

Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing

Nihar B. Shah

*Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720 USA*

NIHAR@EECS.BERKELEY.EDU

Dengyong Zhou

*Machine Learning Department
Microsoft Research
One Microsoft Way, Redmond 98052 USA*

DENGYONG.ZHOU@MICROSOFT.COM

Editor: Qiang Liu

Abstract

Crowdsourcing has gained immense popularity in machine learning applications for obtaining large amounts of labeled data. Crowdsourcing is cheap and fast, but suffers from the problem of low-quality data. To address this fundamental challenge in crowdsourcing, we propose a simple payment mechanism to incentivize workers to answer only the questions that they are sure of and skip the rest. We show that surprisingly, under a mild and natural “no-free-lunch” requirement, this mechanism is the one and only incentive-compatible payment mechanism possible. We also show that among all possible incentive-compatible mechanisms (that may or may not satisfy no-free-lunch), our mechanism makes the smallest possible payment to spammers. We further extend our results to a more general setting in which workers are required to provide a quantized confidence for each question. Interestingly, this unique mechanism takes a “multiplicative” form. The simplicity of the mechanism is an added benefit. In preliminary experiments involving over 900 worker-task pairs, we observe a significant drop in the error rates under this unique mechanism for the same or lower monetary expenditure.

Keywords: high-quality labels, supervised learning, crowdsourcing, mechanism design, proper scoring rules

1. Introduction

Complex machine learning tools such as deep learning are gaining increasing popularity and are being applied to a wide variety of problems. These tools require large amounts of labeled data (Hinton et al., 2012; Raykar et al., 2010; Deng et al., 2009; Carlson et al., 2010). These large labeling tasks are being performed by coordinating crowds of semi-skilled workers through the Internet. This is known as crowdsourcing. Generating large labeled data sets through crowdsourcing is inexpensive and fast as compared to employing experts. Furthermore, given the current platforms for crowdsourcing such as Amazon Mechanical Turk and many others, the initial overhead of setting up a crowdsourcing task is minimal. Crowdsourcing as a means of collecting labeled training data has now become indispensable to the engineering of intelligent systems. The crowdsourcing of labels is also often used to supplement automated algorithms, to perform the tasks that are too difficult

to accomplish by machines alone (Khatib et al., 2011; Lang and Rio-Ross, 2011; Bernstein et al., 2010; Von Ahn et al., 2008; Franklin et al., 2011).

Most workers in crowdsourcing are not experts. As a consequence, labels obtained from crowdsourcing typically have a significant amount of error (Kazai et al., 2011; Vuurens et al., 2011; Wais et al., 2010). It is not surprising that there is significant emphasis on having higher quality labeled data for machine learning algorithms, since a higher amount of noise implies requirement of more labels for obtaining the same accuracy in practice. Moreover, several algorithms and settings are not very tolerant of data that is noisy (Long and Servedio, 2010; Hanneke and Yang, 2010; Manwani and Sastry, 2013; Baldridge and Palmer, 2009); for instance, Long and Servedio (2010) conclude that “a range of different types of boosting algorithms that optimize a convex potential function satisfying mild conditions cannot tolerate random classification noise.” Recent efforts have focused on developing statistical techniques to post-process the noisy labels in order to improve its quality (e.g., Raykar et al., 2010; Zhou et al., 2012; Chen et al., 2013; Dawid and Skene, 1979; Karger et al., 2011; Liu et al., 2012; Zhang et al., 2014; Ipeirotis et al., 2014; Zhou et al., 2015; Khetan and Oh, 2016; Shah et al., 2016c). However, when the inputs to these algorithms are highly erroneous, it is difficult to guarantee that the processed labels will be reliable enough for subsequent use by machine learning or other applications. In order to avoid “garbage in, garbage out”, we take a complementary approach to this problem: cleaning the data at the time of collection.

We consider crowdsourcing settings where the workers are paid for their services, such as in the popular crowdsourcing platforms of Amazon Mechanical Turk (mturk.com), Crowdflower (crowdflower.com) and other commercial platforms, as well as internal crowdsourcing platforms of companies such as Google, Facebook and Microsoft. These commercial platforms have gained substantial popularity due to their support for a diverse range of tasks for machine learning labeling, varying from image annotation and text recognition to speech captioning and machine translation. We consider problems that are objective in nature, that is, have a definite answer. Figure 1a depicts an example of such a question where the worker is shown a set of images, and for each image, the worker is required to identify if the image depicts the Golden Gate Bridge.

Our approach builds on the simple insight that in typical crowdsourcing setups, workers are simply paid in proportion to the amount of tasks they complete. As a result, workers attempt to answer questions that they are not sure of, thereby increasing the error rate of the labels. For the questions that a worker is not sure of, her answers could be very unreliable (Wais et al., 2010; Kazai et al., 2011; Vuurens et al., 2011; Jagabathula et al., 2014). To ensure acquisition of only high-quality labels, we wish to encourage the worker to skip the questions about which she is unsure, for instance, by providing an explicit “I’m not sure” option for every question (see Figure 1b). Given this additional option, one must also ensure that the worker is indeed incentivized to skip the questions that she is not confident about. In a more general form, we consider eliciting the confidence of the worker for each question at multiple levels. For instance, in addition to “I’m not sure”, we may also provide options like “absolutely sure”, and “moderately sure” (see Figure 1c). The goal is to design payment mechanisms that incentivize the worker to attempt only those questions for which they are confident enough, or alternatively, report their confidences truthfully. As we will see later, this significantly improves the aggregate quality of the labels that are input to the machine learning algorithms. We will term any payment mechanism that incentivizes the worker to do so as “incentive compatible”.

In addition to incentive compatibility, preventing spammers is another desirable requirement from incentive mechanisms in crowdsourcing. Spammers are workers who answer randomly with-




<p>a Is this the Golden Gate Bridge?</p>  <p><input type="radio"/> Yes <input type="radio"/> No</p>	<p>b Is this the Golden Gate Bridge?</p>  <p><input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> I'm not sure</p>
<p>c Is this the Golden Gate Bridge?</p>  <p>Yes <input type="radio"/> Moderately sure <input type="radio"/> Absolutely sure No <input type="radio"/> Moderately sure <input type="radio"/> Absolutely sure <input type="radio"/> I'm not sure</p>	

Figure 1: Different interfaces for a task that requires the worker to answer the question “Is this the Golden Gate Bridge?”: (a) the conventional interface; (b) with an option to skip; (c) with multiple confidence levels.

out regard to the question being asked, in the hope of earning some free money, and are known to exist in large numbers on crowdsourcing platforms (Wais et al., 2010; Bohannon, 2011; Kazai et al., 2011; Vuurens et al., 2011). The presence of spammers can significantly affect the performance of any machine learning algorithm that is trained on this data. It is thus of interest to deter spammers by paying them as low as possible. An intuitive objective, to this end, is to ensure a minimum possible payment to spammers who answer randomly. For instance, in a task with binary-choice questions, a spammer is expected to have half of the attempted answers incorrect; one may thus wish to set the payment to its minimum possible value if half or more of the attempted answers are wrong. In this paper, however, we impose *strictly and significantly weaker requirement*, and then show that there is one and only one incentive-compatible mechanism that can satisfy this weak requirement. Our requirement is referred to as the “no-free-lunch” axiom. In the skip-based setting, it says that if *all* the questions attempted by the worker are answered incorrectly, then the payment must be the minimum possible. The no-free-lunch axiom for the general confidence-based setting is even weaker: if the worker indicates the highest confidence level for *all* the questions she attempts in the gold standard, and furthermore if all these responses are incorrect, then the payment must be the minimum possible. We term this condition the “no-free-lunch” axiom. In the general confidence-based setting, we want to make the minimum possible payment if the worker indicates the *highest confidence level* for *all* the questions she attempts *and* if *all* these responses are incorrect.

In order to test whether our mechanism is practically viable, and to assess the quality of the final labels obtained, we conducted experiments on the Amazon Mechanical Turk crowdsourcing platform. In our preliminary experiments that involved several hundred workers, we found that the quality of data consistently improved by use of our schemes as compared to the standard settings, often by two-fold or higher, with the total monetary expenditure being the same or lower as compared to the conventional baseline.

1.1 Summary of Contributions

We propose a payment mechanism for the aforementioned setting (“incentive compatibility” plus “no-free-lunch”), and show that surprisingly, this is the *only* possible mechanism. We also show that additionally, our mechanism makes the smallest possible payment to spammers among all possible incentive compatible mechanisms that may or may not satisfy the no-free-lunch axiom. Interestingly, our payment mechanism takes a multiplicative form: the evaluation of the worker’s response to each question is a certain score, and the final payment is a *product* of these scores. This mechanism has additional appealing features in that it is simple to compute, and is also simple to explain to the workers. Our mechanism is applicable to any type of objective questions, including multiple choice annotation questions, transcription tasks, etc. In preliminary experiments on Amazon Mechanical Turk involving over 900 worker-task pairs, the quality of data improved significantly under our unique mechanism, with the total monetary expenditure being the same or lower as compared to the conventional baseline.

1.2 Related Literature

The framework of “strictly proper scoring rules” (Brier, 1950; Savage, 1971; Gneiting and Raftery, 2007; Lambert and Shoham, 2009) provides a general theory for eliciting information for settings where this information can subsequently be verified by the mechanism designer, for example, by observing the true value some time in the future. In our work, this verification is performed via the presence of some “gold standard” questions in the task. Consequently, our mechanisms can also be called “strictly proper scoring rules”. It is important to note that the framework of strictly proper scoring rules, however, provides a large collection of possible mechanisms and does not guide the choice of a specific mechanism from this collection (Gneiting and Raftery, 2007). In this work, we show that for the crowdsourcing setups considered, under a very mild condition we term the “no-free-lunch” axiom, the mechanism proposed in this paper is the one and only strictly proper scoring rule.

Interestingly, proper scoring rules have another interesting connection with machine learning techniques: quoting Buja et al. (2005), “proper scoring rules comprise most loss functions currently in use: log-loss, squared error loss, boosting loss, and as limiting cases cost-weighted misclassification losses.” The present paper does not investigate this aspect of proper scoring rules, and we refer the reader to Bühlmann and Hothorn (2007); Mease et al. (2007); Buja et al. (2005) for more details.

