Approval Voting and Incentives in Crowdsourcing

Nihar B. Shah NIHAR@EECS.BERKELEY.EDU
University of California, Berkeley, CA 94720

Dengyong Zhou DENZHO@MICROSOFT.COM
Microsoft Research, Redmond, WA 98052

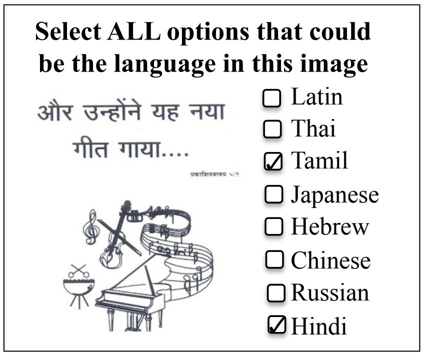
Yuval Peres PERES@MICROSOFT.COM
Microsoft Research, Redmond, WA 98052

Abstract

The growing need for labeled training data has made crowdsourcing an important part of machine learning. The quality of crowdsourced labels is, however, adversely affected by three factors: (1) the workers are not experts; (2) the incentives of the workers are not aligned with those of the requesters; and (3) the interface does not allow workers to convey their knowledge accurately, by forcing them to make a single choice among a set of options. In this paper, we address these issues by introducing approval voting to utilize the expertise of workers who have partial knowledge of the true answer, and coupling it with a ("strictly proper") incentive-compatible compensation mechanism. We show rigorous theoretical guarantees of optimality of our mechanism together with a simple axiomatic characterization. We also conduct preliminary empirical studies on Amazon Mechanical Turk which validate our approach.

1. Introduction

In the big data era, with the ever increasing complexity of machine learning models such as deep learning, the demand for large amounts of labeled data is growing at an unprecedented scale. A primary means of label collection is crowdsourcing, through commercial web services like Amazon Mechanical Turk where crowdsourcing workers or annotators perform tasks in exchange for monetary payments. Unfortunately, the data obtained via crowdsourcing is typically highly erroneous (Kazai et al., 2011; Vuurens et al., 2011; Wais et al., 2010) due to the lack of expertise of workers, lack of appropriate incentives, and often the lack of an appropriate interface for the workers to express their knowledge. Several statistical aggregation methods (Dawid & Skene, 1979; Whitehill et al., 2009; Raykar et al., 2010; Karger et al., 2011; Liu et al., 2012; Zhou et al., 2012; Shah et al., 2015) have been proposed in the literature for improving the quality of the data. Our approach complements these techniques in that we endeavor to obtain higher-quality labels directly via novel interface and incentive mechanisms while not increasing the labeling cost.



Select ALL options that could be the language in this image

और उन्होंने यह नया गीत गाया....

- □ Latin
- □ Thai
- ☑ Tamil
- □ Japanese
- □ Hebrew
- □ Chinese
- □ Russian
- ☑ Hindi

Figure 1. An illustration of a task with an approval-voting interface, asking the worker to select all the options that she believes may be correct.

The typical crowdsourcing labeling task consists of a set of questions such as images to be labeled, and each question is associated with a set of options. Each option is the name of a category and the true label for any question is one of these options. In principle, for each question, the worker is required to select the option that she believes is most likely to be correct. More formally, it involves eliciting the *mode* of the worker's belief. Such a "single-selection" crowdsourcing setting has been studied extensively, both

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

empirically and theoretically.

In this paper, we consider an alternative "approval-voting" means of eliciting labels from the workers, wherein the worker is allowed to select multiple options for every question.[1] See Figure 1 for an example. Approval voting is known to have many advantages over single-selection systems in psychology and social choice theory (Horst, 1932; Coombs, 1953; Coombs et al., 1956; Collet, 1971; Brams & Fishburn, 1978; Gibbons et al., 1979): it provides workers more flexibility to express their beliefs, and utilizes the expertise of workers with partial knowledge more effectively. For instance, Coombs (1953) posits that "It seems to be a common experience of individuals taking objective tests to feel confident about eliminating some of the wrong alternatives and then guess from among the remaining ones" and that "Individuals taking the test should be instructed to cross out all the alternatives which they consider wrong." Under this approval-voting interface, we will require a worker to select every option which she believes could possibly be correct. Mathematically, we formulate this problem as eliciting the *support* of the beliefs of workers for each question. In the setting of crowdsourcing, as compared to single-selection, selecting multiple options would allow for obtaining more information about the partial knowledge of these non-expert workers. This additional information is particularly valuable for difficult labeling questions, allowing for the identification of the sources of difficulty. Indeed, Coombs et al. (1956) conclude that under such a questionnaire, "clear evidence for the existence of partial information mediating responses to multiple choice items was obtained."

Let us illustrate the utility of approval voting using the example of Figure 1. Assume that there are two workers. The first worker recognizes that the language spoken in India, but is confused between "Tamil" and "Hindi". The second worker is confused about some other aspect of the displayed text, and thinks it is either "Hindi" or "Thai". If every worker is allowed to select only a single answer, it may turn out that the first worker selects "Tamil", while the second worker correctly selects "Hindi". Their responses will thus not provide any definitive answer about the true label. In contrast, if we fully elicit their knowledge by letting them select multiple options, that is, ("Tamil", "Hindi") from the first worker and ("Hindi", "Thai") from the second, then one can infer "Hindi" to be the correct answer. Indeed, "Hindi" is the language in Figure 1.

