibrium strategies with a market maker. *Algorithmica* 58(4):930–969.
- Chen, Y.; Devanur, N. R.; Pennock, D. M.; and Vaughan, J. W. 2014. Removing arbitrage from wagering mechanisms. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, 377–394.
- Diaconis, P.; Ylvisaker, D.; et al. 1979. Conjugate priors for exponential families. *The Annals of statistics* 7(2):269–281.
- Fang, F.; Stinchcombe, M.; and Whinston, A. 2007. Putting your money where your mouth is—a betting platform for better prediction. *Review of Network Economics* 6(2).
- Fink, D. 1997. A compendium of conjugate priors. *Unpublished*.
- Gao, X. A.; Zhang, J.; and Chen, Y. 2013. What you jointly know determines how you act: Strategic interactions in prediction markets. In *ACM Conference on Electronic Commerce*, EC '13, 489–506.
- Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, D. B.; Vehtari, A.; and Rubin, D. B. 2013. *Bayesian data analysis*. CRC press.
- Gneiting, T., and Raftery, A. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477):359–378.
- Hansen, J.; Schmidt, C.; and Strobel, M. 2001. Manipulation in political stock markets — preconditions and evidence. Technical Report.
- Hanson, R. D. 2003. Combinatorial information market design. *Information Systems Frontiers* 5(1):107–119.
- Hanson, R. D. 2007. Logarithmic market scoring rules for modular combinatorial information aggregation. *Journal of Prediction Markets* 1(1):1–15.
- Johnstone, D. J. 2007. The parimutuel Kelly probability scoring rule. *Decision Analysis* 4(2):66–75.
- Kilgour, D. M., and Gerchak, Y. 2004. Elicitation of probabilities using competitive scoring rules. *Decision Analysis* 1(2):108–113.
- Lambert, N.; Langford, J.; Wortman, J.; Chen, Y.; Reeves, D. M.; Shoham, Y.; and Pennock, D. M. 2008. Self-financed wagering mechanisms for forecasting. In *ACM Conference on Electronic Commerce*, 170–179. New York, NY, USA: ACM.
- Lambert, N.; Langford, J.; Vaughan, J. W.; Chen, Y.; Reeves, D. M.; Shoham, Y.; and Pennock, D. M. 2014. An axiomatic characterization of wagering mechanisms. *Journal of Economic Theory*. (Forthcoming).
- Matheson, J. E., and Winkler, R. L. 1976. Scoring rules for continuous probability distributions. *Management Science* 22(10):1087–1096.
- Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* 66(336):783–801.
- Sullivan, B. L.; Wood, C. L.; Iliff, M. J.; Bonney, R. E.; Fink, D.; and Kelling, S. 2009. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142(10):2282–2292.
- Wainwright, M. J., and Jordan, M. I. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 1(1-2):1–305.
- Winkler, R. L. 1969. Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association* 64(327):1073–1078.
- Wolfers, J., and Zitzewitz, E. 2004. Prediction markets. *Journal of Economic Perspective* 18(2):107–126.

A Exponential families

Perhaps the most important class of distributions which admit conjugate priors are the exponential families, a broad class which includes many common distributions such as normal, log-normal, Poisson, and many more. We briefly review exponential families and their conjugate priors, which as it turns out are themselves exponential families.

Let $\phi : \mathcal{X} \rightarrow \mathbb{R}^k$ be the *sufficient statistic* (a term justified below). We assume that ϕ is *minimal*, meaning $\langle \theta, \phi(x) \rangle$ cannot be a constant function of x for any $\theta \neq 0$. (Minimality is thus equivalent to affine independence.) Now define

$$g(\theta) = \log \int_{\mathcal{X}} \exp\{\langle \phi(x), \theta \rangle\} dx \quad (12)$$

$$p(x|\theta) = \exp\{\langle \phi(x), \theta \rangle - g(\theta)\}. \quad (13)$$

This family $\{p(x|\theta)\}$ is the exponential family with respect to ϕ . We refer to Θ as the *natural parameters*, as contrast to the *mean parameters* $\mu(\theta) = \mathbb{E}[\phi|\theta]$, which also parameterize the family provided certain regularity conditions are met (Wainwright and Jordan 2008). The function g is called the *cumulant*, and happens to generate the moments of ϕ under $p(x|\theta)$. In particular, we have $\nabla g(\theta) = \mathbb{E}[\phi|\theta] = \mu(\theta)$.

Turning now to the conjugate prior for this family, let

$$h(\nu, n) = \log \int_{\Theta} \exp\{\langle \theta, \nu \rangle - ng(\theta)\} d\theta \quad (14)$$

$$p(\theta|\nu, n) = \exp\{\langle \theta, \nu \rangle - ng(\theta) - h(\nu, n)\}. \quad (15)$$

One can verify directly that $p(\theta|\nu, n)$ is a conjugate family to $p(x|\theta)$. Moreover, the priors are themselves exponential families, with respect to statistic $\psi(\theta) = \begin{bmatrix} \theta \\ -g(\theta) \end{bmatrix}$.

