The Wisdom of Multiple Guesses

JOHAN UGANDER, Microsoft Research, Redmond, WA; Stanford University, Stanford, CA
RYAN DRAPEAU, University of Washington, Seattle, WA
CARLOS GUESTRIN, University of Washington, Seattle, WA

The "wisdom of crowds" dictates that aggregate predictions from a large crowd can be surprisingly accurate, rivaling predictions by experts. Crowds, meanwhile, are highly heterogeneous in their expertise. In this work, we study how the heterogeneous uncertainty of a crowd can be directly elicited and harnessed to produce more efficient aggregations from a crowd, or provide the same efficiency from smaller crowds. We present and evaluate a novel strategy for eliciting sufficient information about an individual's uncertainty: allow individuals to make multiple simultaneous guesses, and reward them based on the accuracy of their closest guess. We show that our *multiple guesses scoring rule* is an incentive-compatible elicitation strategy for aggregations across populations under the reasonable technical assumption that the individuals all hold symmetric log-concave belief distributions that come from the same location-scale family. We first show that our multiple guesses scoring rule is strictly proper for a fixed set of quantiles of any log-concave belief distribution. With properly elicited quantiles in hand, we show that when the belief distributions are also symmetric and all belong to a single location-scale family, we can use interquantile ranges to furnish weights for certainty-weighted crowd aggregation. We evaluate our multiple guesses framework empirically through a series of incentivized guessing experiments on Amazon Mechanical Turk, and find that certainty-weighted crowd aggregations using multiple guesses outperform aggregations using single guesses without certainty weights.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics

Additional Key Words and Phrases: wisdom of crowds; crowdsourcing; scoring rules; uncertainty; k-medians

1. INTRODUCTION

In 1907, Francis Galton famously observed that when 787 people at a country fair were asked to guess the weight of an ox, the median of the guesses (1207 pounds) was impressively close to the true value (1198 pounds) [Galton 1907b]. The phenomena behind this observation has since come to be known as the "wisdom of crowds," whereby aggregations of non-expert estimates are surprisingly accurate, and has been studied extensively in many disparate contexts [Lorge et al. 1958; Surowiecki 2005].

In this work, we explore how one can harness the heterogeneous uncertainty of crowds to improve crowd aggregations by complementing the traditional procedure with an additional request that elicits each individual's certainty, and make use of those certainties to usefully weight crowd aggregations. We contribute and analyze a simple novel mechanism for eliciting uncertainty through a request that is highly compatible with generic crowd estimation procedures: ask individuals to make multiple simultaneous guesses.

This work was supported in part by NSF grant IIS-1258741, ONR PECASE N00014-13-1-0023, and a grant from the TerraSwarm Research Center. Author's addresses: J. Ugander, Microsoft Research, Redmond, WA 98052, jugander@stanford.edu; R. Drapeau and C. Guestrin, University of Washington, Computer Science & Engineering, Seattle, WA 98195 {drapeau,guestrin}@cs.washington.edu.

EC'15, June 15–19, 2015, Portland, OR, USA.
http://dx.doi.org/10.1145/2764468.2764529

Our theoretical contribution is to show that under reasonable assumptions on their beliefs, when individuals are rewarded according to a simple *multiple guesses scoring rule* — scoring them in proportion to the accuracy of their closest guess — they are properly incentivized to spread out their guesses in a manner that usefully reveals the certainty of their beliefs in well-structured ways. More specifically, we show that if a population of individuals hold belief distributions that are symmetric, log-concave, and all belonging to some single location-scale family, then the above scoring rule elicits a specific set of quantiles for all individuals in the population, quantiles that can be used to measure interquantile ranges. We then show that these interquantile ranges can be used to produce weights for both weighted mean and weighted median aggregations that exhibit favorable statistical efficiencies.

The significance of the wisdom of crowds phenomena is in many ways empirical: the basic observation that noisy unbiased measurements can be aggregated to form consistent estimators is not surprising, but do non-expert crowds really produce unbiased samples? Similarly, are people really capable of spreading out their multiple guesses to minimize their loss according to scoring rules? To evaluate the practical significance of our theoretical results, we contribute an experimental evaluation of our multiple guesses scoring rule through a pair of guessing experiments.

We show that not only does our two guesses scoring rule have favorable theoretical properties, it is also empirically performative across a series of guessing games conducted on Amazon Mechanical Turk. Our theoretical analysis predicts that under our assumptions, individuals should place two guesses as if reporting a [25%,75%] confidence interval. We find that responses to our scoring rule are statistically indistinguishable from responses under an interval scoring rule that incentivizes and explicitly asks for [25%,75%] confidence intervals. Furthermore, we find that using inter-guess ranges to weight aggregations significantly reduces estimation error.

Model of crowd beliefs. Consider an unknown quantity μ of interest, and a population of n individuals who are all hold independent uncertain and generally distinct beliefs regarding μ. In practical settings, μ may be the current population of a city, the high temperature tomorrow in a specific place, or the weight of an ox at a county fair. In this work we use a characterization of crowd wisdom built upon the notion of subjective belief distributions [Wallsten et al. 1997; Vul and Pashler 2008], that individuals maintain internal probabilistic representations of their beliefs.

We employ a model of belief distributions in crowds based on noisy signals, where each individual is viewed as possessing a single noisy observation of the unknown quantity μ of interest. Let each individual make one observation x_i from the random variable $S_i = \mu + \epsilon_i$, where μ is constant and the additive noise term ϵ_i is an individualized zero-mean random variable. In general we assume individuals have different distributions for ϵ_i, with personalized variances $\text{Var}[\epsilon_i] = \sigma_i^2$ that are known to them, modeling their relative knowledge about μ. Meanwhile, we let $X_1, ..., X_n$ be individual belief distributions, the beliefs about μ held by each individual based on their single observation from S_i. In the absence of priors or other assumptions, the mean and variance of X_i are then $\mathbb{E}[X_i] = x_i$ and $\text{Var}[X_i] = \sigma_i^2$.

Related work. The study of crowd elicitation and aggregation has been pursued in many directions, including the use of competitive games, scoring using "gold standard" questions, reputations, and the psychology as well as sociology of uncertainty.

Rewarding individuals according to a scoring rule forms a non-competitive game, with the guesses only being evaluated relative to a ground truth, not relative any other guessers. An alternative to using scoring rules to elicit truthful guesses is to create a competitive game between multiple individuals, and structure the game in a manner that incentivizes honest guesses. We do not consider competitive games here,

but this was in fact how the original Galton study was framed [Galton 1907b], a competition that awarded a prize to the individuals with the closest guesses. In the context of a single guess, a competitive game rewarding the "closest guess" is closely related to Hotelling's facility location game [Hotelling 1929] and can have a highly non-trivial strategy space [Mendes and Morrison 2014]. With many players, the optimal symmetric mixed strategy is to draw guesses from the common public prior [Osborne and Pitchik 1986]. Under a mixture of public and private information, as in the "forecasting contest" model [Ottaviani and Sørensen 2006], it is optimal to over-emphasize private information relative public information. It has been shown that aggregating over competitive crowds with a mixture of public and private information can outperform aggregating over non-competitive crowds, at least under certain models of information [Lichtendahl Jr et al. 2013]. It is possible that aggregations from a properly structured "multiple guesses competition" may be more efficient than aggregation from our non-competitive scoring rule, but we do not explore this question.

In settings where the goal is to make inferences about unknown quantities, scoring functions generally rely on the use of "gold standard" questions [Shah and Zhou 2014] with known answers that are interspersed with unknown answers. The participant does not know which of the questions are gold standard questions, but scoring rules are applied only to those questions with known answers. A useful strategy in the complete absence of ground truth knowledge is the so-called "Bayesian Truth Serum" (BTS) scoring rule [Prelec 2004], which asks agents for their estimate and also for their estimate of the sample distribution of estimates from other individuals. By rewarding agents for their knowledge of the estimates of others, the BTS scoring rule is truth-eliciting under mild assumptions, even when the truth is unknown. An analysis of a BTS-like multiple guesses rule, or more generally how incentives change if our scoring rule is evaluated purely relative to other responses of the crowd [Kamar and Horvitz 2012], would be an interesting direction for future research.

In contexts where the same individuals are observed multiple times, a broad range of traditional strategies for discerning latent reputations become admissible [Dekel and Shamir 2009]. A burst of literature has recently examined variations on this theme of identifying "smaller, smarter crowds" in contexts ranging from policy forecasting [Jose et al. 2013; Budescu and Chen 2014] to sports betting [Goldstein et al. 2014; Davis-Stober et al. 2014]. Our context differs in that we have no repeated judgements upon which to judge quality. It would be interesting to study our multiple guesses framework in a repeated judgement setting, to see how effectively one could infer how well individuals judge their own certainty, potentially enabling improved aggregation.