Albeit its great flexibility in eliciting partial knowledge, approval voting alone is not sufficient for high quality crowdsourcing. A worker may have no incentives to truthfully disclose her partial knowledge on the crowdsourcing question. For instance, the worker may simply choose all provided options as her answer and get paid. To address this problem, we need to couple approval voting with an appropriate "incentive-compatible" payment mechanism such that a worker receives her maximum expected payment if and only if she truthfully discloses her partial knowledge (that is, the support of her belief) on the crowdsourcing question. In other words, the payment mechanism has to be a "strictly proper scoring rule". Moreover, we want the mechanism to be "frugal", paying as less as possible to a worker who simply selects all provided options as her answer. The problem setting for incentive mechanism design is formally described in Section 3.

Our first result is negative, proving that unfortunately no mechanism can be incentive compatible for this setting (Section 4). This impossibility result leads us to introduce a "coarse belief" assumption (Section 5) that relies on a certain granularity in people's beliefs. We propose a payment mechanism that is incentive-compatible and frugal (Section 6), and show that it is the only mechanism which satisfies these two requirements. We then generalize the analysis of our mechanism to settings where the coarse belief assumption may not be satisfied, and show that our mechanism simply incentivizes workers to select options for which their belief is relatively high enough (Section 7). This perspective also leads to a simple axiomatic characterization of our mechanism. The paper concludes with a report on preliminary experiments verifying certain basic hypotheses underlying our approach (Section 8) and a discussion on future work (Section 9).

2. Related Literature

Approval voting (Ottewell, 1977; Kellett & Mott, 1977; Weber, 1977; Brams & Fishburn, 1978) is a form of voting in which each voter can "approve of" (that is, select) multiple candidates. No further preferences among these candidates is specified by the voter. Our proposed interface for crowdsourcing elicits approvals on the candidate options for each question. Closer to our setting of crowdsourcing, approval voting has been studied in the context of question and answer forums (Jain et al., 2009) and Doodle polls (Zou et al., 2014). The focus of the present paper is on the design of incentive mechanisms with properties that fundamentally hold irrespective of the nature of the setting.

The framework of scoring rules (Brier, 1950; Savage, 1971; Gneiting & Raftery, 2007; Lambert & Shoham, 2009) considers the design of payment mechanisms to elicit predictions about an event whose actual outcome will be observed in the future. The payment is a function of the agent's response and the outcome of the event. The payment is called "strictly proper" if its expectation, with respect to the belief of the agent about the event, is strictly maximized when

[1] The literature on psychology often refers to approval voting as "subset selection".

the agent reports her true belief. Proper scoring rules however provide a very broad class of mechanisms, and do not specify any specific mechanism for use. The mechanism proposed in the present paper may alternatively be viewed as the "optimal" proper scoring rules for eliciting supports of workers' beliefs across multiple questions.

Shah & Zhou (2014) consider a crowdsourcing setup with the traditional single-selection setting, also eliciting the workers' confidences for each response. They propose a mechanism to suitably incentivize workers and show that their proposed mechanism is shown to be the only one satisfying a proposed "no-free-lunch" axiom. While the setting of our work is different from that of Shah & Zhou (2014), interestingly, our mechanism that was derived for a different interface and under a different set of assumptions, turns out to be the only mechanism that can satisfy the no-free-lunch axiom (adapted to our setting).

The mechanisms presented subsequently in the present paper assume the presence of some "gold standard" questions whose answers are known apriori to the system designer. There is a parallel line of literature (Prelec, 2004; Miller et al., 2005; Faltings et al., 2014; Miller et al., 2005; Dasgupta & Ghosh, 2013) that explores the design of mechanisms that operate in the absence of any gold standard questions. These works typically elicit additional information from the workers, such as asking them to predict the responses of other workers. The mechanisms designed therein can generally provide only weaker guarantees due to the absence of a gold standard answer to compare with.

3. Problem Setup

Consider $N \geq 1$ questions, each of which has $B \geq 2$ options to choose from. For each option, exactly one of the B options is correct. We assume that these N questions contain G $(1 \leq G \leq N)$ "gold standard" questions, that is, questions to which the mechanism designer knows the answers apriori. These gold standard questions are assumed to be mixed uniformly at random among the N questions, and the worker is evaluated based on her performance on these G questions. For every individual question, we assume that the worker has, in her mind, a distribution over the B options representing her beliefs of the probabilities of the respective options being correct. We assume that these belief-distributions of a worker are independent across questions (Gibbons et al., 1979). For any integer K, we will use $[K]$ as a shorthand for the set $\{1, \ldots, K\}$.