In this paper, we assume the existence of some gold standard questions whose answers are known a priori to the system designer. As a result, the payment to a worker is determined solely by her own work. There are settings where gold standard questions may not be available, for instance, when obtaining gold standard questions is too expensive, or when the questions pertain to subjective preferences (Shah and Wainwright, 2015; Shah et al., 2016b; Chen et al., 2016) instead of labeling data. A parallel line of literature (Miller et al., 2005; Dasgupta and Ghosh, 2013; Prelec, 2004; Wolfers and Zitzewitz, 2004; Conitzer, 2009) addresses such settings without gold standard questions. The idea in the mechanisms designed therein is to reward the agents based on certain criteria that compares certain elicited data from the agents with each other, and typically involves asking agents to predict other agents’ responses. The mechanisms designed often provide weaker guarantees (such as that of truth-telling being a Nash equilibrium) due to the absence of a gold standard answer to compare with. This line of literature includes work on peer-prediction (Miller

et al., 2005; Dasgupta and Ghosh, 2013), the Bayesian truth serum (Prelec, 2004) and prediction markets (Wolfers and Zitzewitz, 2004; Conitzer, 2009).

The design of statistical inference algorithms for denoising the data obtained from workers is an active topic of research (Raykar et al., 2010; Zhou et al., 2012; Wauthier and Jordan, 2011; Chen et al., 2013; Khetan and Oh, 2016; Dawid and Skene, 1979; Karger et al., 2011; Liu et al., 2012; Zhang et al., 2014; Vempaty et al., 2014; Ipeirotis et al., 2014; Zhou et al., 2015; Shah et al., 2016c). In addition, several machine learning algorithms accommodating errors in the data have also been designed (Angluin and Laird, 1988; Cano et al., 2001; Lee et al., 2004; Chu et al., 2004). These algorithms are typically oblivious to the elicitation procedure. Our work nicely complements this line of research in that these inference algorithms may now additionally employ the higher quality data and the specific structure of the elicited data for an improved denoising efficiency.

Another relevant problem in crowdsourcing is that of choosing which workers to hire or efficiently matching workers to tasks, and such problems are studied in Yuen et al. (2011); Ho et al. (2013); Zhou et al. (2014); Anari et al. (2014) under different contexts. Our work assumes that a worker is already matched, and focuses on incentivizing that worker to respond in a certain manner. A recent line of work has focused on elicitation of data from multiple agents in order to perform certain specific estimation tasks (Fang et al., 2007; Dekel et al., 2008; Cai et al., 2015). In contrast, our goal is to ensure that workers censor their own low-quality (raw) data, without restricting our attention to any specific downstream algorithm or task.

1.3 Organization

The organization of this paper is as follows. We present the formal problem setting in Section 2. In Section 3 we consider the skip-based setting: We present our proposed mechanism and show that it is the only mechanism which satisfies the requirements discussed above. In Section 4, we then consider the more general setting of eliciting a quantized value of the worker’s confidence. We construct a mechanism for this setting, which also takes a multiplicative form, and prove its uniqueness. In Section 5 we prove that imposing a requirement that is only slightly stronger than our proposed no-free-lunch axiom leads to impossibility results. In Section 6 we present synthetic simulations and real-world experiments on Amazon Mechanical Turk to evaluate the potential of our setting and algorithm to work in practice. We conclude the paper with a discussion on the various modeling choices, future work, and concluding remarks in Section 7.

The paper contains three appendices. In Appendix A we prove all theoretical results whose proofs are not presented in the main text. We provide more details of the experiments in Appendix B. In Appendix C we extend our results to a setting where workers aim to maximize the expected value of some “utility” of their payments.

2. Setting and Notation

In the crowdsourcing setting that we consider, one or more workers perform a *task*, where a task consists of multiple *questions*. The questions are objective, by which we mean, each question has precisely one correct answer. Examples of objective questions include multiple-choice classification questions such as Figure 1, questions on transcribing text from audio or images, etc.

For any possible answer to any question, we define the worker’s *confidence about an answer* as the probability, according to her belief, of this answer being correct. In other words, one can assume that the worker has (in her mind) a probability distribution over all possible answers to a question,

and the confidence for an answer is the probability of that answer being correct. As a shorthand, we also define the *confidence about a question* as the confidence for the answer that the worker is most confident about for that question. We assume that the worker’s confidences for different questions are independent. Our goal is that for every question, the worker should be incentivized to skip if her confidence for that question is below a certain pre-defined threshold, otherwise select the answer that she is most confident about, and if asked, also indicate a correct (quantized) value of her confidence for the answer.

Specifically, we consider two settings:

- **Skip-based.** For each question, the worker can either choose to ‘skip’ the question or provide an answer (Figure 1b).
- **Confidence-based.** For each question, the worker can either ‘skip’ the question or provide an answer, and in the latter case, indicate her confidence for this answer as a number in $\{1, \dots, L\}$ (Figure 1c). We term this indicated confidence as the ‘confidence-level’. Here, L represents the highest confidence-level, and ‘skip’ is considered to be a confidence-level of 0.¹

One can see from the aforementioned definition that the confidence-based setting is a generalization of the skip-based setting (the skip-based setting corresponds to $L = 1$). The goal is to ensure that for a given set of intervals that partition $[0, 1]$, for every question the worker is incentivized to indicate ‘skip’ or choose the appropriate confidence-level when her confidence for that question falls in the corresponding interval. The choice of these intervals will be defined subsequently in the skip-based and confidence-based sections (Section 3 and Section 4) respectively.

Let N denote the total number of questions in the task. Among these questions, we assume the existence of some “gold standard” questions, that is, a set of questions whose answers are known to the requester. Let G ($1 \leq G \leq N$) denote the number of gold standard questions. The G gold standard questions are assumed to be distributed uniformly at random in the pool of N questions (of course, the worker does not know which G of the N questions form the gold standard). The payment to a worker for a task is computed after receiving her responses to all the questions in the task. The payment is based on the worker’s performance on the gold standard questions. Since the payment is based on known answers, the payments to different workers do not depend on each other, thereby allowing us to consider the presence of only one worker without any loss in generality.

We will employ the following standard notation. For any positive integer K , the set $\{1, \dots, K\}$ is denoted by $[K]$. The indicator function is denoted by $\mathbf{1}$, i.e., $\mathbf{1}\{z\} = 1$ if z is true, and 0 otherwise.

Let x_1, \dots, x_G denote the evaluations of the answers that the worker gives to the G gold standard questions, and let f denote the scoring rule, i.e., a function that determines the payment to the worker based on these evaluations x_1, \dots, x_G .

In the skip-based setting, $x_i \in \{-1, 0, +1\}$ for all $i \in [G]$. Here, “0” denotes that the worker skipped the question, “−1” denotes that the worker attempted to answer the question and that answer was incorrect, and “+1” denotes that the worker attempted to answer the question and that answer was correct. The payment function is $f : \{-1, 0, +1\}^G \rightarrow \mathbb{R}$.

In the confidence-based setting, $x_i \in \{-L, \dots, +L\}$ for all $i \in [G]$. Here, we set $x_i = 0$ if the worker skipped the question, and for $l \in \{1, \dots, L\}$, we set $x_i = l$ if the question was answered

1. When the task is presented to the workers, the word ‘skip’ or the numbers $\{1, \dots, L\}$ are replaced by more comprehensible phrases such as “I don’t know”, “moderately sure”, “absolutely sure”, etc.

correctly with confidence l and $x_i = -l$ if the question was answered incorrectly with confidence l . The function $f : \{-L, \dots, +L\}^G \rightarrow \mathbb{R}$ specifies the payment to be made to the worker.

The payment is further associated to two parameters, μ_{\max} and μ_{\min} . The parameter μ_{\max} denotes the *budget*, i.e., the maximum amount that is paid to any individual worker for this task:

$$\max_{x_1, \dots, x_G} f(x_1, \dots, x_G) = \mu_{\max}.$$

The amount μ_{\max} is thus the amount of compensation paid to a perfect worker for her work. Further, one may often also have the requirement of paying a certain minimum amount to any worker. The parameter μ_{\min} ($\leq \mu_{\max}$) denotes this minimum payment: the payment function must also satisfy

$$\min_{x_1, \dots, x_G} f(x_1, \dots, x_G) \geq \mu_{\min}.$$

For instance, crowdsourcing platforms today allow payments to workers, but do not allow imposing penalties: this condition gives $\mu_{\min} = 0$.

We assume that the worker attempts to maximize her overall expected payment. In what follows, the expression ‘the worker’s expected payment’ will refer to the expected payment from the worker’s point of view, and the expectation will be taken with respect to the worker’s confidences about her answers and the uniformly random choice of the G gold standard questions among the N questions in the task. A payment function f is called *incentive compatible* if the expected payment of the worker under this payment function is *strictly* maximized when the worker answers in the manner desired.² The specific requirements of the skip-based and the confidence-based settings are discussed subsequently in their respective sections to follow. In the remainder of this section, we formally define the concepts of the worker’s expected payment and incentive compatibility; the reader interested in understanding the paper at a higher level may skip directly to the next section without loss in continuity.

Let Ω denote the set of options for each question. We assume that Ω is a finite set, for instance, the set $\{\text{Yes}, \text{No}\}$ for a task with binary-choice questions, or the set of all strings of at most a certain length for a task with textual responses. Let $Q \in [0, 1]^{|\Omega| \times N}$ denote the beliefs of a worker for the N questions asked. Specifically, for any question $i \in [N]$ and any option $\omega \in \Omega$, let $Q_{\omega, i}$ represent the probability, according to the worker’s belief, that option ω is the correct answer to question i . Then from the law of total probability, any valid Q must have $\sum_{\omega \in \Omega} Q_{\omega, i} = 1$ for every $i \in [N]$. The value of Q is unknown to the mechanism.