Regarding prior investigations into multiple guessing, recent psychology research has explored the concept of "dialectical crowds within" [Herzog and Hertwig 2009], showing that when people provide multiple *sequential* guesses, the average is a better guess than their first guess. This sequential crowds-within approach does not consider certainty or weighting schemes, and moreover there is active debate about why the approach works [White and Antonakis 2013; Herzog and Hertwig 2013]. We feel our analysis of multiple simultaneous guesses usefully enhances this discussion of "crowds within", providing a theory for how multiple guesses are dispersed under our scoring rule. Regarding the sociology of crowds, we do not consider the potential impacts (negative or positive) of social interference between individuals [Lorenz et al. 2011; Das et al. 2013], or how it can possibly be overcome when it is present.

2. ELICITING UNCERTAINTY

In this section we begin by briefly reviewing the theory of proper scoring rules and how it applies to eliciting a person's uncertainty, the variance of their belief distribution. We then use the language of proper scoring rules to present our novel strategy in terms of

multiple guesses, and prove conditions under which it is proper. Throughout this work we assume all agents are expected utility maximizers; scoring rules for non-expected utility maximizers [Offerman et al. 2009] are outside the scope of this work.

2.1. Strictly proper scoring rules

From the perspective of an individual in the crowd, their beliefs about μ are described by a subjective belief distribution X_i with mean $\mathbb{E}[X_i] = x_i$ and variance $\text{Var}[X_i] = \sigma_i^2$. Our goal in this section is to elicit x_i and σ_i^2 by rewarding the individual using an incentive-compatible scoring rule based on their responses to one or more questions.

Informally, a strictly proper scoring rule for a property of a distribution is a mechanism that correctly incentivizes individuals to truthfully report their beliefs regarding that property, with a uniquely optimal answer. For example, the Brier scoring rule for a response r for a belief distribution X is given by $S_{\text{Brier}}(r; X) = 2rX - r^2$, and is a strictly proper scoring rule for the expectation of a distribution [Brier 1950; Savage 1971]. By convention we view score functions as loss functions, where a lower score is better. The score is a function of a random variable X, and so individuals aim to choose r to minimize the expected value of their score, minimizing their expected loss.

More formally, a scoring rule $S(r; X)$ is said to be *strictly proper* for a response statistic r (here, the expected value) when, for a person holding beliefs about the distribution of X, the expected score of reporting $\mathbb{E}[X]$ under their beliefs is the best possible expected score:

$$\mathbb{E}_X[S(\mathbb{E}[X]; X)] \leq \mathbb{E}_X[S(r; X)], \forall r,$$

with equality if and only if $r = \mathbb{E}[X]$.

Beyond eliciting expectations of distributions, an extension of the Brier rule for higher moments also provides a strictly proper scoring function for jointly eliciting the first k moments of a distribution [Frongillo et al. 2015]:

$$S_{\text{Brier,k}}(r_1, ..., r_k; X) = \sum_{j=1}^{k} 2r_j X^j - r_j^2. \quad (1)$$

This Brier rule for higher moments can be used to elicit a person's joint beliefs about the mean and variance of a distribution by eliciting $\{\mathbb{E}[X], \mathbb{E}[X^2]\}$, which can be transformed to the mean and variance since $\{\mathbb{E}[X], \text{Var}[X]\} = \{\mathbb{E}[X], \mathbb{E}[X^2] - \mathbb{E}[X]^2\}$.

While Brier rules are provably strictly proper, incentive compatible scoring rules are not a sufficient condition for empirically accurate predictions in practice, and may not even be necessary. Individuals have been shown experimentally to be overconfident even under incentive compatible mechanisms [Keren 1991]. Different mechanisms attempting to access the same information can be very differently calibrated empirically. For example, under different incentive compatible framings, individuals may or may not be able to reliably assess higher moments of distributions [Goldstein and Rothschild 2014]. Simply because a mechanism is incentive compatible does not mean that individuals will understand it, and a richer understanding of human computation has a great deal to contribute to empirical mechanism design. We therefore seek a simple scoring rule that makes an intuitive request under transparent incentives.

2.2. The multiple guesses scoring rule for quantiles

Our proposed strategy for eliciting the mean and variance of belief distributions is to solicit "multiple guesses" $r_1, ..., r_k$, and score the guesser based on the absolute deviation between the true value and their closest guess:

$$S_{\text{MG,k}}(\{r_1, ..., r_k\}; X) = \min\{|r_1 - X|, ..., |r_k - X|\}.$$

Both the request (make multiple simultaneous guesses) and scoring (minimize the error of your closest guess) are simple to communicate and understand. Intuitively,

guesses that are close together represent certainty, while guesses that are far apart represent uncertainty. We show that under reasonable assumptions the scoring rule is in fact strictly proper, and the responses reveal sufficient information for performing certainty-weighted crowd aggregation.

More specifically, we show that for log-concave belief distributions, soliciting k guesses and offering rewards based on the absolute deviation of the closest guess is a strictly proper scoring rule for k quantiles of the belief distribution. The quantiles elicited do not have a simple form in general, but we show that they must be the same fixed quantiles across all distributions within the same location-scale family. In the special case of *symmetric* log-concave distributions and $k = 2$, we show that the quantiles do have a simple form: $\{p_1, p_2\} = \{F_X^{-1}(1/4), F_X^{-1}(3/4)\}$, where F_X^{-1} is the quantile function of the belief distribution X.

To compare this rule to other known proper scoring rules for quantiles, it is known that quantiles are elicitable for general distributions (not merely log-concave distributions) under more complex scoring rules [Lambert et al. 2008]. An example of a strictly proper scoring rule for jointly eliciting the quantiles $\{F_X^{-1}(\alpha_1), ..., F_X^{-1}(\alpha_k)\}$ of a general distribution X, with all $\alpha_i \in (0, 1)$, is [Gneiting and Raftery 2007]:

$$S_{\text{quantile,k}}(r_1, ..., r_k; X) = \sum_{j=1}^{k} (X - r_j)\mathbf{1}[X \geq r_j] - \alpha_j r_j. \tag{2}$$

The related *interval scoring rule* is a known strictly proper scoring rule for two symmetric quantiles $\{F_X^{-1}(\frac{\alpha}{2}), F_X^{-1}(1 - \frac{\alpha}{2})\}$, defined as:

$$S_{\text{interval}}(\ell, u; X) = (u - \ell) + 2\alpha^{-1}(\ell - X)\mathbf{1}[X < \ell] \ + 2\alpha^{-1}(X - u)\mathbf{1}[X > u]. \tag{3}$$

This interval scoring rule effectively asks for a confidence interval $[\ell, u]$, and penalizes outcomes that fall outside the interval while concurrently penalizing a wide interval. For the case of $\alpha = 1/2$, the interval score elicits $\{F_X^{-1}(1/4), F_X^{-1}(3/4)\}$, the same quantiles as our two guesses scoring rule for symmetric log-concave distributions. While the two guesses scoring rule is only strictly proper for log-concave distributions, it is intuitively simpler than the interval scoring rule, for which the simplest description is arguably "your score is the width of your interval, and if the observation falls outside your interval, add on four times how far outside it falls." In the experimental portion of this work, we evaluate the relative empirical performance of the two guesses scoring rule and the interval scoring rule when targeted to elicit the same two quantiles.

2.3. The properness of multiple guesses

We now present our main result, the strict properness of our multiple guesses scoring rule for eliciting specific quantiles of any log-concave uncertainty distribution, where log-concavity is a sufficient but not necessary condition. We also present a second proposition regarding additional results for symmetric log-concave distributions.

PROPOSITION 2.1. *For any log-concave distribution X, the k guesses scoring rule*

$$S_{MG,k}(\{r_1, \ldots, r_k\}; X) = \min\{|X - r_1|, \ldots, |X - r_k|\},$$

is strictly proper for some set of points $\{p_1, \ldots, p_k\}$, meaning

$$\mathbb{E}_X[S_{MG,k}(\{p_1, ..., p_k\}; X)] \leq \mathbb{E}_X[S_{MG,k}(\{r_1, ..., r_k\}; X)]$$

for all $\{r_1, \ldots, r_k\}$, with equality if and only if $p_i = r_i, \forall i$. For all k, the ordered quantiles of $\{p_1, \ldots, p_k\}$ are fixed for all distributions within the same location-scale family.