As we saw above, it is easy to verify by direct calculation that the cumulant $g(\theta)$ satisfies $\nabla g(\theta) = \mathbb{E}_{x \sim p(x|\theta)}[\phi(x)]$. A much less obvious fact, but a very useful one, is that the implied mean ν/n of the conjugate prior $p(\theta|\nu, n)$ is *credible*, in the sense that the expected value of ϕ is in fact ν/n .

Theorem 5 ((Diaconis, Ylvisaker, and others 1979)). *Let $p(x|\theta)$ be an exponential family with cumulant $g(\theta)$ and let $p(\theta|\nu, n)$ be its conjugate prior. Then*

$$\int_{\mathcal{X}} \phi(x) p(x|\nu, n) dx = \int_{\Theta} \nabla g(\theta) p(\theta|\nu, n) = \nu/n. \quad (16)$$

B Conjectures for Exponential Families

Here we give intuition for the conjectures stated in Section 5. For the first, Conjecture 1, note that when $\dim \phi = |\mathcal{X}| - 1$, and the statistic is minimal, then Θ is just a reparameterization of the categorical distributions, $\Delta_{\mathcal{X}}$. As we saw in Section 4 that a single sample is insufficient for the categorical case, Conjecture 1 is implied by the following alternate conjecture.

Conjecture 3. *The map $\varphi : (\nu, n) \mapsto p(x|\nu, n)$ is injective for an exponential family conjugate prior if and only if $\dim \phi < |\mathcal{X}| - 1$.*

There is considerable intuition for this conjecture. By Theorem 5 (the credible mean property of exponential family conjugate priors), to examine the injectivity of φ we may

restrict our attention to a fixed value of $\mu = \nu/n$. This is because if $\nu/n \neq \nu'/n'$, then $\varphi(\nu, n) \neq \varphi(\nu', n')$. Given this fact, it is clear that the injectivity cannot hold whenever $k = \dim \phi \geq |\mathcal{X}| - 1$, because by minimality of ϕ , the mean $\mathbb{E}[\phi] = \nu/n = \mu$ must uniquely identify the distribution, and thus scaling n and taking $\nu = n\mu$ and yields the same PPD for all $n > 0$. Conversely, one can show by the form of the conjugate prior (15) that for our fixed value of μ , we have

$$\frac{p(\theta|n\mu, n)}{p(\theta'|n\mu, n)} = \left(\frac{p(\theta|\mu, 1)}{p(\theta'|\mu, 1)} \right)^n, \quad (17)$$

for all $\theta, \theta' \in \Theta$ and all n, μ . Thus, as n increases there is strong concentration in the prior about the mode $\hat{\theta}$, which one can show is equal to $\nabla g^*(\nu/n)$ by convex conjugacy, so that $\mu(\hat{\theta}) = \mu$. It is clear then that the limit of $p(x|n\mu, n)$ as $n \rightarrow \infty$ is simply $p(x|\hat{\theta})$. It would therefore be natural to show that $\text{KL}(p(x|\nu, n); p(x|\hat{\theta}))$, or some other notion of distance, is monotone decreasing in n , which would then imply injectivity of φ .

For Conjecture 2, the intuition lies in a reparameterization of the conjugate prior distribution. Let $\mu(\theta) = \nabla g(\theta)$ denote the mean parameter corresponding to θ , and recall from the credible mean property that $\mathbb{E}[\phi(x)|\nu, n] = \mathbb{E}[\mu(\theta)|\nu, n] = \nu/n$. Then by independence of x_1, x_2 , we have $\mathbb{E}[\phi(x_1)\phi(x_2)^\top|\nu, n] = \mathbb{E}[\mu(\theta)\mu(\theta)^\top|\nu, n]$. Thus, letting r_1 and R_2 be the reported values for the $\mathbb{E}[\phi(x_1)]$ and $\mathbb{E}[\phi(x_1)\phi(x_2)^\top]$, we see that $\text{Var}[\mu(\theta)|\nu, n]$ is simply $R_2 - r_1 r_1^\top$. In other words, we can use this information to compute the variance of the posterior distribution of the mean parameters. That is, if we thought of the posterior as being a distribution $p(\mu|\nu, n)$ over mean parameters instead of over natural parameters θ , we would be able to elicit the variance of this posterior. Intuitively, this variance should correspond to the confidence of the agent, and in particular should be monotone decreasing in n , which would allow us to compute n and thus optimally aggregate.