Let us first define the notion of the expected payment (from the worker’s point of view) for any given response of the worker to the questions. For any question $i \in [N]$, suppose the worker indicates the confidence-level $\xi_i \in \{0, \dots, L\}$. For every question $i \in [N]$ such that $\xi_i \neq 0$, let $\omega_i \in \Omega$ denote the option selected by the worker; whenever $\xi_i = 0$, indicating a skip, we let ω_i take any arbitrary value in Ω . Furthermore, let $p_i = Q_{\omega_i, i}$ denote the probability, according to the worker’s belief, that the chosen option ω_i is the correct answer to question i . For notational purposes, we also define a vector $E = (\epsilon_1, \dots, \epsilon_G) \in \{-1, 1\}^G$. Then for the given responses, for the worker beliefs Q , and under payment mechanism f , the worker’s expected payment $\Gamma_{Q, f} : (\{0, \dots, L\} \times \Omega)^N \rightarrow \mathbb{R}$

2. Such a notion of incentive compatibility is often called “strict incentive compatibility”; we drop the prefix term “strict” for brevity.

is given by the expression:

$$\begin{aligned} & \Gamma_{Q,f}(\xi_1, \omega_1, \dots, \xi_N, \omega_N) \\ &= \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1, \dots, j_G) \\ \subseteq \{1, \dots, N\}}} \sum_{E \in \{-1, 1\}^G} \left(f(\epsilon_1 \xi_{j_1}, \dots, \epsilon_G \xi_{j_G}) \prod_{i=1}^G (p_{j_i})^{\frac{1+\epsilon_i}{2}} (1-p_{j_i})^{\frac{1-\epsilon_i}{2}} \right). \end{aligned} \quad (1)$$

In the expression (1), the outermost summation corresponds to the expectation with respect to the randomness arising from the unknown positions of the gold standard questions. The inner summation corresponds to the expectation with respect to the worker’s beliefs about the correctness of her responses. Note that the right hand side of (1) implicitly depends on $(\omega_1, \dots, \omega_N)$ through the values (p_1, \dots, p_N) . Also note that for every question i such that $\xi_i = 0$, the right hand side of (1) does not depend on the values of ω_i and p_i ; this is because the choice $\xi_i = 0$ of skipping question i implies that the worker did not select any particular option.

We will now use the the definition of the expected payment of the worker to define the notion of incentive compatibility. To this end, for any valid probabilities Q , let $\mathcal{A}(Q) \subseteq (\{0, \dots, L\} \times \Omega)^N$ denote an associated set of “desired” responses. By this we mean that every $a \in (\{0, \dots, L\} \times \Omega)^N$ represents a possible response to the set of N questions, and the goal is to incentivize the worker to provide any one response in the set $\mathcal{A}(Q)$. Then a mechanism f is termed incentive compatible if

$$\Gamma_{Q,f}(a) > \Gamma_{Q,f}(a') \quad \text{for every } a \in \mathcal{A}(Q), \text{ every } a' \notin \mathcal{A}(Q), \text{ and every valid } Q.$$

The goal is to design mechanisms that are incentive compatible, that is, incentivize the workers to respond in a certain manner. The specific choice of “desired responses” for the skip-based and the confidence-based settings are discussed subsequently in their respective sections. We begin with the skip-based setting.

3. Skip-based Setting

In this section, we consider the setting where for every question, the worker can choose to either answer the question or to skip it; no additional information is asked from the worker. See Figure 1b for an illustration.

3.1 Setting

Let $T \in (0, 1)$ be a predefined value. The goal is to design payment mechanisms that incentivize the worker to skip the questions for which her confidence is lower than T , and answer those for which her confidence is higher than T .³ Moreover, for the questions that she attempts to answer, she must be incentivized to select the answer that she believes is most likely to be correct. The value of T is chosen a priori based on factors such as budget constraints, the targeted quality of labels, and/or the choice of the algorithm used to subsequently aggregate the responses of multiple workers. In this paper, we assume that the value of the threshold T is specified to us.

Now the space of all possible mechanisms for this problem may be rather wide. Thus in order to narrow down our search, we impose the following additional simple and natural requirement:

3. In the event that the confidence about a question is exactly equal to T , the worker may choose to answer or skip.

Axiom 1 (No-free-lunch Axiom) *If all the answers attempted by the worker in the gold standard are wrong, then the payment is the minimum possible. More formally, $f(x_1, \dots, x_G) = \mu_{\min}$ for every evaluation (x_1, \dots, x_G) such that $0 < \sum_{i=1}^G \mathbf{1}\{x_i \neq 0\} = \sum_{i=1}^G \mathbf{1}\{x_i = -1\}$.*

One may expect a payment mechanism to impose the restriction of minimum payment to spammers who answer randomly. For instance, in a task with binary-choice questions, a spammer is expected to have 50% of the attempted answers incorrect; one may thus wish to set a the minimum possible payment if 50% or more of the attempted answers were incorrect. The no-free-lunch axiom which we impose is however a *significantly weaker condition*, mandating minimum payment if *all* attempted answers are incorrect.

3.2 Payment Mechanism

We now present our proposed payment mechanism in Algorithm 1.

Algorithm 1: Incentive mechanism for skip-based setting

- Inputs:
 - ▶ Threshold T
 - ▶ Budget parameters μ_{\max} and μ_{\min}
 - ▶ Evaluations $(x_1, \dots, x_G) \in \{-1, 0, +1\}^G$ of the worker’s answers to the G gold standard questions
- Set $\alpha_{-1} = 0$, $\alpha_0 = 1$, $\alpha_{+1} = \frac{1}{T}$
- The payment is

$$f(x_1, \dots, x_G) = \kappa \prod_{i=1}^G \alpha_{x_i} + \mu_{\min},$$

where $\kappa = (\mu_{\max} - \mu_{\min})T^G$.

The proposed mechanism has a *multiplicative* form: each answer in the gold standard is given a score based on whether it was correct (score = $\frac{1}{T}$), incorrect (score = 0) or skipped (score = 1), and the final payment is simply a product of these scores (scaled and shifted by constants). The mechanism is easy to describe to workers: For instance, if $T = \frac{1}{2}$, $G = 3$, $\mu_{\max} = 80$ cents and $\mu_{\min} = 0$ cents, then the description reads:

“The reward starts at 10 cents. For every correct answer in the 3 gold standard questions, the reward will double. However, if any of these questions are answered incorrectly, then the reward will become zero. So please use the ‘I’m not sure’ option wisely.”

Observe how this payment rule is similar to the popular ‘double or nothing’ paradigm (Double or Nothing, 2014).

The algorithm makes a minimum payment if *one or more* attempted answers in the gold standard are wrong. Note that this property is significantly stronger than the no-free-lunch axiom which we originally required, where we wanted a minimum payment only when *all* attempted answers were wrong. Surprisingly, as we prove shortly, Algorithm 1 is the only incentive-compatible mechanism that satisfies no-free-lunch.

The following theorem shows that this mechanism indeed incentivizes a worker to skip the questions for which her confidence is below T , while answering those for which her confidence is greater than T . In the latter case, the worker is incentivized to select the answer which she thinks is most likely to be correct.

Theorem 2 *The mechanism of Algorithm 1 is incentive-compatible and satisfies the no-free-lunch axiom.*

In the remainder of this subsection, we present the proof of Theorem 2. The reader may go directly to subsection 3.3 without loss in continuity.

Proof of Theorem 2. The proposed payment mechanism satisfies no-free-lunch since the payment is μ_{\min} when there are one or more wrong answers in the gold standard. It remains to show that the mechanism is incentive compatible. To this end, observe that the property of incentive-compatibility does not change upon any shift of the mechanism by a constant value or any scaling by a positive constant value. As a result, for the purposes of this proof, we can assume without loss of generality that $\mu_{\min} = 0$.

We will first assume that, for every question that the worker does not skip, she selects the answer which she believes is most likely to be correct. Under this assumption we will show that the worker is incentivized to skip the questions for which her confidence is smaller than T and attempt if it is greater than T . Finally, we will show that the mechanism indeed incentivizes the worker to select the answer which she believes is most likely to be correct for the questions that she doesn't skip. In what follows, we will employ the notation $\kappa = \mu_{\max} T^G$.

Let us first consider the case when $G = N$. Let p_1, \dots, p_N be the confidences of the worker for questions $1, \dots, N$ respectively. Further, let $p_{(1)} \geq \dots \geq p_{(m)} > T > p_{(m+1)} \geq \dots \geq p_{(N)}$ be the ordered permutation of these confidences (for some number m). Let $\{(1), \dots, (N)\}$ denote the corresponding permutation of the N questions. If the mechanism is incentive compatible, then the expected payment received by this worker should be maximized when the worker answers questions $(1), \dots, (m)$ and skips the rest. Under the mechanism proposed in Algorithm 1, this action fetches the worker an expected payment of

$$\kappa \frac{p_{(1)}}{T} \dots \frac{p_{(m)}}{T}.$$

Alternatively, if the worker answers the questions $\{i_1, \dots, i_\beta\}$, with $p_{i_1} > \dots > p_{i_\nu} > T > p_{i_{\nu+1}} > \dots > p_{i_\beta}$ for some value ν , then the expected payment is

$$p_{i_1} \dots p_{i_\beta} \frac{\kappa}{T^\beta} = \kappa \frac{p_{i_1}}{T} \dots \frac{p_{i_\beta}}{T} \tag{2}$$

$$\leq \kappa \frac{p_{i_1}}{T} \dots \frac{p_{i_\nu}}{T} \tag{3}$$

$$\leq \kappa \frac{p_{(1)}}{T} \dots \frac{p_{(m)}}{T}, \tag{4}$$

where inequality (3) holds because $\frac{p_{i_j}}{T} \leq 1 \ \forall j > \nu$ and holds with equality only when $\beta = \nu$. Inequality (4) is a result of $\frac{p_{(j)}}{T} \geq 1 \ \forall j \leq m$ and holds with equality only when $\nu = m$. It follows that the expected payment is (strictly) maximized when $i_1 = (1), \dots, i_\beta = (m)$ as required.