Our goal is to elicit, for every question, the *support* of the worker's distribution over the B options. In other words, we wish to incentivize the worker such that for each question, the worker should select the *smallest subset of the set of options* such that the correct answer according to her belief lies in the selected subset. Formally, suppose that for any question $i \in [N]$, the worker believes that the probability of option $b \in [B]$ being correct is p_{ib}, for some non-negative values $p_{i1}, \ldots, p_{iB}$ that sum to one. Then the goal is to incentivize the worker to, for each question $i \in [N]$, select precisely the set of options

$$\{b \in [B] \mid p_{ib} \neq 0\}. \tag{1}$$

Payment function. As mentioned earlier, the worker's performance is evaluated based on her responses to the gold standard questions. For any question in the gold standard, we denote the evaluation of the worker's performance on this question by a value in the set $\{-(B - 1), \ldots, -1, 1, \ldots, B\}$: the magnitude of this value represents the number of options she had selected and the sign is positive if the correct answer was in that subset and negative otherwise. For instance, if the worker selected four options for a certain gold standard question but none of them was correct, then the evaluation of this response is denoted as "-4"; if the worker selects two options for a gold standard question and one of them turns out to be the correct option then the evaluation of this response is denoted as "$+2$". Let

$$f : \{-(B-1), \ldots, -1, 1, \ldots, B\}^G \to \mathbb{R}_+$$

denote the payment function. It is this function f which must be designed in order to incentivize the worker. Note that we restrict the payment f to be non-negative to adhere to present crowdsourcing setups which deal solely in monetary compensations. Finally, we let $\alpha > 0$ denote the value of perfect data: a worker who answers everything perfectly should be paid an amount α, that is,

$$f(1, \ldots, 1) = \alpha. \tag{2}$$

Throughout the paper, we will consider only those mechanisms that satisfy (2).

Expected payment. A quantity central to our analysis is the *expected payment*, where the expectation is from the point of view of the worker, and is taken over the randomness in the choice of the G gold standard questions among the N questions and over the N probability distributions representing her beliefs for the N questions. Let us formalize this notion. Suppose that for question $i \in [N]$, the worker has selected some $y_i \in [B]$ of the B options. Further, let $s_i \in [0, 1]$ denote the probability, under the worker's beliefs, that the correct answer to question i lies in this set of y_i selected options. In other words, s_i denotes the sum of the beliefs for the y_i options selected by the worker (consequently, the sum of the beliefs for the options not selected is $(1 - s_i)$). Then from the worker's point of view, her expected payment for this selection is

$$\frac{1}{\binom{N}{G}} \sum_{\substack{(j_1,\ldots,j_G)\\ \subseteq[N]}} \sum_{\substack{(\epsilon_1,\ldots,\epsilon_G)\\ \in\{-1,1\}^G}} \Big(\prod_{i=1}^{G} (1-s_{j_i})^{\mathbf{1}\{\epsilon_i=-1\}} s_{j_i}^{\mathbf{1}\{\epsilon_i=1\}} f(\epsilon_1 y_{j_1},\ldots,\epsilon_G y_{j_G})\Big). \qquad (3)$$

The outer summation in (3) corresponds to the expectation with respect to the random distribution of the G gold standard questions in the N total questions, and the inner summation corresponds to the expectation with respect to the worker's beliefs of her choices being correct. In this paper, we assume that the workers aim to maximize their expected rewards; extending our theory to more general utility functions is straightforward.

Definition 1 (Incentive compatibility). *A mechanism is incentive compatible if the expected payment (Equation* (3)*), from the worker's point of view, is strictly maximized when she selects precisely the support (Equation* (1)*) of her belief for each question.*

Observe that a worker who selects all the options for all the questions doesn't give any useful information. In order to deter such behavior, one would like to ensure that in addition to paying a (large enough) amount α to a good worker, the mechanism should expend as small an amount as possible on such a worker. This leads to a notion of "frugality".

Definition 2 (Frugality). *An incentive-compatible mechanism f is frugal if*

$$f(B,\ldots,B) \leq f'(B,\ldots,B)$$

for every incentive-compatible mechanism f' that has $f'(1,\ldots,1) = f(1,\ldots,1)$.

Our goal is to design mechanisms that are incentive-compatible, and whenever they exist, find the mechanism(s) that is(are) most frugal.

4. An Impossibility Result

It turns out that, unfortunately, we must face a roadblock in the first step: We can show that there exists no mechanism that is incentive compatible.

Theorem 4.1. *For any N, G and $B \geq 2$, there is no mechanism that can guarantee that the worker will be incentivized to select precisely the support of her distribution for each question.*

In order to circumvent this impossibility result, we appeal to a certain well-understood property of human belief described in the following section.

5. Coarse Belief Assumption

There is an extensive literature in psychology establishing the coarseness of processing and perception in humans. For instance, Miller's celebrated paper (Miller, 1956) establishes the information and storage capacity of humans, that an average human being can typically distinguish at most about seven states. This granualrity of human computation is verified in many subsequent experiments (Shiffrin & Nosofsky, 1994; Saaty & Ozdemir, 2003). Jones & Loe (2013) establish the ineffectiveness of finer-granularity response elicitation. Mullainathan et al. (2008) hypothesize that humans often group things into categories; this hypothesis is experimentally verified by Siddiqi (2011) in a specific setting. We incorporate this established notion of coarseness of human processing in our model in terms of a simple assumption.

Consider some (fixed and known) value $\rho > 0$, and assume that the probability of any option for any question, according to the worker's belief, is either zero or greater than ρ. The impossibility shown in Theorem 4.1 pertains to $\rho = 0$. Also, one must necessarily take into account situations when a worker is totally clueless about a question, that is, when her belief is distributed uniformly over all options. Hence we restrict $\rho < \frac{1}{B}$. To summarize, we make the following "coarse belief" assumption.