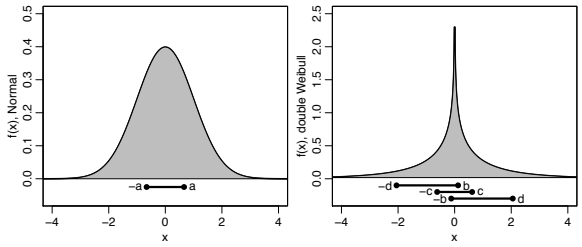

Fig. 1. Left: for the two guesses scoring rule, the unique optimal response set for a zero-mean Normal distribution is $\{-0.674\sigma, 0.674\sigma\}$. Right: as an example of how even symmetric and unimodal non-log-concave distributions may not have a unique optima, for the zero-mean double Weibull distribution [Mease et al. 2004] there are two asymmetric optima $\{-d, b\}$ and $\{-b, d\}$ that both outperform the best symmetric answer $\{-c, c\}$.

PROPOSITION 2.2. *For any symmetric log-concave distribution X the unique quantiles are symmetric, meaning $p_i = F_X^{-1}(q) \Leftrightarrow p_{k-i+1} = F_X^{-1}(1-q)$ for $i = 1, ..., k$, where indices are ordered. For k odd, the median response is the median of X, $p_{(k+1)/2} = F_X^{-1}(1/2)$. For $k = 2$, $\{p_1, p_2\} = \{F_X^{-1}(1/4), F_X^{-1}(3/4)\}$.*

Log-concave distributions include the Normal, Laplace, Logistic, Gamma, and Uniform distributions. Location-scale families include the Normal distribution family, the Laplace distribution family, the Uniform distribution family, and any elliptical distribution family. The proof of the uniqueness in Proposition 2.1 is an application of a known result from the literature on optimal quantizers; establishing that the quantiles are fixed within a location-scale family requires some basic novel analysis. We first provide a discussion that usefully frames the results before giving the proofs.

Notice that the expectation of the multiple guesses score function can be written as:

$$\mathbb{E}_X[\min\{|X - r_1|, ..., |X - r_k|\}] = \int_{-\infty}^{\frac{r_1+r_2}{2}} |x - r_1| f_X(x)dx + ... + \int_{\frac{r_{k-1}+r_k}{2}}^{\infty} |x - r_k| f_X(x)dx,$$

where we've assumed, without loss of generality, that $r_1 \leq \ldots \leq r_k$ are ordered. This objective function is precisely the objective function of the k-median problem for a continuous univariate distribution, the task of determining k ordered points $r_1, ..., r_k$ that minimize the expected absolute deviation under that distribution. Therefore, conditions under which the continuous k-medians problem has a unique optimal set are conditions under which we know that our score function elicits a unique response set. Slight variations on the k-medians problem have been studied across a broad range of literatures: it is a variation on the facility location problem in operations research [Shmoys et al. 1997], the Fermat-Weber problem in geometry [Fekete et al. 2005], and the optimal quantizer problem in signal processing [Dalenius 1950; Fleischer 1964].

For k-medians problems in general — beyond univariate log-concave distributions — there can be multiple globally optimal sets that are equivalently optimal, meaning that a rational individual with beliefs that are not log-concave could be equally justified in returning any one of several globally optimal sets. Figure 1 presents an example of how uniqueness can fail in general even for symmetric unimodal distributions.

There are many known sufficient conditions for uniqueness for the optimal quantizer problem, of which the univariate k-medians problem is a special case, but they are not neatly nested. Results can generally be divided into sufficient conditions on the probability distribution and sufficient conditions on the loss function, where the loss function for k-medians is simply absolute deviation. In 1964, Fleischer gave an analysis of Lloyd's algorithm for the k-means problem (quadratic loss [Lloyd 1982]), and found that for log-concave univariate distributions the optima of the objective function is unique, and Lloyd's algorithm finds that optima [Fleischer 1964]. After Fleischer,

log-concavity was shown to also be a sufficient condition for uniqueness of local optima for any symmetric convex loss function $L(z)$ where $L(z) = 0$ iff $z = 0$ [Trushkin 1982; Kieffer 1983], which includes the absolute deviation loss function used in the k-medians objective function. Recently, even weaker conditions have been given, showing that uniqueness holds for log-concave distributions under loss functions that are slightly more general than convex [Delattre et al. 2004; Cohort 2000]. These conditions are arguably the broadest concise sufficient conditions known in the literature, though it is also known that log-concavity is *not* a necessary condition for uniqueness: other methods have shown uniqueness for e.g. Pareto distributed uncertainty (under convex loss functions), which is not a log-concave distribution [Fort and Pagès 2002].

We now state a special case of the general result of Cohort [Cohort 2000], for which the strict properness of our multiple guesses score function is a corollary.

PROPOSITION 2.3. *For the absolute deviation loss function $L(x) = |x|$ and $f_X(x)$ log-concave, there exists a local minima $r_1, ..., r_k$ for the expected loss*

$$\mathbb{E}_X[\min\{L(X - r_1), ..., L(X - r_k)\}],$$

for any k, and it is unique.

PROOF. See any of [Trushkin 1982; Kieffer 1983; Cohort 2000], with Cohort providing the most general result. □

Using this uniqueness, we now prove the results presented in Proposition 2.1 and Proposition 2.2.

PROOF (OF PROPOSITION 2.1). By Proposition 2.3, we know the optima exists and is unique. For general k, there is no simple expression for a universal set of quantiles, as the quantiles can be distribution-dependent. We now show that for a fixed k, the same quantiles are elicited for all distributions belonging to the same location-scale family. For general k, the stationary conditions of the expected loss $g(\{p_1, ..., p_k\}; X) = \mathbb{E}_X[S_{\text{MG,k}}(\{p_1, ..., p_k\}; X)]$ are:

$$\begin{cases} \frac{\partial g}{\partial p_1} = 2F_X(p_1) - F_X(\frac{p_1+p_2}{2}) = 0 \\ \frac{\partial g}{\partial p_i} = 2F_X(p_i) - F_X(\frac{p_{i-1}+p_i}{2}) - F(\frac{p_i+p_{i+1}}{2}) = 0, \qquad i = 2, \dots, k-1 \\ \frac{\partial g}{\partial p_k} = 2F_X(p_k) - F_X(\frac{p_{k-1}+p_k}{2}) - 1 = 0. \end{cases}$$

By setting $q_i = F_X(p_i)$, $\forall i$, we have the following system of equations:

$$\begin{cases} 2q_1 = F_X(\frac{F_X^{-1}(q_1)+F_X^{-1}(q_2)}{2}) \\ 2q_i = F_X(\frac{F_X^{-1}(q_{i-1})+F_X^{-1}(q_i)}{2}) + F_X(\frac{F_X^{-1}(q_i)+F_X^{-1}(q_{i+1})}{2}), \qquad i = 2, \dots, k-1 \\ 2q_k = F_X(\frac{F_X^{-1}(q_{k-1})+F_X^{-1}(q_k)}{2}) + 1. \end{cases}$$

For every distribution X (with location parameter μ and scale parameter σ) in a location-scale family $\mathcal{F}$, we can transform the cumulative distribution function and its inverse to a standardized distribution $Z \in \mathcal{F}$ (with location 0 and scale 1) using the properties $F_X(p) = F_Z(\frac{p-\mu}{\sigma})$ and $F_X^{-1}(q) = \mu + \sigma F_Z^{-1}(q)$. By substitution, we can reduce the system of equations to only depend on properties of Z, obtaining:

$$\begin{cases} 2q_1 = F_Z(\frac{F_Z^{-1}(q_1)+F_Z^{-1}(q_2)}{2}) \\ 2q_i = F_Z(\frac{F_Z^{-1}(q_{i-1})+F_Z^{-1}(q_i)}{2}) + F_Z(\frac{F_Z^{-1}(q_i)+F_Z^{-1}(q_{i+1})}{2}), \qquad i = 2, \dots, k-1 \\ 2q_k = F_Z(\frac{F_Z^{-1}(q_{k-1})+F_Z^{-1}(q_k)}{2}) + 1. \end{cases}$$

Thus, for all distributions in the same location-scale family, the points $p_1, \ldots, p_k$ must have the same quantiles $q_1, \ldots, q_k$ as the quantiles for the standard distribution Z of the family $\mathcal{F}$. □

PROOF (OF PROPOSITION 2.2). For X symmetric log-concave, existence and uniqueness again follow from Proposition 2.3. The symmetry $p_i = F_X^{-1}(q) \Leftrightarrow p_{k-i+1} = F_X^{-1}(1-q)$ for $i = 1, ..., k$ follows from the uniqueness of the optimal point set: if the points were not symmetric then a reflection of the point set about the median would have the same expected loss, a contradiction of uniqueness. This result in turn gives us that the median $F_X^{-1}(1/2)$ belonging to the optimal set for k odd.