The case of $G < N$ is a direct consequence of the result for $G = N$, as follows. When $G < N$, from a worker's point of view, the set of G questions is distributed uniformly at random in

the superset of N questions. However, for every set of G questions, the relations (2), (3), (4) and their associated equality/strict-inequality conditions hold. The expected payment is thus (strictly) maximized when the worker answers the questions for which her confidence is greater than T and skips those for which her confidence is smaller than T .

One can see that for every question that the worker chooses to answer, the expected payment increases with an increase in her confidence. Thus, the worker is incentivized to select the answer that she thinks is most probably correct.

Finally, since $\kappa = \mu_{\max} T^G > 0$ and $T \in (0, 1)$, the payment is always non-negative and satisfies the μ_{\max} -budget constraint.

3.3 Uniqueness of this Mechanism

While we started out with a very weak condition of no-free-lunch of that requires a minimum payment when *all* attempted answers are wrong, the mechanism proposed in Algorithm 1 is significantly more strict and pays the minimum amount when *any* of the attempted answers is wrong. A natural question that arises is: can we design an alternative mechanism satisfying incentive compatibility and no-free-lunch that operates somewhere in between? The following theorem answers this question in the negative.

Theorem 3 *The mechanism of Algorithm 1 is the only incentive-compatible mechanism that satisfies the no-free-lunch axiom.*

Theorem 3 gives a strong result despite imposing very weak requirements. To see this, recall our earlier discussion on deterring spammers, that is, making a low payment to workers who answer randomly. For instance, when the task comprises binary-choice questions, one may wish to design mechanisms which make the minimum possible payment when the responses to 50% or more of the questions in the gold standard are incorrect. The no-free-lunch axiom is a much weaker requirement, and the only mechanism that can satisfy this requirement is the mechanism of Algorithm 1.

The proof of Theorem 3 is based on the following key lemma, establishing a condition that any incentive-compatible mechanism must necessarily satisfy. Note that this lemma does *not* require the no-free-lunch axiom.

Lemma 4 *Any incentive-compatible mechanism f must satisfy, for every gold standard question $i \in \{1, \dots, G\}$ and every $(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_G) \in \{-1, 0, 1\}^{G-1}$,*

$$\begin{aligned} T f(y_1, \dots, y_{i-1}, 1, y_{i+1}, \dots, y_G) + (1 - T) f(y_1, \dots, y_{i-1}, -1, y_{i+1}, \dots, y_G) \\ = f(y_1, \dots, y_{i-1}, 0, y_{i+1}, \dots, y_G) . \end{aligned}$$

The proof of Lemma 4 is provided in Appendix A.1. Using this lemma, we will now prove Theorem 3. The reader interested in further results and not the proof may feel free to jump to Subsection 3.4 without any loss in continuity.

Proof of Theorem 3. The property of incentive-compatibility does not change upon any shift of the mechanism by a constant value or any scaling by a positive constant value. As a result, for the purposes of this proof, we can assume without loss of generality that $\mu_{\min} = 0$.

We will first prove that any incentive-compatible mechanism satisfying the no-free-lunch axiom must make a zero payment if one or more answers in the gold standard are incorrect. The proof proceeds by induction on the number of skipped questions S in the gold standard. Let us assume for now that in the G questions in the gold standard, the first question is answered incorrectly, the next $(G - 1 - S)$ questions are answered by the worker and have arbitrary evaluations, and the remaining S questions are skipped. The proof proceeds by an induction on S . Suppose $S = G - 1$. In this case, the only attempted question is the first question and the answer provided by the worker to this question is incorrect. The no-free-lunch axiom necessitates a zero payment in this case, thus satisfying the base case of our induction hypothesis. Now we prove the hypothesis for some S under the assumption of it being true when the number of questions skipped in the gold standard is $(S + 1)$ or more. From Lemma 4 (with $i = G - S - 1$) we have

$$\begin{aligned} T f(-1, y_2, \dots, y_{G-S-2}, 1, 0, \dots, 0) + (1 - T) f(-1, y_2, \dots, y_{G-S-2}, -1, 0, \dots, 0) \\ = f(-1, y_2, \dots, y_{G-S-2}, 0, 0, \dots, 0) \\ = 0, \end{aligned}$$

where the final equation is a consequence of our induction hypothesis: The induction hypothesis is applicable since $f(-1, y_2, \dots, y_{G-S-2}, 0, 0, \dots, 0)$ corresponds to the case when the last $(S + 1)$ questions are skipped and the first question is answered incorrectly. Now, since the payment f must be non-negative and since $T \in (0, 1)$, it must be that

$$f(-1, y_2, \dots, y_{G-S-2}, 1, 0, \dots, 0) = 0,$$

and

$$f(-1, y_2, \dots, y_{G-S-2}, -1, 0, \dots, 0) = 0.$$

This completes the proof of our induction hypothesis. Furthermore, each of the arguments above hold for any permutation of the G questions, thus proving the necessity of zero payment when any one or more answers are incorrect.

We will now prove that when no answers in the gold standard are incorrect, the payment must be of the form described in Algorithm 1. Let κ be the payment when all G questions in the gold standard are skipped. Let C be the number of questions answered correctly in the gold standard. Since there are no incorrect answers, it follows that the remaining $(G - C)$ questions are skipped. Let us assume for now that the first C questions are answered correctly and the remaining $(G - C)$ questions are skipped. We repeatedly apply Lemma 4, and the fact that the payment must be zero when one or more answers are wrong,

$$\begin{aligned} f(\underbrace{1, \dots, 1}_{C-1}, 1, \underbrace{0, \dots, 0}_{G-C}) &= \frac{1}{T} f(\underbrace{1, \dots, 1}_{C-1}, 0, \underbrace{0, \dots, 0}_{G-C}) - \frac{1-T}{T} f(\underbrace{1, \dots, 1}_{C-1}, -1, \underbrace{0, \dots, 0}_{G-C}) \\ &= \frac{1}{T} f(\underbrace{1, \dots, 1}_{C-1}, 0, \underbrace{0, \dots, 0}_{G-C}), \end{aligned}$$

and so on to obtain

$$\begin{aligned} f(\underbrace{1, \dots, 1}_{C-1}, 1, \underbrace{0, \dots, 0}_{G-C}) &= \frac{1}{T^C} f(\underbrace{0, \dots, 0}_G) \\ &= \frac{1}{T^C} \kappa. \end{aligned}$$

In order to abide by the budget, we must have the maximum payment as $\mu_{\max} = \kappa \frac{1}{T^G}$. It follows that $\kappa = \mu_{\max} T^G$. Finally, the arguments above hold for any permutation of the G questions, thus proving the uniqueness of the mechanism of Algorithm 1.

3.4 Optimality against Spamming Behavior

As discussed earlier, crowdsourcing tasks, especially those with multiple choice questions, often encounter spammers who answer randomly without heed to the question being asked. For instance, under a binary-choice setup, a spammer will choose one of the two options uniformly at random for every question. A highly desirable objective in crowdsourcing settings is to deter spammers. To this end, one may wish to impose a condition of making the minimum possible payment when the responses to 50% or more of the attempted questions in the gold standard are incorrect. A second desirable metric could be to minimize the expenditure on a worker who simply skips all questions. While the aforementioned requirements were deterministic functions of the worker's responses, one may alternatively wish to impose requirements that depend on the distribution of the worker's answering process. For instance, a third desirable feature would be to minimize the expected payment to a worker who answers all questions uniformly at random. We now show that interestingly, our unique multiplicative payment mechanism *simultaneously* satisfies all these requirements. The result is stated assuming a multiple-choice setup, but extends trivially to non-multiple-choice settings.

Theorem 5.A (Distributional) *Consider any value $A \in \{0, \dots, G\}$. Among all incentive-compatible mechanisms (that may or may not satisfy no-free-lunch), Algorithm 1 pays strictly the smallest amount to a worker who skips some A of the questions in the the gold standard, and chooses answers to the remaining $(G - A)$ questions uniformly at random.*

Theorem 5.B (Deterministic) *Consider any value $B \in (0, 1]$. Among all incentive-compatible mechanisms (that may or may not satisfy no-free-lunch), Algorithm 1 pays strictly the smallest amount to a worker who gives incorrect answers to a fraction B or more of the questions attempted in the gold standard.*

We see from this result that the multiplicative payment mechanism of Algorithm 1 thus possesses very useful properties geared to deter spammers, while ensuring that a good worker will be paid a high enough amount. To illustrate this point, let us compare the mechanism of Algorithm 1 with the popular additive class of payment mechanisms.

Example 1 *Consider the popular class of “additive” mechanisms, where the payments to a worker are added across the gold standard questions. This additive payment mechanism offers a reward of $\frac{\mu_{\max}}{G}$ for every correct answer in the gold standard, $\frac{\mu_{\max}T}{G}$ for every question skipped, and 0 for every incorrect answer. Importantly, the final payment to the worker is the sum of the rewards across the G gold standard questions. One can verify that this additive mechanism is incentive compatible. One can also see that that as guaranteed by our theory, this additive payment mechanism does not satisfy the no-free-lunch axiom.*

Suppose each question involves choosing from two options. Let us compute the payment that these two mechanisms make under a spamming behavior of choosing the answer randomly to each question. Given the 50% likelihood of each question being correct, one can compute that the additive

mechanism makes a payment of $\frac{\mu_{\max}}{2}$ in expectation. On the other hand, our mechanism pays an expected amount of only $\mu_{\max}2^{-G}$. The payment to spammers thus reduces exponentially with the number of gold standard questions under our mechanism, whereas it does not reduce at all in the additive mechanism.

Now, consider a different means of exploiting the mechanism(s) where the worker simply skips all questions. To this end, observe that if a worker skips all the questions then the additive payment mechanism will make a payment of $\mu_{\max}T$. On the other hand, the proposed payment mechanism of Algorithm 1 pays an exponentially smaller amount of $\mu_{\max}T^G$ (recall that $T < 1$).

We prove Theorem 5 in the rest of this subsection. The reader may feel free to jump directly to Section 4 without any loss in continuity.

Proof of Theorem 5. The property of incentive-compatibility does not change upon any shift of the mechanism by a constant value or any scaling by a positive constant value. As a result, for the purposes of this proof, we can assume without loss of generality that $\mu_{\min} = 0$.