Definition 3 (Coarse belief assumption). *The worker's belief for any option for any question lies in the set $\{0\}\cup(\rho,1]$ for some (fixed and known) $\rho \in \left(0, \frac{1}{B}\right)$.*

We wish to elicit the full support of the workers' beliefs, given a coarseness of belief that assigns a value of zero to very low probability categories. The goal is to design mechanisms that are incentive-compatible and frugal, assuming the coarse belief assumption holds true.

6. Incentive Mechanism

We now present a mechanism for the problem at hand, under the coarse belief assumption. The mechanism is described in Algorithm 1.

Algorithm 1 Incentive mechanism for approval voting

- **Input:** evaluations of the worker's answers to the G gold standard questions $(x_1,\ldots,x_G) \in \{-(B-1),\ldots,-1,1,\ldots,B\}^G$
- **Output:** the worker's payment

$$f(x_1,\ldots,x_G) = \alpha(1-\rho)^{\sum_{i=1}^{G}(x_i-1)} \prod_{i=1}^{G} \mathbf{1}\{x_i \geq 1\}$$

The payment is based only on the evaluation of the worker's responses to the gold standard questions. It is easy to describe the mechanism in words:

1. The payment is zero if the correct answer is not selected for any of the questions;
2. Otherwise, it equals α reduced by $(100\rho)\%$ for each option selected except the correct option.

The following pair of theorems present our main results, proving that this mechanism is *the one and only* mechanism that satisfies our requirements.

Theorem 6.1. *The mechanism of Algorithm 1 is incentive-compatible and frugal.*

The following theorem shows that our mechanism is *strictly* better than any other mechanism.

Theorem 6.2. *There is no other incentive-compatible mechanism that expends as small an amount as Algorithm 1 on a worker who does not attempt any question.*

To show the optimality and uniqueness properties claimed in Theorem 6.1 and Theorem 6.2 respectively, we prove the absence of other good mechanisms via contradiction-based arguments. Specifically, for any candidate mechanism, we identify a specific set of beliefs for which the worker will not be incentivized to act as required. In line with our earlier argument of beliefs being "coarse", the beliefs considered in these proofs are simple enough: the worker has some belief about one of the options, knows for sure that certain other options are incorrect, and is indifferent among the rest of the options.

To put things in perspective, observe that $\rho = 0$ eliminates the dependence of the payment in Algorithm 1 on $\sum_i x_i$ and makes the mechanism incentive incompatible. The impossibility result of Theorem 4.1 proves that every possible mechanism must necessarily suffer this fate.

The remainder of this section is devoted to the proof of Theorem 6.1. The reader may feel free to jump to Section 7 without any loss in continuity.

6.1. Proof of Theorem 6.1

Incentive compatibility. First consider the case $N = G = 1$. In this case, the mechanism of Algorithm 1 reduces to

$$f(x) = \alpha(1-\rho)^{(x_1-1)}\mathbf{1}\{x_1 \geq 1\}.$$

Suppose without loss of generality that the worker's beliefs for the B options are $p_1 \geq \cdots \geq p_m > \rho > p_{m+1} = \cdots = p_B = 0$ for some $m \in [B]$. An incentive-compatible mechanism must strictly maximize the worker's expected payment when she selects the support of her belief, that is, the options $\{1,\ldots,m\}$. The expected payment, $\$_{\text{sup}}$, under this selection is

$$\begin{aligned}\$_{\text{sup}} &= \alpha \sum_{i=1}^{m} p_i (1-\rho)^{m-1} \\ &= (1-\rho)^{m-1}.\end{aligned}$$

Suppose the worker selects some other set of options $\{o_1,\ldots,o_\ell\} \subseteq [B]$, $\{o_1,\ldots,o_\ell\} \neq [m]$. Then her expected payment $\$_{\text{oth}}$ under the proposed mechanism for this selection is

$$\begin{aligned}\$_{\text{oth}} &= \alpha \sum_{i=1}^{\ell} p_{o_i} (1-\rho)^{\ell-1} \\ &\leq \alpha \sum_{i=1}^{\ell} p_i (1-\rho)^{\ell-1}, \qquad (4)\end{aligned}$$

since $p_1 \geq \cdots \geq p_B$. If $\ell = m$ then the inequality in (4) is strict since $p_j < p_i$ for all $(j > m,\, i \leq m)$. Thus the expected payment under the choice $\ell = m$ but with a selection different from the support is strictly lower than $\$_{\text{sup}}$. Also observe that the expected payment on selecting $\ell > m$ is upper bounded by $(1-\rho)^{\ell-1}$, which is strictly smaller than $\$_{\text{sup}}$. Let us now consider the remaining, interesting case of $\ell < m$. Since $p_i > \rho$ for all $i \in [m]$, we have

$$\begin{aligned}\$_{\text{oth}} &< \alpha\left(\sum_{i=1}^{m} p_i - (m-\ell)\rho\right)(1-\rho)^{\ell-1} \\ &= \alpha\,(1-(m-\ell)\rho)\,(1-\rho)^{\ell-1} \\ &\leq \alpha\,(1-(m-(\ell+1))\rho)\,(1-\rho)^{\ell} \\ &\vdots \\ &\leq \alpha(1-\rho)^{m-1} \\ &= \$_{\text{sup}}.\end{aligned}$$

This completes the proof for the case $N = G = 1$.