Meanwhile for $k = 2$, the explicit optima $\{p_1, p_2\} = \{F_X^{-1}(1/4), F_X^{-1}(3/4)\}$ can be easily derived, and here we verify by ansatz that $\{F_X^{-1}(1/4), F_X^{-1}(3/4)\}$ satisfy the stationary conditions of the expected loss for all symmetric X:

$$\begin{cases} \frac{\partial g}{\partial p_1} = 2F_X(p_1) - F_X(\frac{p_1+p_2}{2}) = 0 \\ \frac{\partial g}{\partial p_2} = 2F_X(p_2) - F_X(\frac{p_1+p_2}{2}) - 1 = 0. \end{cases} \Rightarrow \begin{cases} 1/2 = F_X(\frac{p_1+p_2}{2}) \\ 1/2 = F_X(\frac{p_1+p_2}{2}) \end{cases}$$
$$\Leftrightarrow F_X^{-1}(1/2) = [F_X^{-1}(1/4) + F_X^{-1}(3/4)]/2.$$

This last condition holds for all symmetric distributions X, while log-concavity continues to act as a sufficient condition for uniqueness. □

Proposition 2.3 establishes that the unique global optima is in fact the *only local optima*. The existence of only a single local optima is a strong statement about computability that relates closely to human reasoning and the capabilities of individuals with so-called "bounded rationality" [Simon 1972]. The uniqueness of the local optima follows not from convexity, as the objective function is not convex, but rather from more subtle uniqueness arguments such as the Mountain Pass Theorem [Courant 1950].

Having established that our multiple guesses scoring rule can be used to elicit specific sets of quantiles, in the next section we will establish that interquantile ranges — simply taking the gap between any two quantiles — can provide sufficient information for constructing weights in weighted wisdom of the crowd aggregation settings.

3. AGGREGATION WITH UNCERTAINTY

In this section, we study certainty-weighted aggregation strategies for aggregating the beliefs and uncertainties of a population with regard to an unknown quantity. In particular, we derive certainty-weighted estimators that outperform their unweighted analogs in terms of their asymptotic relative statistical efficiency in various settings.

Recall that we are considering a population of individuals who each hold a single observation x_i from a differently corrupted signal $S_i = \mu + \epsilon_i$, with the goal of estimating μ. We generally assume all the observations x_i and variances of the noise distributions $\text{Var}[\epsilon_i] = \sigma_i^2$ are known to us, having successfully elicited them from the individual belief distributions using the strategy outlined in the previous section. From the perspective of an aggregator, we thus have n observations x_i from the differently corrupted signals $S_i = \mu + \epsilon_i$, and our goal is to estimate μ.

The two main aggregation strategies we consider are the certainty-weighted mean and the certainty-weighted median of the population. In Francis Galton's original 1907 study of weight guessing, which did not study certainty, Galton strongly advocated aggregation using the sample median, since the sample mean "would give a voting power to cranks in proportion to their crankiness" [Galton 1907a]. In more formal parlance, basic results from robust statistics that would have been known to Galton [de Laplace 1820] dictate that the asymptotic variance of the sample median is $1/(4f_X(0)^2 n)$, while the asymptotic variance of the sample mean is $E[X^2]/n$. As a result, the median is more

efficient than the mean if and only if $4f_X(0)^2 > \mathbb{E}[X^2]^{-1}$. For normal distributions the mean is more efficient, while for distributions with heavy tails, such as populations containing cranks, the median is more efficient.

A key difference between what Laplace was studying and the crowd context we are studying is that our population of observations x_i do not come from a single distribution, but instead from a family of distributions with a shared location parameter. How can we incorporate both the elicited observations and the elicited variances to maximize our efficiency?

To answer this question, we first derive maximum likelihood estimators under known heterogeneous variances for two basic families of noise distributions: Normal distributions and Laplace distributions. We observe that the *variance-weighted mean* is the MLE for the Normal family, while the *standard deviation-weighted median* is the MLE for the Laplace family. We then establish that for families of distributions that belong to the same location-scale family, the variance in the weighting scheme can be replaced by any interquantile range, enabling us to employ the quantiles deduced in the previous section. Lastly, we consider several contamination models in the style of Tukey [Tukey 1960], where some observations have much greater variance than others, and observe that the weighted median is the most efficient of our estimators in contaminated contexts.

3.1. Maximum likelihood estimators: weighted mean and weighted median

We now consider the cases of Normal family uncertainty and Laplace family uncertainty, deriving the optimal weighting schemes for maximum likelihood estimation when the noise distributions come from these families.

We first consider Normal uncertainty, where all the individual signals $S_i = \mu + \epsilon_i$ have noise terms ϵ_i that are $N(0, \sigma_i^2)$. For this case we have the following proposition.

PROPOSITION 3.1. *Given $x_1, \ldots, x_n$ drawn from $S_1, \ldots, S_n$ where $S_i \sim N(\mu, \sigma_i^2)$ independently with σ_i^2 known $\forall i$, the MLE for μ is given by $\hat{\mu}_N = (\sum_{i=1}^n \frac{x_i}{\sigma_i^2})/(\sum_{i=1}^n \frac{1}{\sigma_i^2})$.*

PROOF. The log-likelihood function is simply:

$$\ell(\mu; \{x_i\}_{i=1}^n, \{\sigma_i\}_{i=1}^n) = -\sum_{i=1}^n \ln(\sigma_i) - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma_i^2}, \tag{4}$$

which makes the estimator $\hat{\mu}_N = (\sum_{i=1}^n \frac{x_i}{\sigma_i^2})/(\sum_{j=1}^n \frac{1}{\sigma_j^2})$. □

Intuitively, this estimator discounts estimates from individuals who are uncertain and focuses weight on those who are certain. When the uncertainties are homogeneous ($\sigma_i = \sigma, \forall i$), the above estimator $\hat{\mu}_N$ reduces to a homogeneous estimator $\hat{\mu}_{N,hom} = \frac{1}{n}\sum_{i=1}^n x_i$, which is independent of σ, even when σ is known. As a result, collecting information about individual uncertainties would be a waste of surveying resources if the uncertainties were homogeneous, as it would not figure in the eventual estimator.

What if the uncertainties σ_i were heterogeneous and we incorrectly assumed they were homogeneous? The following proposition shows that if a population has heterogeneous uncertainty, then the heterogeneous estimator dominates the homogeneous estimator, in the sense that it always has strictly lower variance.

PROPOSITION 3.2. *Given $x_1, \ldots, x_n$ drawn from $S_1, \ldots, S_n$ where $S_i \sim N(\mu, \sigma_i^2)$ independently with σ_i^2 known $\forall i$, the variances of the homogeneous and heterogeneous estimators $\hat{\mu}_N$ and $\hat{\mu}_{N,hom}$ are:*

$$\mathit{Var}[\hat{\mu}_{N,hom}] = \tfrac{1}{n^2}\textstyle\sum_{i=1}^n \sigma_i^2, \qquad \mathit{Var}[\hat{\mu}_N] = \frac{1}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}, \tag{5}$$

where $\mathit{Var}[\hat{\mu}_N] \leq \mathit{Var}[\hat{\mu}_{N,hom}]$, with equality iff $\sigma_i = \sigma, \forall i$.

PROOF. For the two estimators, we have

$$\text{Var}[\hat{\mu}_{N,hom}] = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2, \quad \text{Var}[\hat{\mu}_N] = \frac{\sum_{i=1}^{n}\frac{1}{\sigma_i^4}\text{Var}[X_i]}{[\sum_{i=1}^{n}\frac{1}{\sigma_i^2}]^2} = \frac{1}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}.$$

The inequality $\text{Var}[\hat{\mu}_N] \leq \text{Var}[\hat{\mu}_{N,hom}]$ follows from Cauchy-Schwartz, and is strict if and only if the vectors $(\sigma_1^2, \ldots, \sigma_n^2)$ and $(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_n^2})$ are linearly independent. Equality therefore requires that for some constant c, $\sigma_i^2 = c/\sigma_i^2, \forall i$. Since σ_i^2 is non-negative, this is true if and only if $\sigma_i = \sigma, \forall i$. □

As a corollary of this proposition, when a population has heterogeneous Gaussian uncertainty and there are a few disproportionately uncertain individuals in the crowd, the relative efficiency of the weighted vs. unweighted estimator would be enormous, highlighting the value of knowing uncertainties when heterogeneity is rampant.