Part A (Distributional). Let m denote the number of options in each question. One can verify that under the mechanism of Algorithm 1, a worker who skips A questions and answers the rest uniformly at random will get a payment of $\frac{\mu_{\max}T^A}{m^{G-A}}$ in expectation. This expression arises due to the fact that Algorithm 1 makes a zero payment if any of the attempted answers are incorrect, and a payment of $\mu_{\max}T^A$ if the worker skips A questions and answers the rest correctly. Under uniformly random answers, the probability of the latter event is $\frac{1}{m^{G-A}}$.

Now consider any other mechanism, and denote it as f' . Let us suppose without loss of generality that the worker attempts the first $(G - A)$ questions. Since the payment must be non-negative, a repeated application of Lemma 4 gives

$$\begin{aligned} f'(\underbrace{1, \dots, 1}_{G-A}, 0, \dots, 0) &\geq T f'(\underbrace{1, \dots, 1}_{G-A+1}, 0, \dots, 0) & (5) \\ &\vdots \\ &\geq T^A f'(1, \dots, 1) \\ &= T^A \mu_{\max}, & (6) \end{aligned}$$

where (6) is a result of the μ_{\max} -budget constraint. Since there is a $\frac{1}{m^{G-A}}$ chance of the $(G - A)$ attempted answers being correct, the expected payment under any other mechanism f' must be at least $\frac{\mu_{\max}T^A}{m^{G-A}}$.

We will now show that if any mechanism f' that makes an expected payment of $\frac{\mu_{\max}T^A}{m^{G-A}}$ to such a spammer, then the mechanism must be identical to Algorithm 1. We split the proof of this part into two cases, depending on the value of the parameter A .

Case I ($A < G$): In order to make an expected payment of $\frac{\mu_{\max}T^A}{m^{G-A}}$, the mechanism must achieve the bound (6) with equality, and furthermore, the mechanism must have zero payment if any of the $(G - A)$ attempted questions are answered incorrectly. In other words, the mechanism f' under consideration must satisfy

$$f'(y_1, \dots, y_{G-A}, 0, \dots, 0) = 0 \quad \forall (y_1, \dots, y_{G-A}) \in \{-1, 1\}^{G-A} \setminus \{1\}^{G-A}.$$

A repeated application of Lemma 4 then implies

$$f'(0, 0, \dots, -1) = 0. \quad (7)$$

Note that so far we considered the case when the worker attempts the first $(G - A)$ questions. The arguments above hold for any choice of the $(G - A)$ attempted questions, and consequently the results shown so far in this proof hold for all permutations of the arguments to f' . In particular, the mechanism f' must make a zero payment when any $(G - 1)$ questions in the gold standard are skipped and the remaining question is answered incorrectly. Another repeated application of Lemma 4 to this result gives

$$f'(y_1, \dots, y_G) = 0 \quad \forall (y_1, \dots, y_G) \in \{0, -1\}^G \setminus \{0\}^G.$$

This condition is precisely the no-free-lunch axiom, and in Theorem 3 we had shown that Algorithm 1 is the only incentive-compatible mechanism that satisfies this axiom. It follows that our mechanism, Algorithm 1 strictly minimizes the expected payment in the setting under consideration.

Case II ($A = G$): In order to achieve the bound (6) with equality, the mechanism f' must also achieve the bound (5) with equality. Noting that we have $A = G$ in this case, it follows that the mechanism f' must satisfy

$$f'(-1, 0, \dots, 0) = 0.$$

This condition is identical to (7) established for Case I earlier, and the rest of the argument now proceeds in a manner identical to the subsequent arguments in Case I.

Part B (Deterministic). Given our result of Theorem 3, the proof for the deterministic part is straightforward. Algorithm 1 makes a payment of zero when one or more of the answers to questions in the gold standard are incorrect. Consequently, for every value of parameter $B \in (0, 1]$, Algorithm 1 makes a zero payment when a fraction B or more of the attempted answers are incorrect. Any other mechanism doing so must satisfy the no-free-lunch axiom. In Theorem 3 we had shown that Algorithm 1 is the only incentive-compatible mechanism that satisfies this axiom. It follows that our mechanism, Algorithm 1, strictly minimizes the payment in the event under consideration.

4. Confidence-based Setting

In this section, we will discuss incentive mechanisms when the worker is asked to select from more than one confidence-level for every question (Figure 1c). In particular, for some $L \geq 1$, the worker is asked to indicate a confidence-level in the range $\{0, \dots, L\}$ for every answer. Level 0 is the ‘skip’ level, and level L denotes the highest confidence. Note that we do not solicit an answer if the worker indicates a confidence-level of 0 (skip), but the worker must provide an answer if she indicates a confidence-level of 1 or higher. This makes the case of having only a ‘skip’ as considered in Section 3 a special case of this setting, and corresponds to $L = 1$.

We generalize the requirement of no-free-lunch to the confidence-based setting as follows.

Axiom 6 (Generalized-no-free-lunch axiom) *If all the answers attempted by the worker in the gold standard are selected as the highest confidence-level (level L), and all of them turn out to be wrong, then the payment is μ_{\min} . More formally, we require the mechanism f to satisfy $f(x_1, \dots, x_G) = \mu_{\min}$ for every evaluation (x_1, \dots, x_G) that satisfies $0 < \sum_{i=1}^G \mathbf{1}\{x_i \neq 0\} = \sum_{i=1}^G \mathbf{1}\{x_i = -L\}$.*

In the confidence-based setting, we require specification of a set of thresholds $\{S_l, T_l\}_{l=1}^L$ that determine the confidence-levels that the workers should indicate. These thresholds are used to choose the payment mechanism in a principled manner. In particular, we will require specification of two reference points for each confidence level, and this specification generalizes the skip-based setting.

- The first set of thresholds specifies a comparison of any confidence level with the skipping option as a fixed reference. To this end, recall that in the skip-based setting, the threshold T specified when the worker should skip a question and when she should attempt to answer. This is generalized to the confidence-based setting where for every level $l \in [L]$, a fixed threshold S_l specifies the ‘strength’ of confidence-level l : If restricted to only the two options of skipping or selecting confidence-level l for any question, the worker should be incentivized to select confidence-level l if her confidence is higher than S_l and skip if her confidence is lower than S_l .
- The second set of thresholds specifies a comparison of any confidence level with its neighbors. If a worker decides to not skip a question, she must choose one of multiple confidence-levels. A set $\{T_l\}_{l=1}^L$ of thresholds specify the boundaries between different confidence-levels. In particular, when the confidence of the worker for a question lies in (T_{l-1}, T_{l+1}) , then the worker must be incentivized to indicate confidence-level $(l - 1)$ if her confidence is lower than T_l and to indicate confidence-level l if her confidence is higher than T_l . This includes selecting level L if her confidence is higher than T_L and selecting level 0 if her confidence is lower than T_1 .

We will call a payment mechanism as incentive-compatible if it satisfies the two requirements listed above, and also incentivizes the worker to select the answer that she believes is most likely to be correct for every question for which her confidence is higher than T_1 .

The problem setting inherently necessitates certain restrictions in the choice of the thresholds. Since we require the worker to choose a higher level when her confidence is higher, the thresholds must necessarily be monotonic and satisfy $0 < S_1 < S_2 < \dots < S_L < 1$ and $0 < T_1 < T_2 < \dots < T_L < 1$. Also observe that the definitions of S_1 and T_1 coincide, and hence $S_1 = T_1$. Additionally, we can show (Proposition 18 in Appendix A.5) that for incentive-compatible mechanisms to exist, it must be that $T_l > S_l \forall l \in \{2, \dots, L\}$. As a result, the thresholds must also satisfy $T_1 = S_1, T_2 > S_2, \dots, T_L > S_L$. These thresholds may be chosen based on various factors of the problem at hand, for example, on the post-processing algorithms, any statistics on the distribution of worker abilities, budget constraints, etc. In this paper, we will assume that these values are given to us.

4.1 Payment Mechanism

In this section, we present our proposed payment mechanism, and prove that it is guaranteed to satisfy our requirements. We begin with a description of the mechanism in Algorithm 2.

Algorithm 2: Incentive mechanism for the confidence-based setting

• Inputs:

- ▶ Thresholds S_1, \dots, S_L and T_1, \dots, T_L
- ▶ Budget parameters μ_{\max} and μ_{\min}
- ▶ Evaluations $(x_1, \dots, x_G) \in \{-L, \dots, +L\}^G$ of the worker's answers to the G gold standard questions

• Set $\alpha_{-L}, \dots, \alpha_L$ as

- ▶ $\alpha_L = \frac{1}{S_L}, \alpha_{-L} = 0$
- ▶ For $l \in \{L-1, \dots, 1\}$,

$$\alpha_l = \frac{(1 - S_l)T_{l+1}\alpha_{l+1} + (1 - S_l)(1 - T_{l+1})\alpha_{-(l+1)} - (1 - T_{l+1})}{T_{l+1} - S_l} \quad \text{and} \quad \alpha_{-l} = \frac{1 - S_l\alpha_l}{1 - S_l}$$

- ▶ $\alpha_0 = 1$

• The payment is

$$f(x_1, \dots, x_G) = \kappa \prod_{i=1}^G \alpha_{x_i} + \mu_{\min}$$

where $\kappa = (\mu_{\max} - \mu_{\min}) \left(\frac{1}{\alpha_L}\right)^G$.

The following theorem shows that this mechanism indeed incentivizes a worker to select answers and confidence-levels as desired.

Theorem 7 *The mechanism of Algorithm 2 is incentive-compatible and satisfies the generalized-no-free-lunch axiom.*

The proof of Theorem 7 follows in a manner similar to that of the proof of Theorem 2, and is provided in Appendix A.2.