Let us now consider the case of $N = G \geq 1$. By our assumption of the independence of the beliefs of the worker across the questions, the expected payment equals

$$\prod_{i=1}^{G} \mathbf{E}\left[\alpha(1-\rho)^{(x_i-1)}\mathbf{1}\{x_i \geq 1\}\right].$$

Since the payments are non-negative, if each individual component in the product is maximized then the product is also necessarily maximized. Each individual component simply corresponds to the setting of $N = G = 1$ discussed earlier. Thus calling upon our earlier result, we get that the expected payment for the case $N = G > 1$ is maximized when the worker acts as desired for every question.

Let us finally consider the case of $N > G \geq 1$. Recall from (3) that the expected payment for the general case is a cascade of two expectations: the outer expectation is with respect to the uniformly random distribution of the G gold standard questions among the N total questions, while the inner expectation is taken over the worker's beliefs of the different questions conditioned on the choice of the gold standard questions. The arguments above for the case $N = G$ prove that every individual term in the inner expectation is maximized when the worker acts as desired. The expected payment is thus maximized when the worker acts as desired.

Frugality. We first present a lemma that forms the workhorse of this and other subsequent proofs.

Lemma 6.3. *Consider some $y, y' \in [B]^N$ and some $\mathcal{I} \subseteq [N]$ such that $y_i = y'_i + 1$ for all $i \in \mathcal{I}$, and $y_i = y'_i$ for all $i \notin \mathcal{I}$. Then any incentive compatible mechanism f must necessarily satisfy*

$$\frac{1}{\binom{N}{G}} \sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1},\ldots,y_{j_G}) \geq \frac{1}{\binom{N}{G}} \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1},\ldots,y'_{j_G}).$$

Furthermore, a necessary condition for the above equation to be satisfied with equality is

$$f(\epsilon_1 y'_{j_1},\ldots,\epsilon_G y'_{j_G}) = 0$$

for all $(j_1,\ldots,j_G) \subseteq [N]$, and all $\{(\epsilon_1,\ldots,\epsilon_G) \in \{-1,1\}^G \backslash \{1\}^G \mid \epsilon_i = 1$ whenever $j_i \notin \mathcal{I}\}$.

We now prove the frugality of our proposed mechanism using this lemma. Consider any incentive compatible mechanism f such that $f(1,\ldots,1) = \alpha$. Consider any $x_0 \in [B-1]$. Applying Lemma 6.3 with $y = (x_0+1,\ldots,x_0+1)$ and $y' = (x_0,\ldots,x_0)$ gives

$$f(x_0+1,\ldots,x_0+1) \geq (1-\rho)^G f(x_0,\ldots,x_0).$$

A repeated application of this inequality for all $x_0 \in [B-1]$ gives

$$\begin{aligned} f(B,\ldots,B) &\geq (1-\rho)^{(B-1)G} f(1,\ldots,1) \\ &= (1-\rho)^{(B-1)G}\alpha. \end{aligned}$$

The mechanism of Algorithm 1 achieves this lower bound on $f(B,\ldots,B)$ with equality, thereby completing the proof.

7. Generalization

We earlier made the "coarse belief" assumption that the worker's belief for any option, when non-zero, is atleast ρ. We then designed the mechanism of Algorithm 1 that is incentive compatible with respect to eliciting the supports of the beliefs of the worker. In this section, we generalize the results presented earlier in the paper to the setting where workers may have arbitrary beliefs.

7.1. Incentivizing Workers with Finer Beliefs

Suppose the mechanism of Algorithm 1, for a certain value of ρ, is encountered by a worker who may have arbitrary beliefs. Interestingly, it turns out that the mechanism doesn't break down, but instead does something desirable: it incentivizes the worker to select all options for which the relative belief of the worker is high enough.

Theorem 7.1. *Under the mechanism of Algorithm 1, for any question, a worker with beliefs $1 \geq p_1 \geq \ldots \geq p_B \geq 0$ will be incentivized to select options $\{1,\ldots,m\}$ where*

$$m = \underset{z\in[m]}{\mathit{arg\,max}} \left(\frac{p_z}{\sum_{i=1}^{z} p_i} > \rho \right).$$

It is not hard to interpret this incentivized action. The worker selects options one by one in decreasing order of her beliefs as long as the selected option contributes a fraction more than ρ to the total belief of the selected options.

Let us now verify that the earlier result of Theorem 6.1 for "coarse beliefs" is indeed a special case of Theorem 7.1. To this end, suppose the beliefs of the worker for any particular question are $p_1 \geq \cdots p_k > \rho > p_{k+1} = \cdots = p_B = 0$ for some $k \in [B]$. Then we have

$$\frac{p_z}{\sum_{i=1}^{z} p_i} = \frac{0}{\sum_{i=1}^{z} p_i} = 0 < \rho \qquad \text{for all } z \geq k+1,$$

and

$$\frac{p_z}{\sum_{i=1}^{z} p_i} \geq \frac{p_z}{1} > \rho \qquad \text{for all } z \leq k.$$

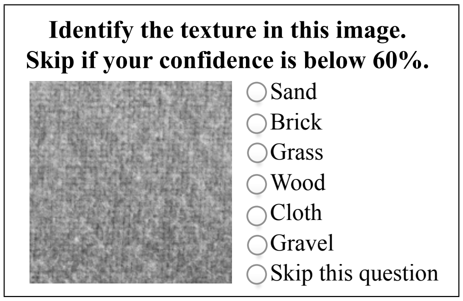
It follows that under the result of Theorem 7.1, a worker with "coarse beliefs" will be incentivized to select precisely the support of her beliefs.