We next consider Laplace uncertainty, where all the individual signals $S_i = \mu + \epsilon_i$ have noise terms ϵ_i that are Laplace$(0, \sigma_i^2)$.

PROPOSITION 3.3. *Given $x_1, \ldots, x_n$ drawn from $S_1, \ldots, S_n$ where $S_i \sim$ Laplace(μ, σ_i^2) independently with variance σ_i^2 known $\forall i$, the MLE for μ is given by the weighted median $\hat{\mu}_L = \text{argmin}_m \sum_{i=1}^{n}\frac{1}{\sigma_i}|x_i - m|$.*

PROOF. The log-likelihood function is simply:

$$\ell(\mu; \{x_i\}_{i=1}^n, \{\sigma_i\}_{i=1}^n) = -\sum_{i=1}^{n}\ln(\sigma_i) - \sqrt{2}\sum_{i=1}^{n}\frac{1}{\sigma_i}|x_i - \mu|. \tag{6}$$

By definition a median of a set of values $\{x_1, ..., x_n\}$ is a value, not necessarily unique, that minimizes the absolute deviation of the set, $\text{argmin}_m \sum_{i=1}^{n}|x_i - m|$. Analogously, the weighted median of the set $\{x_1, ..., x_n\}$ with corresponding weights $\{w_1, ..., w_n\}$ is a value, not necessarily unique, that minimizes the weighted absolute deviation of the set, $\text{argmin}_m \sum_{i=1}^{n} w_i|x_i - m|$. It is immediately clear that the weighted median maximizes the log-likelihood in (6). □

It is important to highlight that the weights used for the weighted mean of the Normal family are not the same weights as those used in the weighted median of the Laplace family: the former uses inverse variance, while the latter uses inverse standard deviation.

3.2. Weights from interquantile range

We now show how we don't need to know the actual variances σ_i^2 or standard deviations σ_i in order to correctly weight the mean or median: because of cancellations in the estimators, any measure of uncertainty that is uniformly proportional to the variances or standard deviations of a family of distributions can be used to weight the mean or median estimator, respectively, to achieve the same statistical efficiency.

We begin with the following basic observation: for any unscaled uncertainty measures $s_i = c\sigma_i^2$, with c constant $\forall i$, the weights $w_i^s = (1/s_i)/(1/\sum_{k=1}^{n} s_k)$ are exactly equal to the variance weights $w_i^{\sigma^2} = (1/\sigma_i^2)/(1/\sum_{k=1}^{n}\sigma_k^2)$. An analogous statement is clearly also true for uncertainty measures proportional to the standard deviation.

This observation has two practically important consequences for our work. First, if people are uniformly biased in their estimate of their certainty (e.g. they are uniformly overconfident) we see it will not impact the quality of our estimator. Second, it implies that eliciting any response with a known transformation to a quantity that is proportional to the variance (or standard deviation) can serve as an equally accurate weight for weighted mean (or weighted median) aggregation. In particular, we now

show how within any family of distributions that constitute a *location-scale family*, any interquantile range is proportional to the standard deviation of that distribution.

PROPOSITION 3.4. *For any random variable X with $\mathbb{E}[X] = \mu < \infty$ and $Var[X] = \sigma^2 < \infty$ with a distribution belonging to a location-scale family $\mathcal{F}$, any interquantile range $IQR(X; p, q) = F_X^{-1}(p) - F_X^{-1}(q)$, with p and q fixed, is proportional to the standard deviation,*

$$IQR(X; p, q) = c_{\mathcal{F}}(p, q)\sqrt{Var(X)}.$$

The constant $c_{\mathcal{F}}(p, q)$ depends only on the family $\mathcal{F}$ of X but not the specific distribution.

PROOF. Let Z be a random variable with the "standard" distribution of the family $\mathcal{F}$, i.e. location 0 and scale 1. Then for any random variable X with a distribution in $\mathcal{F}$ and $\mathbb{E}[X] = \mu$, $\text{Var}[X] = \sigma^2$, a basic property of location-scale families is that $F_X^{-1}(p) = \mu + \sigma F_Z^{-1}(p), \forall p \in (0, 1)$. We therefore have that:

$$IQR(X; p, q) = \mu + \sigma F_Z^{-1}(p) - \mu - \sigma F_Z^{-1}(q) = (F_Z^{-1}(p) - F_Z^{-1}(q))\sigma,$$

where $c_{\mathcal{F}}(p, q) = (F_Z^{-1}(p) - F_Z^{-1}(q))$ is a constant that depends only on the family and the fixed values of p and q. □

Thus, for populations where all belief distributions come from the same location-scale family, weighting by any IQR is equivalent to weighting by the standard deviation, and weighting by any squared IQR is equivalent to weighting by the variance. Note that this does not apply if individuals have belief distributions that belong to different location-scale families, since the constants can be different. For the Normal distribution family, $c_{\mathcal{N}} = 2\sqrt{2}\text{Erf}^{-1}(1/2) \approx 1.349$ while for the Laplace distribution family $c_{\mathcal{L}} = \ln(2)\sqrt{2} \approx 1.386$.

3.3. Contamination models and robustness

Thus far we have seen that we can elicit quantiles and subsequently use interquantile ranges to perform certainty-weight aggregation. We now investigate the robustness of the two strategies we've presented: weighted-mean and weighted-median aggregation.

In a seminal series of papers, Tukey [Tukey 1960] examined the robustness of estimators to outliers by considering a framework where a set of samples is presumed to come from a certain distribution, but is in fact "contaminated" with samples from another distribution with much higher variance. Typically this *contamination model* framework consists of a mixture of a primary standard Normal distribution, with samples from $N(0, 1)$, and a contaminating distribution, with samples from $N(0, b)$, where b is a parameter controlling the variance of the contamination. The mixture proportion is typically fixed at 80% primary samples / 20% contamination samples, providing a single parameter (the contamination variance b) to guide an examination of how noisy outliers can effect the relative efficiency of different estimators.

In Figure 2 we examine such a contamination model, showing the variance of different estimators based on 50 total samples from the mixture model, averaged over 100,000 iterations. In this first panel it is assumed that the identity of the contaminators is known, i.e. all individuals accurately represent their different certainties. In the second panel the identity of 10% of the contaminators is misattributed, meaning that 10% of the low-certainty weights are attributed to high-certainty individuals, and 10% of the low-certainty individuals are given high-certainty weights. In the third panel, the low-certainty weights are randomly attributed across the population.

From this analysis we observe that the efficiency of our weighted estimators is robust to identifiable contamination, assuming uncertain guessers say they're uncertain. If individuals are mischaracterizing their certainty, however, the efficiency gains are

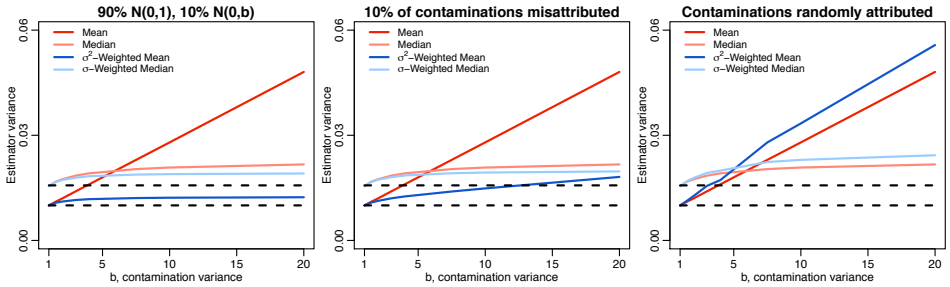

Fig. 2. Estimators variance for contamination models with 100 samples. Left: a contaminated Normal distribution. Center and Right: the same contamination model with partially misattributed weights. Dashed lines indicate variance under uniform certainty.

relatively modest. If the certainty weights are wildly misattributed then the variances for the weighted estimators can be larger than for the naive unweighted estimators.

3.4. From guesses to estimates

We conclude our theoretical contribution with a practical summary of how the various components of our results assemble guesses to form certainty-weighted aggregations. We continue to assume individuals hold symmetric log-concave belief distributions all from the same location-scale family.