Remark 8 *The mechanism of Algorithm 2 also ensures a condition stronger than the ‘boundary-based’ definition of the thresholds $\{T_l\}_{l \in [L]}$ given earlier. Under this mechanism, for every $l \in [L-1]$ the worker is incentivized to select confidence-level l (over all else) whenever her confidence lies in the interval (T_l, T_{l+1}) , select confidence-level 0 (over all else) whenever her confidence is lower than T_1 and select confidence-level L (over all else) whenever her confidence is higher than T_L .*

4.2 Uniqueness of this Mechanism

We prove that the mechanism of Algorithm 2 is unique, that is, no other incentive-compatible mechanism can satisfy the generalized-no-free-lunch axiom.

Theorem 9 *The payment mechanism of Algorithm 2 is the only incentive-compatible mechanism that satisfies the generalized-no-free-lunch axiom.*

The proof of Theorem 9 is provided in Appendix A.3. The proof is conceptually similar to that of Theorem 9 but involves resolving several additional complexities that arise due to elicitation from multiple confidence levels.

5. A Stronger No-free-lunch Axiom: Impossibility Results

Recall that the no-free-lunch axiom under the skip-based mechanism of Section 3 requires the payment to be the minimum possible if all attempted answers in the gold standard are incorrect. However, a worker who skips all the questions may still receive a payment. The generalization under the confidence-based mechanism of Section 4 requires the payment to be the minimum possible if all attempted answers in the gold standard were selected with the highest confidence-level and were incorrect. However, a worker who marked all questions with a lower confidence level may be paid even if her answers to all the questions in the gold standard turn out to be incorrect. One may thus wish to impose a stronger requirement instead, where the minimum payment is made to workers who make no useful contribution. This is the primary focus of this section.

Consider the skip-based setting. Define the following axiom which is slightly stronger than the no-free-lunch axiom defined previously.

Strong-no-free-lunch: If none of the answers in the gold standard are correct, then the payment is μ_{\min} . More formally, strong-no-free-lunch imposes the condition $f(x_1, \dots, x_G) = \mu_{\min}$ for every evaluation (x_1, \dots, x_G) that satisfies $\sum_{i=1}^G \mathbf{1}\{x_i > 0\} = 0$.

The strong-no-free-lunch axiom is only slightly stronger than the no-free-lunch axiom proposed in Section 3 for the skip-based setting. The strong-no-free-lunch axiom can equivalently be written as imposing requiring the payment to be the minimum possible for every evaluation that satisfies $\sum_{i=1}^G \mathbf{1}\{x_i \neq 0\} = \sum_{i=1}^G \mathbf{1}\{x_i = -1\}$. From this interpretation, one can see that to the set of events necessitating the minimum payment under the no-free-lunch axiom, the strong-no-free-lunch axiom adds only one extra event, the event of the worker skipping all questions. Unfortunately, it turns out that this minimal strengthening of the requirements is associated to impossibility results.

In this section we show that no mechanism satisfying the strong-no-free-lunch axiom can be incentive compatible in general. The only exception is the case when (a) all questions are in the gold standard ($G = N$), and (b) it is guaranteed that the worker has a confidence greater than T for at least one of the N questions. These conditions are, however, impractical for the crowdsourcing setup under consideration in this paper. We will first prove the impossibility results under the strong-no-free-lunch axiom. For the sake of completeness (and also to satisfy mathematical curiosity), we will then provide a (unique) mechanism that is incentive-compatible and satisfies the strong-no-free-lunch axiom for the skip-based setting under the two conditions listed above. The proofs of each of the claims made in this section are provided in Appendix A.6.

Let us continue to discuss the skip-based setting. In this section, we will call any worker whose confidences for all of the N questions is lower than T as an *unknowledgeable worker*, and call the worker a *knowledgeable worker* otherwise.

Proposition 10 *No payment mechanism satisfying the strong-no-free-lunch axiom can incentivize an unknowledgeable worker to skip all questions. As a result, no mechanism satisfying the strong-no-free-lunch axiom can be incentive-compatible.*

The proof of this proposition, and that of all other theoretical claims made in this section, are presented in Appendix A.6.

The impossibility result of Proposition 10 relies on trying to incentivize an unknowledgeable worker to act as desired. Since no mechanism can be incentive compatible for unknowledgeable workers, we will now consider only workers who are knowledgeable. The following proposition shows that the strong-no-free-lunch axiom is too strong even for this relaxed setting.

Proposition 11 *When $G < N$, there exists no mechanism that is incentive-compatible for knowledgeable workers and satisfies the strong-no-free-lunch axiom.*

Given this impossibility result for $G < N$, we are left with $G = N$ which means that the true answers to all the questions are known a priori. This condition is clearly not applicable to a crowdsourcing setup; nevertheless, it is mathematically interesting and may be applicable to other scenarios such as testing and elicitation of beliefs about future events.

Proposition 12 below presents a mechanism for this case and proves its uniqueness. We previously saw that an unknowledgeable worker cannot be incentivized to skip all the questions (even when $G = N$). Thus, in our payment mechanism, we do the next best thing: Incentivize the unknowledgeable worker to answer only one question, that which she is most confident about, while incentivizing the knowledgeable worker to answer questions for which her confidence is greater than T and skip those for which her confidence is smaller than T .

Proposition 12 *Let C be the number of correct answers and W be the number of wrong answers (in the gold standard). Let the payment be μ_{\min} if $W > 0$ or $C = 0$, and be $(\mu_{\max} - \mu_{\min})T^{G-C} + \mu_{\min}$ otherwise. Under this mechanism, when $G = N$, an unknowledgeable worker is incentivized to answer only one question, that for which her confidence is the maximum, and a knowledgeable worker is incentivized to answer the questions for which her confidence is greater than T and skip those for which her confidence is smaller than T . Furthermore, when $G = N$, this mechanism is the one and only mechanism that obeys the strong-no-free-lunch axiom and is incentive-compatible for knowledgeable workers.*

The following proposition shows that the strong-no-free-lunch axiom leads to negative results in the confidence-based setting ($L > 1$) as well. The strong-no-free-lunch axiom is still defined as in the beginning of Section 5, i.e., the payment is zero if none of the answers are correct.

Proposition 13 *When $L > 1$, for any values of N and $G (\leq N)$, it is impossible for any mechanism to satisfy the strong-no-free-lunch axiom and be incentive-compatible even when the worker is knowledgeable.*

6. Simulations and Experiments

In this section, we present synthetic simulations and real-world experiments to evaluate the effects of our setting and our mechanism on the final label quality.

6.1 Synthetic Simulations

We employ synthetic simulations to understand the effects of various distributions of the confidences and labeling errors. We consider binary-choice questions in this set of simulations. Whenever a worker answers a question, her confidence for the correct answer is drawn from a distribution \mathcal{P} independent of all else. We investigate the effects of the following five choices of the distribution \mathcal{P} :

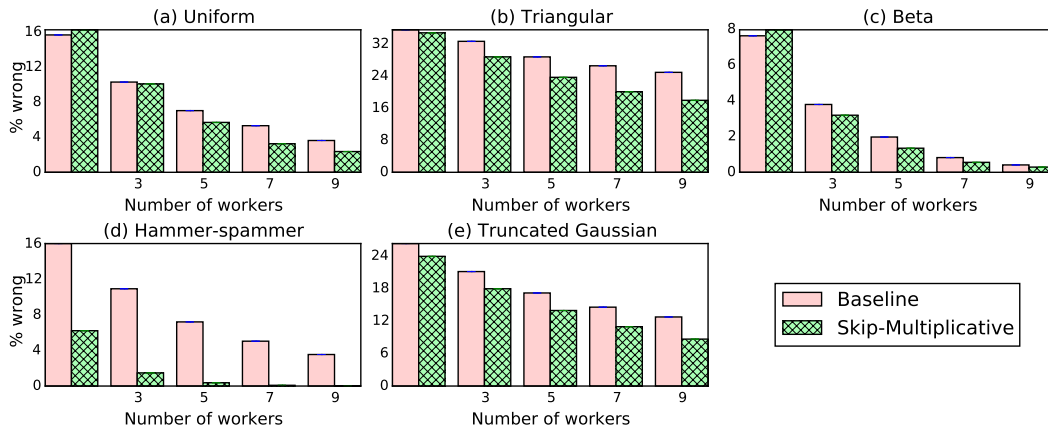


Figure 2: Error under different interfaces for synthetic simulations of five distributions of the workers’ error probabilities.

- The uniform distribution on the support $[0.5, 1]$.
- A triangular distribution with lower end-point 0.2, upper end-point 1 and a mode of 0.6.
- A beta distribution with parameter values $\alpha = 5$ and $\beta = 1$.
- The hammer-spammer distribution (Karger et al., 2011): uniform on the discrete set $\{0.5, 1\}$.
- A truncated Gaussian distribution: a truncation of $\mathcal{N}(0.75, 0.5)$ to the interval $[0, 1]$.

We compare (a) the setting where workers attempt every question, with (b) the setting where workers skip questions for which their confidence is below a certain threshold T . In this set of simulations, we set $T = 0.75$. In either setting, we aggregate the labels obtained from the workers for each question via a majority vote on the two classes. Ties are broken by choosing one of the two options uniformly at random.

Figure 2 depicts the results from these simulations. Each bar represents the fraction of questions that are labeled incorrectly, and is an average across 50,000 trials. (The standard error of the mean is too small to be visible.) We see that the skip-based setting consistently outperforms the conventional setting, and the gains obtained are moderate to high depending on the underlying distribution of the workers’ errors. In particular, the gains are quite striking under the hammer-spammer model: this result is not surprising since the mechanism (ideally) screens the spammers out and leaves only the hammers who answer perfectly.