7.2. An Alternate Axiomatic Derivation

We now present an alternative axiomatic derivation of our mechanism when accommodating workers with arbitrary beliefs. The derivation involves a "no-free-lunch axiom" of Shah & Zhou (2014), which when adapted to our approval-voting based setting is defined as follows. We say that a worker has 'attempted' a question if for that question, she doesn't select all the B options. We say that the answer to a question is wrong if the correct option does not lie in the set of selected options.

Definition 4 (No-free-lunch; adapted from Shah & Zhou (2014))**.** *If the answer to every attempted question in the gold standard turns out to be wrong, then the worker gets a payment of zero, namely,*

$$\begin{aligned} &f(x_1,\ldots,x_G) = 0 \\ &\qquad \forall\ (x_1,\ldots,x_G) \in \{-(B-1),\ldots,-1,B\}^G \backslash \{B\}^G. \end{aligned}$$



Identify the texture in this image.
Skip if your confidence is below 60%.

○ Sand
○ Brick
○ Grass
○ Wood
○ Cloth
○ Gravel
○ Skip this question

Figure 2. An illustration of the experiment on identifying textures. The figure depicts the interface for the skip-based mechanism of Shah & Zhou (2014).

The no-free-lunch axiom is quantitatively different from the criterion of frugality proposed in this paper. However, both these notions have the same qualitative goal, namely to minimize the expenditure when no useful data is obtained, while providing higher payments to workers providing better data. Interestingly, as we show below, both these notions lead to the same (unique) mechanism under our setting of approval voting.

Theorem 7.2. *Consider no assumptions on the minimum value of the belief, and suppose the workers must be incentivized to select options* $\{1, \ldots, m\}$ *where* $m = arg\,max_z \left(\frac{p_z}{\sum_{i=1}^{z} p_i} > \rho \right)$. *Then, the mechanism of Algorithm 1 is the one and only mechanism that is incentive compatible and satisfies no-free-lunch.*

8. Preliminary Experiments

This section presents results from an evaluation of our proposed mechanism on the popular Amazon Mechanical Turk (`mturk.com`) commercial crowdsourcing platform. The goal of this preliminary experimental exercise is to perform a basic check on whether our mechanism has the potential to work in practice. Specifically, our goal is to evaluate the primary hypotheses underlying the theory: (a) whether workers are able to make a judicious use of the approval voting setup, (b) whether the existence of the mechanism improves the quality, and (c) if there is a strong opposition from the workers to the interface or the mechanism for any reason.

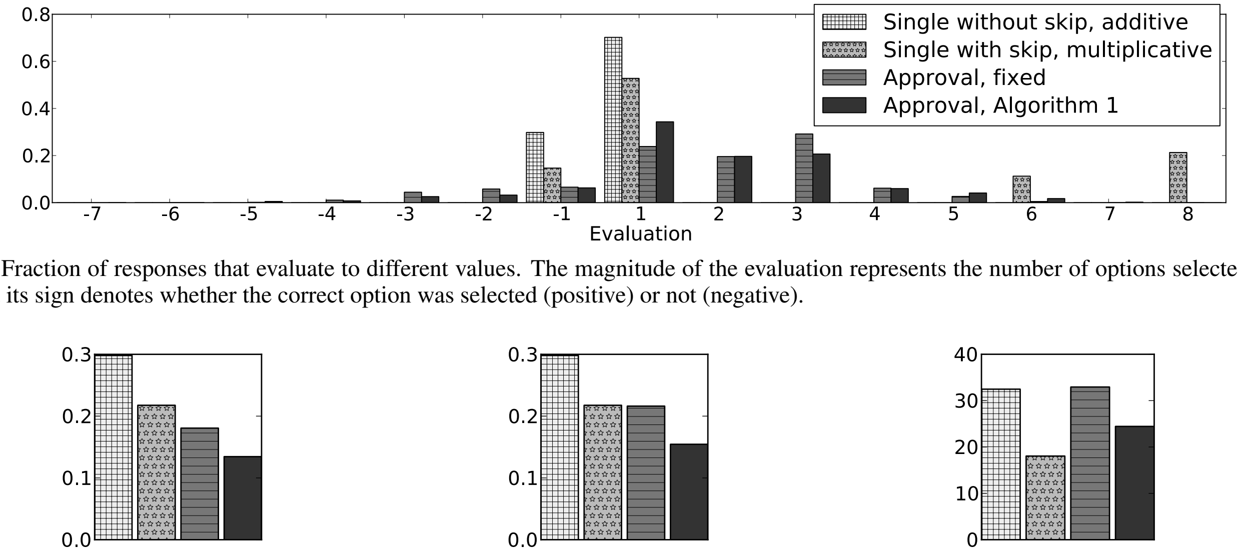
We note that conclusive experiments for mechanism design are in general quite expensive with respect to time (workers may need months to understand a new mechanism) and budget. They are unlike typical machine-learning experiments that require only existing benchmark datasets. Like most mechanism design papers, we position our work primarily as a theoretical study. We expect that more detailed experiments will follow the publication of our work; indeed, it is best if experiments on such incentive schemes are conducted by multiple groups.