For each individual $i = 1, ..., n$ we obtain a set of guesses $\{g_1^i, ..., g_k^i\}$, and our goal is to assemble an estimate of the location, $\hat{\mu}_i$, and of the scale, $\hat{s}_i$, for some $s_i \propto \sigma_i$. For symmetric distributions the location can clearly be obtained as the midpoint of any two symmetric quantiles, while for the scale we've previously established in Proposition 3.4 that the scale can be obtained from any interquantile range. For $k = 2$ guesses this presents a strategy for assembling individual estimates from their guesses, $\hat{\mu}_i = \frac{1}{2}(g_1^i + g_2^i)$ and $\hat{s}_i = |g_1^i - g_2^i|$, allowing us to use the following two estimators to certainty-weight our wisdom of the crowd aggregation:

$$\hat{\mu}_{\text{mean}} = \frac{1}{\sum_{j=1}^n \frac{1}{\hat{s}_j^2}} \sum_{i=1}^n \frac{\hat{\mu}_i}{\hat{s}_i^2}, \qquad \hat{\mu}_{\text{median}} = M_w(\{\hat{\mu}_i\}_{i=1}^n; \{1/\hat{s}_i\}_{i=1}^n).$$

For $k \geq 3$ guesses, individual estimation can become more complicated. When each individual provides three or more guesses, we are presented choices in how we choose to assemble $\hat{\mu}_i$ and $\hat{s}_i$. We would be justified in basing $\hat{s}_i$ on any fixed quantile gap. For $\hat{\mu}_i$ we would be justified in using the overall guess mean, or even the mean of any pair of symmetric quantiles. For odd k we would also be justified in using the median guess $g_{(k+1)/2}$ as the individual's $\hat{\mu}_i$. We briefly consider these individual-level choices from an empirical perspective, as given in the next section.

4. GUESSING EXPERIMENTS

To evaluate the empirical performance of our multiple guesses scoring rule, we implemented an online estimation game for dot guessing we called the *Dot Guessing Game*, similar to previous dot guessing designs [Horton 2010], and recruited 400 participants from Amazon Mechanical Turk to play the game in various configurations. As a general premise, participants were presented a series of images with a large number of dots, see Figure 3, and asked to guess the number of dots in each of the images. Each user started with a tutorial explaining the game, after which they were presented with an initial allotment of points. Once the game began, different scoring rules (the variable condition of the experiment) were used to deduct points based on the user's performance on each image. The tutorial explained the scoring rules to participants, and that they would receive a monetary reward at the end of the game relative to the

amount of points they had remaining. A fixed set of images were generated for each game, with dot counts distributed exponentially over a range of 25 to 250 dots. There was no time limit for how long someone could spend on a single image. At the end of the game, monetary bonuses were awarded to participants that had points remaining.

The task of guessing the number of dots in images was chosen for several reasons. Previous wisdom of the crowd experiments have often studied trivia questions [Lorenz et al. 2011], but because our study was run in an uncontrolled online setting with monetary rewards, it was important that the questions not be easily answered based on internet search results. Secondly, because Amazon Mechanical Turk recruits users from all around the globe, we also wanted to ensure that our questions were not related to any specific cultural context. Dot counting has been previously used in other wisdom of the crowd experiments [Das et al. 2013], and deserves consideration as a useful "model organism" for crowd aggregation research [Horton 2010].

We conducted two experiments using our Dot Guessing Game platform, each deployed to a population of 200 participants, to empirically evaluate the elicitation and estimation strategies presented in this work. In our first experiment, we investigated the performance of the multiple guesses scoring rule by dividing the game into three sections, configured under a 1-, 2-, and 3-guess scoring rule. A new tutorial introduced each section to explain and demonstrate each scoring rule. In our second separate experiment, we used a design involving two sections to investigate the performance of the two guesses scoring rule as compared to the interval scoring rule. Again a new tutorial introduced each section to explain and demonstrate each scoring rule.

Feedback was provided to participants after each image was submitted in both experiments. The feedback displayed the error, the correct answer, and the participant's current score, allowing him or her to track their progress throughout the game. The feedback showed the calculations that went into the deduction to ensure the user knew how they were losing points. The order of the images and the order of the sections were randomly assigned for each user.

Our two experiments were designed to address specific questions. For the multiple guesses scoring rules, did the additional information elicited by the additional guesses result in more accurate estimates of the underlying quantity in practice? When eliciting 3 guesses, how often are the guesses symmetrically positioned about their median? Do responses under the two guesses scoring rule bear any semblance to a [25%, 75%] confidence interval, in terms of calibration? How do the two guesses compare to the responses given under the interval scoring rule? We now evaluate these questions, discussing the two experiments in detail.

Multiple Guesses Experiment. In this experiment, the game was split into 3 sections, presented in a random order. In each section, participants were given 1, 2, or 3 guesses for each set of 12 images, where the first 2 images were tutorial images. Users were not allowed to input the same guess more than once per image (e.g., they could not submit "100" and "100" as their two guesses for an image). After each image, the user was penalized based on the distance of their closest guess to the actual number of dots present, as demonstrated during a tutorial and explained through continuous feedback between images.

The task was designed to take less than 10 minutes, and users were paid a base rate of $0.50 for participating, plus a bonus based on their score. Each participant started the game with 500 points, which were converted to bonus payments at the end of the game at a conversion rate of $0.01 per 10 points. Of 200 participants, 171 finished with points, receiving an average bonus of $0.20. Our main analysis and statistical tests are based on the performance of these 171 participants who finished with points, ensuring that they were properly incentivized by the scoring rules throughout the game (having not run out of points).

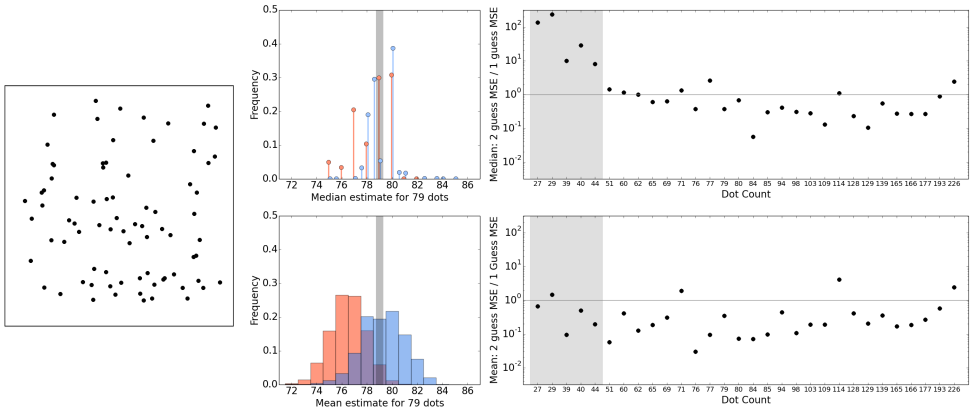

Fig. 3. Experimental results for 1- and 2-guess responses. Left: an image from the game, with 79 dots. Center: bootstrapped distributions of mean and median estimators for the 1-guess (red) and weighted 2-guess (blue) responses for the 79 dot image. Right: ratios of mean squared error (MSE) for bootstrapped population of median and mean estimators. The shaded region indicates dot counts where an "integer constraint" had a pronounced effect, see text. Ratios below 1.0 indicate the weighted 2-guess MSE is lower than the unweighted 1-guess MSE.

In Figure 3 we see the main results of this experiment. The left panel shows a sample image with 79 dots. The central panels show a distribution of estimators based on a sample size of 47 participants making guesses about this image with 79 dots, bootstrapped with replacement from the participants who saw this image instance as a "1 guess" question vs. the participants who saw this image instance as a "2 guess" question. The fixed sample size for each estimator, 47, was set by the size of the smallest population across all (image, scoring rule) exposure pairs. The bootstrapped distribution of 2-guess median estimates has less variance than the 1-guess median estimates, while for mean estimates the 2-guess distribution is significantly less biased but exhibits roughly equivalent variance.

In the right panels of Figure 3, we show the relative mean squared error (MSE) for bootstrapped populations of 10,000 estimates from each estimator, for each image. The game was run with 30 different images, and overall, we see that weighted medians based on 2 guesses generally performs significantly better than unweighted medians based on 1 guess responses, with relative MSEs well below 1.0.