The setup of the simulations described above assumes that the workers confidences equal the true error probabilities. In practice, however, the workers may have incorrect beliefs. The setup also assumes that ties are broken randomly; however in practice, ties may be broken in a more systematic manner by eliciting additional labels for only these hard questions. We now present a second set of simulations that mitigates these biases. In particular, when a worker has a confidence of p , the actual probability of error is assumed to be drawn from a Gaussian distribution with mean p and standard deviation 0.1, truncated to $[0, 1]$. In addition, when evaluating the performance of the

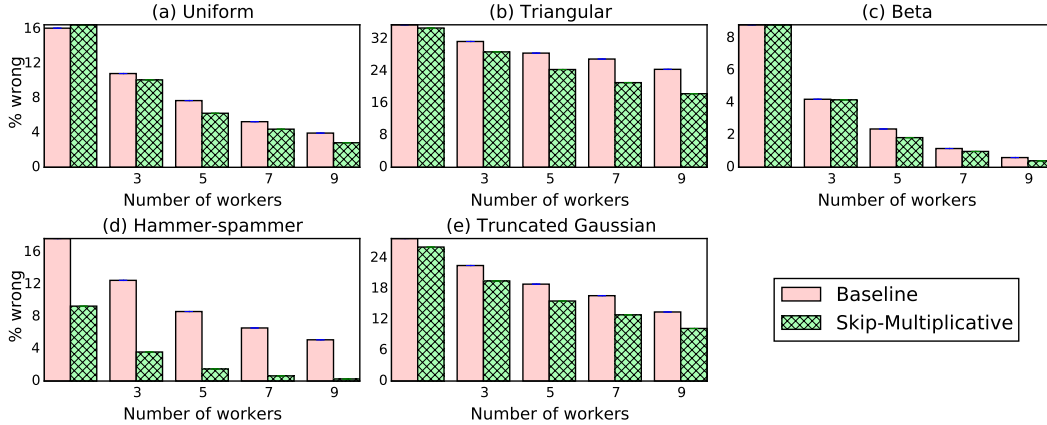


Figure 3: Errors under a model that is a perturbation of the first experiment, where the worker’s confidence is a noisy version of the true error probability and where ties are considered different from random decisions.

majority voting procedure, we consider a tie as having an error of 0.4. Figure 3 depicts the results of these simulations. We observe that the results from these simulations are very similar to those obtained in the earlier simulation setup of Figure 2.

6.2 Experiments on Amazon Mechanical Turk

We conducted preliminary experiments on the Amazon Mechanical Turk commercial crowdsourcing platform (mturk.com) to evaluate our proposed scheme in real-world scenarios. The complete data, including the interface presented to the workers in each of the tasks, the results obtained from the workers, and the ground truth solutions, are available on the website of the first author.

6.2.1 GOAL

Before delving into details, we first note certain caveats relating to such a study of mechanism design on crowdsourcing platforms. When a worker encounters a mechanism for only a small amount of time (a handful of tasks in typical research experiments) and for a small amount of money (at most a few dollars in typical crowdsourcing tasks), we cannot expect the worker to completely understand the mechanism and act precisely as required. For instance, we wouldn’t expect our experimental results to change significantly even upon moderate modifications in the promised amounts, and furthermore, we do expect the outcomes to be noisy. Incentive compatibility kicks in when the worker encounters a mechanism across a longer term, for example, when a proposed mechanism is adopted as a standard for a platform, or when higher amounts are involved. This is when we would expect workers or others (e.g., bloggers or researchers) to design strategies that can game the mechanism. The theoretical guarantee of incentive compatibility then prevents such gaming in the long run.

We thus regard these experiments as preliminary. Our intentions towards this experimental exercise were (a) to evaluate the potential of our algorithms to work in practice, (b) to investigate

the effect of the proposed algorithms on the net error in the collected labeled data, and (c) to identify if there is any major issue of dissatisfaction among the workers.

6.2.2 EXPERIMENTAL SETUP

We conducted our experiments on the “Amazon Mechanical Turk” commercial crowdsourcing platform (`mturk.com`). On this platform, individuals or businesses (called ‘requesters’) can post tasks, and any individual (called a ‘worker’) may complete the task over the Internet in exchange for a pre-specified payment. The payment may comprise of two parts: a fixed component which is identical for all workers performing that task, and a ‘bonus’ which may be different for different workers and is paid at the discretion of the requester.

We designed nine experiments (tasks) ranging from image annotation to text and speech recognition. The individual experiments are described in more detail in Appendix B. All experiments involved objective questions, and the responses elicited were multiple choice in five of the experiments and free-form text in the rest. For each experiment, we tested three settings: (i) the baseline conventional setting (Figure 1a) with a mechanism of paying a fixed amount per correct answer, (ii) our skip-based setting (Figure 1b) with our multiplicative mechanism, and (iii) our confidence-based setting (Figure 1c) with our confidence-based mechanism. For each mechanism in each experiment, we specified the requirement of 35 workers independently performing the task. This amounts to a total of 945 worker-tasks (315 worker-tasks for each mechanism). We also set the following constraints for a worker to attempt our tasks: the worker must have completed at least 100 tasks previously, and must have a history of having at least 95% of her prior work approved by the respective requesters. In each experiment, we offered a certain small fixed payment (in order to attract the workers in the first place) and executed the variable part of our mechanisms via a bonus payment.

6.2.3 RESULTS: RAW DATA

Figure 4 plots, for the baseline, skip-based and confidence-based mechanisms for all nine experiments, the (i) fraction of questions that were answered incorrectly, (ii) fraction of questions that were answered incorrectly among those that were attempted, (iii) the average payment to a worker (in cents), and (iv) break up of the answers in terms of the fraction of answers in each confidence level. The payment for various tasks plotted in Figure 4 is computed as the average of the payments across 100 (random) selections of the gold standard questions, in order to prevent any distortion of the results due to the randomness in the choice of the gold standard questions.

The figure shows that the amount of errors among the attempted questions is much lower in the skip and the confidence-based settings than the baseline setting. Also observe that in the confidence-based setting, as expected, the answers selected with higher confidence-levels are more correct. The total money spent under each of these settings is similar, with the skip and the confidence-based settings faring better in most cases. We also elicited feedback from the workers, in which we received several positive comments (and no negative comments). Examples of comments that we received: “I was wondering if it would possible to increase the maximum number of HITs I may complete for you. As I said before, they were fun to complete. I think I did a good job completing them, and it would be great to complete some more for you.”; “I am eagerly waiting for your bonus.”; “Enjoyable. Thanks.”

6.2.4 RESULTS: AGGREGATED DATA

We saw in the previous section that under the skip-based setting, the amount of error among the attempted questions was significantly lower than the amount of error in the baseline setting. However, the skip-based setting was also associated, by design, to lesser amount of data by virtue of questions being skipped by the workers. A natural question that arises is how the baseline and the skip-based mechanisms will compare in terms of the final data quality, i.e., the amount of error once data from multiple workers is aggregated.

To this end, we considered the five experiments that consisted of multiple-choice questions. We let a parameter `num_workers` take values in $\{3, 5, 7, 9, 11\}$. For each of the five experiments and for each of the five values of `num_workers`, we perform the following actions 1,000 times: for each question, we choose `num_workers` workers and perform a majority vote on their responses. If the correct answer for that question does not lie in the set of options given by the majority, we consider it as an accuracy of zero. Otherwise, if there are m options tied in the majority vote, and the correct answer is one of these m , then we consider it as an accuracy of $\frac{100}{m}\%$ (hence, 100% if the correct answer is the only answer picked by the majority vote). We average the accuracy across all questions and across all iterations.

We choose majority voting as the means of aggregation since (a) it is the simplest and still most popular aggregation method, and (b) to enable an apples-to-apples comparison design since while more advanced aggregation algorithms have been developed for the baseline setting without the skip (Raykar et al., 2010; Zhou et al., 2012; Wauthier and Jordan, 2011; Chen et al., 2013; Khetan and Oh, 2016; Dawid and Skene, 1979; Karger et al., 2011; Liu et al., 2012; Vempaty et al., 2014; Zhang et al., 2014; Ipeirotis et al., 2014; Zhou et al., 2015; Shah et al., 2016c), but design of analogous algorithms for the new skip-based setting is still open.

The results are presented in Figure 5. We see that in most cases, our skip-based mechanism induces a lower labeling error at the aggregate level than the baseline. Furthermore, in many of the instances, the reduction is two-fold or higher.

All in all, in the experiments, we observe a substantial reduction in the error-rates while expending the same or lower amounts and receiving no negative comments from the workers, suggesting that these mechanisms can work; the fundamental theory underlying the mechanisms ensures that the system cannot be gamed in the long run. Our proposed settings and mechanisms thus have the potential to provide much higher quality labeled data as input to machine learning algorithms.

7. Discussion and Conclusions

In this concluding section, we first discuss the modeling assumptions that we made in this paper, followed by a discussion on future work and concluding remarks.

7.1 Modeling Assumptions

When forming the model for our problem, as in any other field of theoretical research, we had to make certain assumptions and choices. In what follows, we discuss the reasons for the modeling choices we made.

- *Use of gold standard questions.* We assume the existence of gold standard questions in the task, i.e., a subset of questions to which the answers are known to the system designer. The existence of gold standard is commonplace in crowdsourcing platforms (Le et al., 2010; Chen et al., 2011).