We conducted two separate sets of experiments, with over 200 workers in each experiment. The first set required workers to identify the languages from displayed text (Figure 1). The second set of experiments required workers to identify the textures in images displayed to them (Figure 8). In each experiment, every worker was assigned one of four mechanisms uniformly at random. The mechanisms were executed as a “bonus payment” based on the evaluation of the worker’s performance on the gold standard questions, on top of a guaranteed payment of 10 cents. The four mechanisms tested were:

- Single-selection interface with additive payments: The bonus starts at zero and is increased by a fixed amount for every correct answer.
- Skip-based single-selection interface with multiplicative payments (Shah & Zhou, 2014): The bonus starts at a certain positive value, is reduced by a certain fraction for each skipped question, and becomes zero in case of an incorrect answer.
- Approval-voting interface with a fixed payment: The bonus is fixed.
- Approval-voting interface with the payment defined in Algorithm 1.

The entire data related to the experiments is available on the website of the first author.

Figure 3 presents aggregate results from the two experiments. Figure 3(a) shows the breakup of the evaluations of all the collected responses. Observe that the workers made judicious use of the approval-voting interface whenever made available to them, with more than 40% responses comprising a selection of two or three options. Figure 3(b) depicts the fraction of responses to attempted questions that turned out to be wrong. Figure 3(c) depicts the fraction of responses that were correct when only one option was selected. A comparison with the results for single-selection in the plots reveals that the approval-voting based approach coupled with our incentive-compatible mechanism was successful in converting a significant fraction of incorrect answers to correct partial information. Figure 3(d) depicts the average payment per worker. We can see that our mechanism is associated to a very little or no increase in the net expenditure. Finally, we also elicited feedback about the task from every worker; we did not receive any negative feedback about either the approval voting interface or our mechanism. All in all, these experiments on Amazon Mechanical Turk indicate that our mechanism is indeed practical and can potentially be useful for many applications in machine learning.



(a) Fraction of responses that evaluate to different values. The magnitude of the evaluation represents the number of options selected and its sign denotes whether the correct option was selected (positive) or not (negative).

(b) Fraction wrong among attempted questions

(c) Fraction wrong when only one option was selected

(d) Average bonus per worker (cents)

Figure 3. Results from the two experiments conducted on Amazon Mechanical Turk.

A standard means of denoising data from crowdsourcing is to ask every question to multiple workers, and employ a statistical aggregation algorithm to aggregate the data so obtained. In the future, we wish to evaluate the performance of our proposed interface and mechanism on such aggregated data. To this end, our goal for the future is to design algorithms designed towards statistical aggregation of data collected through the interface and mechanism proposed in this paper.

9. Discussion and Future Work

Our goal is to deliver high quality labels for machine learning applications, at low costs, by means of incentive mechanisms or aggregation algorithms or both. In this paper, we pursue the former approach. We take an approval-voting based means of gathering labeled data from crowdsourcing. In particular, we use the approval-voting interface to elicit the support of the beliefs of workers. This approach is complementary to that of eliciting a single answer (the mode of the belief), and may often be able to utilize more effectively the expertise of workers who have partial knowledge of the true answer. We design an incentive mechanism via a principled theoretical approach; we prove the appealing properties of optimality and uniqueness of the mechanism. Preliminary experiments conducted on Amazon Mechanical Turk corroborate the usefulness of this mechanism for practical scenarios. Our mechanism may also draw more experts to the crowdsourcing platform since their compensation will be significantly higher than that of mediocre workers, unlike most compensation mechanisms in current use.

For the traditional single-selection setting, there is a long, existing line of work on statistical methods to aggregate redundant noisy data from multiple workers (Dawid & Skene, 1979; Whitehill et al., 2009; Raykar et al., 2010; Karger et al., 2011; Liu et al., 2012; Zhou et al., 2012). An open problem is the design of aggregation algorithms for approval-voting-based data: algorithms that can exploit the specific structure of the responses that arise as a result of the proposed interface and mechanism. There is indeed work on aggregation algorithms (Massó & Vorsatz, 2008; Caragiannis et al., 2010; Brams & Kilgour, 2014) and probabilistic models (Marley, 1993; Falmagne & Regenwetter, 1996; Doignon et al., 2004; Regenwetter & Tsetlin, 2004) for approval-voting in the literature social choice theory; their objective, however, is primarily of fairness and stretgyproofing of the voting procedure, as opposed to our goal of noise removal as required for labeling tasks in crowdsourcing.

References

Brams, Steven J and Fishburn, Peter C. Approval voting. *American Political Science Review*, 72(03):831–847, 1978.

Brams, Steven J and Kilgour, D Marc. Satisfaction approval voting. In *Voting Power and Procedures*, pp. 323–346. Springer, 2014.

Brier, Glenn W. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Caragiannis, Ioannis, Kalaitzis, Dimitris, and Markakis, Evangelos. Approximation algorithms and mechanism design for minimax approval voting. In *AAAI*, 2010.

Collet, Leverne S. Elimination scoring: An empirical evaluation. *Journal of Educational Measurement*, 8(3):209–214, 1971.

Coombs, Clyde H. On the use of objective examinations. *Educational and Psychological Measurement*, 13 (2):308–310, 1953.

Coombs, Clyde H, Milholland, John Edgar, and Womer, Frank Burton. The assessment of partial knowledge. *Educational and Psychological Measurement*, 16(1):13–37, 1956.