We also observe a clear trend, where the 2-guess median MSE was very high (relative to the 1-guess median MSE) when there were very few dots. We note that this is an artifact of the request being easy, giving participants high certainty, and the answers being required to be integers; for example, if a participant is very certain that there are 27 dots in the image, they are forced to choose between answering $\{26, 27\}$ or $\{27, 28\}$. Our estimators interpret such beliefs as centered at 26.5 or 27.5, yielding few or no belief distributions that are centered at 27. No 1-guess median can be anything other than an integer, which makes a comparison of the median MSEs somewhat unfair in high-certainty scenarios. To avoid this artifact, we focus our statistical tests on images with more than 50 dots, where there was a notable jump in general uncertainty among users.

The reductions in MSE are significant: the one-sided p-value for a pairwise ratio t-test of the 2-guess median MSE vs. 1-guess median MSE is $p = 0.0003$, while for the mean estimators the p-value is $p < 0.0001$. For completeness, when all 200 participants are included, instead of just the 171 who finished with points, the median and mean estimator p-values are $p = 0.012$ and $p = 0.129$, respectively. This suggests that our results are not robust to participants who "aren't playing the game," and complemen-

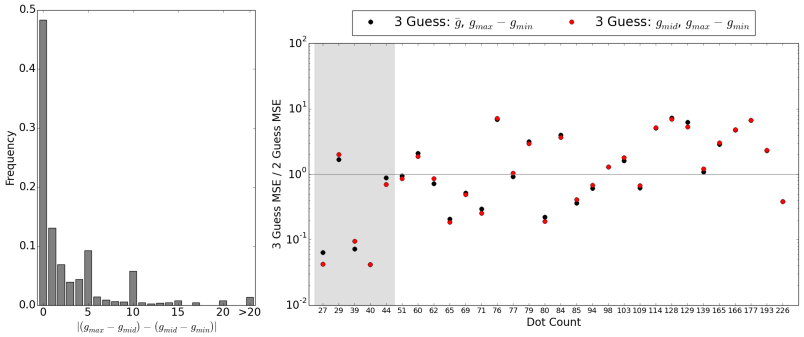

Fig. 4. Experimental results for 3-guess responses. Left: symmetry of gaps in 3-guess triplets. Right: ratios of the mean squared error (MSE) for bootstrapped population of differently configured 3-guess weighted median estimators, compared to a 2-guess MSE. The 3-guess MSE does not vary much between different estimator configurations, and is not clearly better or worse than the 2-guess MSE. The shaded region indicates dot counts where an "integer constraint" effect is suspected, see text.

tary approaches to ensuring worker quality are still advisable in highly heterogeneous crowds such as the Mechanical Turk general population. A possible estimation-based approach, not explored, would be to Winsorize the weights [Hastings et al. 1947].

Next, Figure 4 examines the behavior of participants when they were asked to make 3 guesses. In the left panel we study the symmetry of the guesses: by ordering the three guesses as $(g_{min}, g_{mid}, g_{max})$, we compare the difference between $g_{max} - g_{mid}$ and $g_{mid} - g_{min}$, finding that 48.2% of participants positioned their three guesses symmetrically about their middle guess, with few triplets distributed with large asymmetries. In the right panel, we evaluate the ratio of MSEs for estimators bootstrapped from the participants who were asked for three guesses compared to guesses from those asked for two guesses. We consider both g_{mid} and $\bar{g} = \frac{1}{3}(g_{min} + g_{mid} + g_{min})$ as individual location estimates, while for weights we use $g_{max} - g_{min}$. Estimators weighted by sample standard deviation $\sqrt{\sum_i (g_i - \bar{g})^2}$, not shown, were nearly indistinguishable from estimators weighted by $g_{max} - g_{min}$. This indistinguishability is to be expected since for three symmetric guesses the standard deviation is proportional to the max-min gap. Results from this panel are inconclusive; there is no clear benefit or drawback to 3-guess over 2-guesses, except possibly for low dot counts where the request for 3 guesses appears to avoid the integer constraint effect seen in 2-guess requests. A more detailed investigation of higher guess-count elicitation is left as future work.

Interval Comparison Experiment. In our second experiment, the game was configured with two sections, presented in a random order, where one section asked participants for 2 guesses for every image, analogous to the 2 guesses section from the previous experiment, and the other section asked for a [25%,75%] confidence interval regarding the number of dots believed to be in each image. Participants were not allowed to provide a confidence interval with a width of 0. In the interval section, points were deducted according to the interval scoring rule discussed in Section 2.2. The magnitude of the 2-guess scoring rule was scaled up to deduct 5 times the error (to the closest guess) in points, providing approximately the same expected point penalty per image as the interval rule (in theory, under optimal performance with Gaussian belief distributions).

The task was again designed to take less than 10 minutes, and users were paid a base rate of $0.50 for participating, plus a bonus based on their score. Each participant started the game with 2000 points, which were converted to bonus payments at the end of the game at a conversion rate of $0.01 per 50 points. Of 200 participants, 149 finished with points, receiving an average bonus of $0.14. Our main analysis and statistical tests are again based on the performance of these 149 participants who fin-

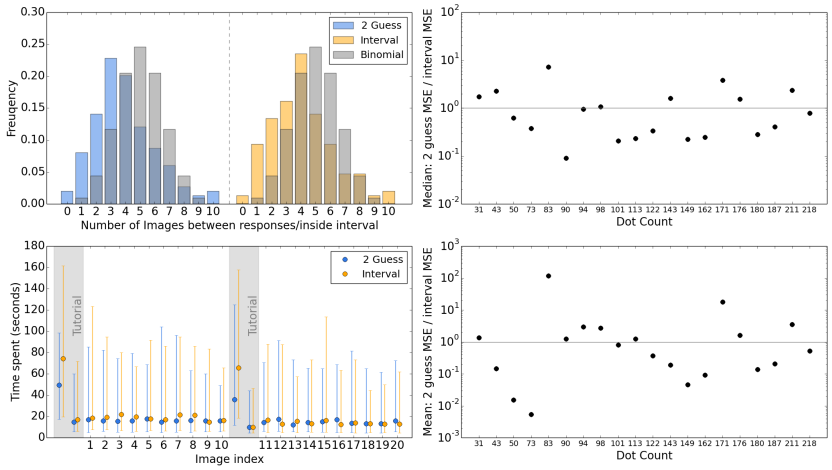

Fig. 5. Results for interval vs. 2-guess comparison. Top left: frequency with which the true dot count was inside/between intervals/guesses. Bottom left: time taken to play the game. Half the users played the interval game, then the 2 guesses game; the other half played the 2 guesses game, then the interval game. Circles indicate medians and error bars are empirical 5th/95th percentiles of the population. Right: ratios of MSE for bootstrapped populations of 2-guess and interval weighted median (top) and weighted mean (bottom) estimators. Ratios below 1.0 indicate the 2-guess-weighted MSE is lower than the interval-weighted MSE.

ished with points, ensuring that they were properly incentivized by the scoring rules throughout the game.

The results of the interval comparison game are presented in Figure 5. In the top left, we see that the two scoring rules were calibrated equally (poorly). If intervals/guesses were properly calibrated, these distributions would both be Binomial distributions with $n = 10$, $p = 0.5$ (here p is the width of the requested interval). The estimated probabilities of being inside/between the interval/guesses were $p_{int} = 0.395$ and $p_{2g} = 0.377$, with a test for a difference between these two parameters is not statistically significant (two-sided p-value 0.5422). We conclude that participants were spreading out their guesses very similar to a [25%,75%] confidence interval, as predicted theoretically for symmetric log-concave belief distributions, but still exhibited the typical overconfidence common in confidence interval elicitation [Keren 1991].

Another question we were interested in was whether participants took significantly longer time to respond to the interval scoring rule than the 2-guesses scoring rule. In the bottom left panel of Figure 5, we see the time spent per image, including the tutorial images. We see that the interval scoring rule tutorial instructions took users slightly longer to process, but overall response times were equivalent.

Lastly, in the right panel we see the ratio of mean squared errors for a population of estimators bootstrapped from the 2-guess responses and from the interval responses. We see that the 2-guess scoring rule appears to generate lower error weighted median estimators than the interval scoring rule, but the differences are not significant under a one-sided paired ratio t-test ($p = 0.1011$ for median, $p = 0.141$ for mean). The "integer constraint" for low dot counts (discussed earlier) impacts both "2-response" median estimators equally, so all dot counts are considered.

In summary, we observe that the multiple guesses scoring rule is an empirically performative approach to eliciting uncertainty for weighted wisdom of the crowd aggregations, as performative as the interval scoring rule. We observe that individuals responded to our guessing game as if their belief distributions were largely symmetric, and the 2-guesses scoring rule elicited responses very similar to the interval scoring

rule configured for [25%,75%] percentiles, as predicted by our theoretical analysis under symmetric log-concave belief distributions.

5. CONCLUSIONS

It is well known that aggregate predictions from large crowds can rival predictions by experts, but crowds are generally heterogeneous in their expertise for any given task at hand. In this paper we have developed and evaluated a new theory of how this heterogeneous "uncertainty of the crowd" can be elicited and utilized to assemble more efficient predictions. Our investigation involved rewarding users according to the *multiple guesses scoring rule* that scores users in proportion to the accuracy of their closest of several simultaneous guesses. We have shown that by simply weighting individuals based on properties of their multiple guesses, estimation error can be significantly reduced for both mean and median aggregation. By providing both theoretical justifications and empirical evaluations, we contribute a novel technique for harnessing heterogeneous crowd certainty in real world crowd estimation tasks. Lastly, our multiple-guesses elicitation and weighting strategy suggests compelling extensions to existing crowd inference techniques, and we leave research into extensions such as learning latent reputations for confidence reporting [Dekel and Shamir 2009], using multiple guesses in a competitive game [Lichtendahl Jr et al. 2013], or modifying the Bayesian Truth Serum mechanism [Prelec 2004] all as future work.

ACKNOWLEDGMENTS

We thank Gilles Pagès for providing a scan of Pierre Cohort's thesis, and Daniel Gorrie, Eric Horvitz, Jon Kleinberg, Robert Kleinberg, Martin Larsson, and Sean Taylor for discussions.

REFERENCES

Glenn Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev* 78, 1 (1950), 1–3.

David Budescu and Eva Chen. 2014. Identifying Expertise to Extract the Wisdom of Crowds. *Management Science* (2014).

Pierre Cohort. 2000. *Sur quelques problemes de quantification*. Ph.D. Dissertation. Univ. Paris 6.

Richard Courant. 1950. *Dirichlet's principle, conformal mapping, and minimal surfaces*. Vol. 3. Springer.

Tore Dalenius. 1950. The Problem of Optimum Stratification. *Scand Actuarial J* 1950, 3-4 (1950), 203–213.

Abhimanyu Das, Sreenivas Gollapudi, Rina Panigrahy, and Mahyar Salek. 2013. Debiasing social wisdom. In *KDD*. ACM, 500–508.

Clintin P Davis-Stober, David V Budescu, Jason Dana, and Stephen B Broomell. 2014. When is a crowd wise? *Decision* 1, 2 (2014), 79.

Pierre Simon de Laplace. 1820. *Théorie analytique des probabilités*. Courcier.

Ofer Dekel and Ohad Shamir. 2009. Vox populi: Collecting high-quality labels from a crowd. In *COLT*.

Sylvain Delattre, Siegfried Graf, Harald Luschgy, Gilles Pages, and others. 2004. Quantization of probability distributions under norm-based distortion measures. *Statistics and Decisions* 22 (2004), 261–282.

Sándor P Fekete, Joseph SB Mitchell, and Karin Beurer. 2005. On the continuous Fermat-Weber problem. *Operations Research* 53, 1 (2005), 61–76.

P Fleischer. 1964. Sufficient conditions for achieving minimum distortion in a quantizer. *IEEE Int. Conv. Rec* 12 (1964), 104–111.

Jean-Claude Fort and Gilles Pagès. 2002. Asymptotics of optimal quantizers for some scalar distributions. *J. Comput. Appl. Math.* 146, 2 (2002), 253–275.

Rafael M Frongillo, Yiling Chen, and Ian A Kash. 2015. Elicitation for Aggregation. In *AAAI*.

Francis Galton. 1907a. One vote, one value. *Nature* 75 (1907), 414.

Francis Galton. 1907b. Vox populi. *Nature* 75 (1907), 450.

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

Daniel Goldstein, R Preston McAfee, and Siddharth Suri. 2014. The Wisdom of Smaller, Smarter Crowds. In *EC*. ACM.

Daniel G Goldstein and David Rothschild. 2014. Lay understanding of probability distributions. *Judgment and Decision Making* 9, 1 (2014), 1–14.

Cecil Hastings, Frederick Mosteller, John W Tukey, and Charles P Winsor. 1947. Low moments for small samples: a comparative study of order statistics. *Annals of Mathematical Statistics* (1947), 413–426.

Stefan M Herzog and Ralph Hertwig. 2009. The wisdom of many in one mind improving individual judgments with dialectical bootstrapping. *Psychological Science* 20, 2 (2009), 231–237.

Stefan M Herzog and Ralph Hertwig. 2013. The Crowd Within and the Benefits of Dialectical Bootstrapping A Reply to White and Antonakis (2013). *Psychological Science* 24, 1 (2013), 117–119.

John J Horton. 2010. The Dot-Guessing Game: A 'Fruit Fly' for Human Computation Research. *SSRN 1600372* (2010).

Harold Hotelling. 1929. Stability in Competition. *The Economic Journal* 39, 153 (1929), 41–57.

Victor Richmond R Jose, Yael Grushka-Cockayne, and Kenneth C Lichtendahl Jr. 2013. Trimmed opinion pools and the crowd's calibration problem. *Management Science* 60, 2 (2013), 463–475.

Ece Kamar and Eric Horvitz. 2012. *Incentives and truthful reporting in consensus-centric crowdsourcing*. Technical Report. MSR-TR-2012-16, Microsoft Research.

Gideon Keren. 1991. Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica* 77, 3 (1991), 217–273.

John Kieffer. 1983. Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function. *IEEE Transactions on Information Theory* 29, 1 (1983), 42–47.

Nicolas S Lambert, David M Pennock, and Yoav Shoham. 2008. Eliciting properties of probability distributions. In *EC*. ACM, 129–138.

Kenneth C Lichtendahl Jr, Yael Grushka-Cockayne, and Phillip E Pfeifer. 2013. The Wisdom of Competitive Crowds. *Operations Research* 61, 6 (2013), 1383–1398.

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Trans on Inf Theory* 28, 2 (1982), 129–137.

Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *PNAS* 108, 22 (2011), 9020–9025.

Irving Lorge, David Fox, Joel Davitz, and Marlin Brenner. 1958. A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. *Psychological bulletin* 55, 6 (1958), 337.

David Mease, Vijayan N Nair, and Agus Sudjianto. 2004. Selective assembly in manufacturing: statistical issues and optimal binning strategies. *Technometrics* 46, 2 (2004), 165–175.

Anthony Mendes and Kent E Morrison. 2014. Guessing games. *AMM* 121, 1 (2014), 33–44.

Theo Offerman, Joep Sonnemans, Gijs Van de Kuilen, and Peter Wakker. 2009. A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *The Review of Economic Studies* 76, 4 (2009), 1461–1489.

Martin J Osborne and Carolyn Pitchik. 1986. The nature of equilibrium in a location model. *International Economic Review* 27, 1 (1986), 223–37.

Marco Ottaviani and Peter Norman Sørensen. 2006. The strategy of professional forecasting. *Journal of Financial Economics* 81, 2 (2006), 441–466.

Dražen Prelec. 2004. A Bayesian truth serum for subjective data. *Science* 306, 5695 (2004), 462–466.

Leonard J Savage. 1971. Elicitation of personal probabilities and expectations. *JASA* 66 (1971), 783–801.

Nihar B Shah and Dengyong Zhou. 2014. Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing. *arXiv preprint arXiv:1408.1387* (2014).

David B Shmoys, Éva Tardos, and Karen Aardal. 1997. Approximation algorithms for facility location problems. In *STOC*. ACM, 265–274.

Herbert A Simon. 1972. Theories of bounded rationality. *Decision and organization* 1 (1972), 161–176.

James Surowiecki. 2005. *The wisdom of crowds*. Random House LLC.

A Trushkin. 1982. Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions. *IEEE Trans. on Information Theory* 28, 2 (1982), 187–198.

John W Tukey. 1960. A survey of sampling from contaminated distributions. *Contributions to probability and statistics* 39 (1960), 448–485.

Edward Vul and Harold Pashler. 2008. Measuring the crowd within probabilistic representations within individuals. *Psychological Science* 19, 7 (2008), 645–647.

Thomas S Wallsten, David Budescu, Ido Erev, and Adele Diederich. 1997. Evaluating and combining subjective probability estimates. *J Behavioral Decision Making* 10, 3 (1997), 243–268.

Chris M White and John Antonakis. 2013. Quantifying Accuracy Improvement in Sets of Pooled Judgments Does Dialectical Bootstrapping Work? *Psychological science* 24, 1 (2013), 115–116.