Dasgupta, Anirban and Ghosh, Arpita. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 319–330. International World Wide Web Conferences Steering Committee, 2013.

Dawid, Alexander Philip and Skene, Allan M. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics*, pp. 20–28, 1979.

Doignon, Jean-Paul, Pekeč, Aleksandar, and Regenwetter, Michel. The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika*, 69(1):33–54, 2004.

Falmagne, J-Cl and Regenwetter, Michael. A random utility model for approval voting. *Journal of Mathematical Psychology*, 40(2):152–159, 1996.

Faltings, Boi, Jurca, Radu, Pu, Pearl, and Tran, Bao Duy. Incentives to counter bias in human computation. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

Gibbons, Jean D, Olkin, Ingram, and Sobel, Milton. A subset selection technique for scoring items on a multiple choice test. *Psychometrika*, 44(3):259–270, 1979.

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

Horst, Paul. The chance element in the multiple choice test item. *The Journal of General Psychology*, 6(1):209–211, 1932.

Jain, Shaili, Chen, Yiling, and Parkes, David C. Designing incentives for online question and answer forums. In *Proceedings of the 10th ACM conference on Electronic commerce*, pp. 129–138, 2009.

Jones, W Paul and Loe, Scott A. Optimal number of questionnaire response categories more may not be better. *SAGE Open*, 3(2):2158244013489691, 2013.

Karger, David R, Oh, Sewoong, and Shah, Devavrat. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pp. 1953–1961, 2011.

Kazai, Gabriella, Kamps, Jaap, Koolen, Marijn, and Milic-Frayling, Natasa. Crowdsourcing for book search evaluation: impact of HIT design on comparative system ranking. In *ACM SIGIR conference on Research and development in Information Retrieval*, pp. 205–214, 2011.

Kellett, John and Mott, Kenneth. Presidential primaries: Measuring popular choice. *Polity*, pp. 528–537, 1977.

Lambert, Nicolas and Shoham, Yoav. Eliciting truthful answers to multiple-choice questions. In *ACM conference on Electronic commerce*, pp. 109–118, 2009.

Liu, Qiang, Peng, Jian, and Ihler, Alexander T. Variational inference for crowdsourcing. In *NIPS*, pp. 701–709, 2012.

Marley, AAJ. Aggregation theorems and the combination of probabilistic rank orders. In *Probability models and statistical analyses for ranking data*, pp. 216–240. Springer, 1993.

Massó, Jordi and Vorsatz, Marc. Weighted approval voting. *Economic Theory*, 36(1):129–146, 2008.

Miller, George A. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Miller, Nolan, Resnick, Paul, and Zeckhauser, Richard. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

Mullainathan, Sendhil, Schwartzstein, Joshua, and Shleifer, Andrei. Coarse thinking and persuasion*. *The Quarterly journal of economics*, 123(2):577–619, 2008.

Ottewell, Guy. The arthmetic of voting. *In defence of variety*, 1977.

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

Raykar, Vikas C, Yu, Shipeng, Zhao, Linda H, Valadez, Gerardo Hermosillo, Florin, Charles, Bogoni, Luca, and Moy, Linda. Learning from crowds. *The Journal of Machine Learning Research*, 11:1297–1322, 2010.

Regenwetter, Michel and Tsetlin, Ilia. Approval voting and positional voting methods: Inference, relationship, examples. *Social Choice and Welfare*, 22(3):539–566, 2004.

Saaty, Thomas L and Ozdemir, Mujgan S. Why the magic number seven plus or minus two. *Mathematical and Computer Modelling*, 38(3):233–244, 2003.

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

Shah, Nihar B. and Zhou, Dengyong. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. *arXiv:1408.1387*, 2014.

Shah, Nihar B, Balakrishnan, Sivaraman, Bradley, Joseph K, Parekh, Abhay, Ramchandran, Kannan, and Wainwright, Martin. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *AIStats*, 2015.

Shiffrin, RM and Nosofsky, RM. Seven plus or minus two: a commentary on capacity limitations. *Psychological review*, 101(2):357, 1994.

Siddiqi, Hammad. Does coarse thinking matter for option pricing? evidence from an experiment. *IUP Journal of Behavioral Finance*, 8(2), 2011.

Vuurens, Jeroen, de Vries, Arjen P, and Eickhoff, Carsten. How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In *ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, pp. 21–26, 2011.

Wais, Paul, Lingamneni, Shivaram, Cook, Duncan, Fennell, Jason, Goldenberg, Benjamin, Lubarov, Daniel, Marin, David, and Simons, Hari. Towards building a high-quality workforce with Mechanical Turk. *NIPS workshop on computational social science and the wisdom of crowds*, 2010.

Weber, Robert J. Comparison of voting systems. *New Haven: Cowles Foundation Discussion paper A*, 498, 1977.

Whitehill, Jacob, Ruvolo, Paul, Wu, Ting-fan, Bergsma, Jacob, and Movellan, Javier. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pp. 2035–2043, 2009.

Zhou, Dengyong, Platt, John, Basu, Sumit, and Mao, Yi. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems 25*, pp. 2204–2212, 2012.

Zou, James, Meir, Reshef, and Parkes, David. Approval voting behavior in doodle polls. In *The 5th Workshop on Computational Social Choice*, June 2014.