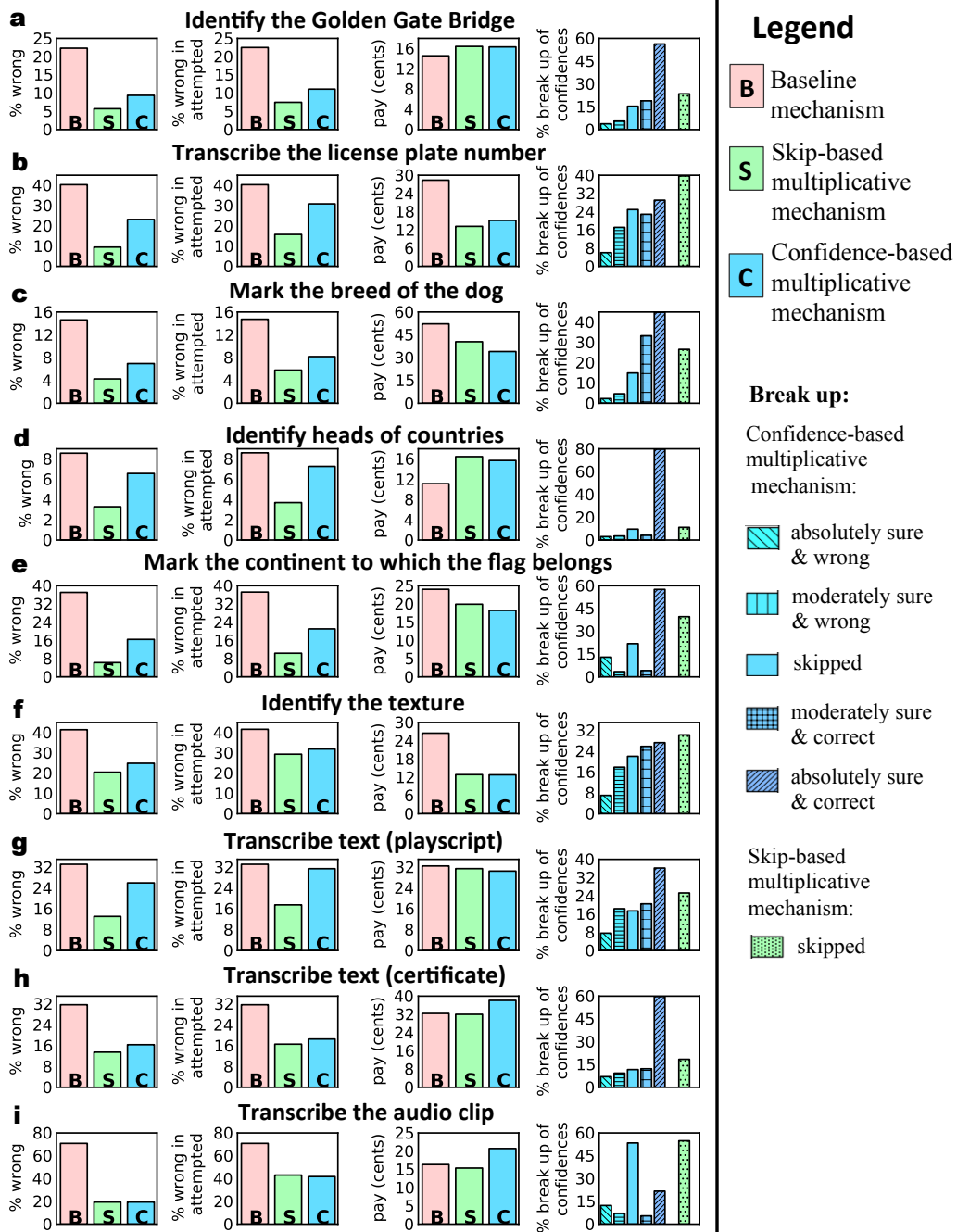


Figure 4: The error-rates in the raw data and payments in the nine experiments. Each individual bar in the plots corresponds to one mechanism in one experiment and is generated from 35 distinct workers (this totals to 945 worker-tasks).

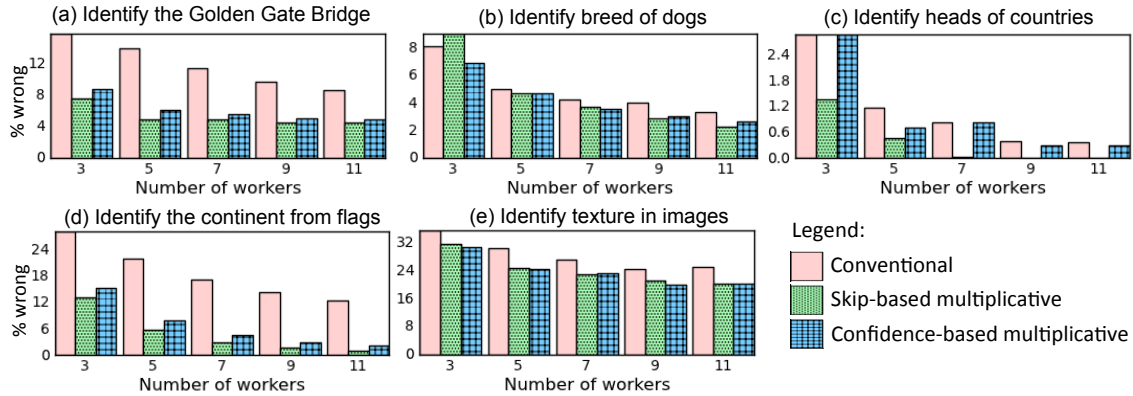


Figure 5: Error-rates in the aggregated data in the five experiments involving multiple-choice tasks.

- Workers aiming to maximize their expected payments:* We assume that the workers aim to maximize their expected payments. In many other problems in game theory, one often makes the assumption that people are “risk-averse”, and aim to maximize the expected value of some “utility function” of their payments. While we extend our results to general utility functions in Appendix C in order to accommodate such requirements, we also think that the assumption of workers maximizing their expected payments is a perfectly reasonable assumption for the crowdsourcing settings considered here. The reason is that each such task lasts for a handful of minutes and is worth a few tens of cents. Workers typically perform tens to hundreds of tasks per day, and consequently their empirical hourly wages very quickly converge to their expectation.
- Workers knowing their confidences:* We understand that in practice the workers will have noisy or granular estimates of their own beliefs. The mathematical assumption of workers knowing their precise confidences is an idealization intended for mathematical tractability. This is one of the reasons why we only elicit a quantized value of the workers’ beliefs (in terms of skipping or choosing one of a finite number of confidence levels), and not try to ask for a precise value.
- Eliciting a quantized version of the beliefs:* We do not directly attempt to elicit the values of the beliefs of the workers, but instead ask them to indicate only a quantization (e.g., “I’m not sure” or “moderately confident”, etc.). We prefer this quantization to direct assessment to real-valued probability, motivated by the extensive literature in psychology on the coarseness of human perception and processing (e.g., (Miller, 1956; Shiffrin and Nosofsky, 1994; Jones and Loe, 2013; Shah et al., 2016a)) establishing that humans are more comfortable at providing quantized responses. This notion is verified by experiments on Amazon Mechanical Turk in Shah et al. (2016a) where it is observed that people are more consistent when giving ordinal answers (comparing pairs of items) as opposed to when they are asked for numeric evaluations.
- Choosing the number of confidence levels L :* In the paper we assume that the number of confidence levels L is specified to us, and we provide mechanisms for any given choice of L . It is an interesting and challenging open problem to choose L for any given application in a principled manner. Up on increasing L , on one hand, we obtain additional nuanced information about the workers’ beliefs, while on the other hand, workers require a greater time and effort to provide

select the confidence level and their answers also tend to get noisier. In other words, both the “signal” and the “noise” in the data increase with an increase in the value of L , and lead to an interesting trade-off.

7.2 Open problems

We discuss two sets of open problems, one from the practical perspective and another on the theoretical front.

First, in the paper, we assumed that the number of total questions N in a task, the number of gold standard questions G , and the threshold T for skipping (or the number and thresholds of the different confidence levels) were provided to the mechanism. While these parameters may be chosen by hand by a system designer based on her own experience, a more principled design of these parameters is an important question. The choices for these parameters may have to be made based on certain tradeoffs. For instance, a higher value of G reduces the variance in the payments but uses more resources in terms of gold standard questions. Or for instance, more number of threshold levels L would increase the amount of information obtained about the workers’ beliefs, but also increase the noise in the workers’ estimates of her own beliefs.

A second open problem is the design of inference algorithms that can exploit the specific structure of the skip-based setting. There are several algorithms and theoretical analyses in the literature for aggregating data from multiple workers in the baseline setting (Raykar et al., 2010; Zhou et al., 2012; Wauthier and Jordan, 2011; Chen et al., 2013; Khetan and Oh, 2016; Dawid and Skene, 1979; Karger et al., 2011; Liu et al., 2012; Vempaty et al., 2014; Zhang et al., 2014; Ipeirotis et al., 2014; Zhou et al., 2015; Shah et al., 2016c). A useful direction of research in the future is to develop algorithms and theoretical guarantees that incorporate information about the workers’ confidences. For instance, for the skip-based setting, the missing labels are not missing “at random” but are correlated with the difficulty of the task; in the confidence-based setting, we elicit information about the workers’ perceived confidence levels. Designing algorithms that can exploit this information judiciously (e.g., via confidence-weighted worker/item constraints in the minimax entropy method of Zhou et al. (2015)) is a useful direction of future research.

7.3 Conclusions

Despite remarkable progress in machine learning and artificial intelligence, many problems are still not solvable by either humans or machines alone. In recent years, crowdsourcing has emerged as a powerful tool to combine both human and machine intelligence. Crowdsourcing is also a standard means of collecting labeled data for machine learning algorithms. However, crowdsourcing is often plagued with the problem of poor-quality output from workers.

We designed a reward mechanism for crowdsourcing to ensure collection of high-quality data. Under a very natural “no-free-lunch” axiom, we mathematically prove that surprisingly, our mechanism is the only feasible reward mechanism. We further show that among all possible incentive-compatible mechanisms, our “multiplicative” mechanism makes the strictly smallest expenditure on spammers. In preliminary experiments, we observe a significant drop in the error rates under this unique mechanism as compared to basic baseline mechanisms, suggesting that our mechanism has the potential to work well in practice. Our mechanisms offer some additional benefits. The pattern of skips or confidence levels of the workers provide a reasonable estimate of the difficulty of each question. In practice, the questions that are estimated to be more difficult may now be delegated

to an expert or to more non-expert workers. Secondly, the theoretical guarantees of the mechanism may allow for better post-processing of the data, incorporating the confidence information and improving the overall accuracy. The simplicity of the rules of our mechanisms may facilitate an easier adoption among the workers.

In conclusion, given the uniqueness in theory, simplicity, and good performance observed in practice, we envisage our ‘multiplicative’ mechanisms to be of interest to machine learning researchers and practitioners who use crowdsourcing to collect labeled data.

Acknowledgments

The work of Nihar B. Shah was funded in part by a Microsoft Research PhD fellowship. We thank John C. Platt, Christopher J. C. Burges and Christopher Meek for many inspiring discussions. We also thank John C. Platt and Martin J. Wainwright for helping in proof-reading and polishing parts of the manuscript. This work was done when Nihar B. Shah was an intern at Microsoft Research.

Appendix A. Proofs

In this section, we prove the claimed theoretical results whose proofs are not included in the main text of the paper.

The property of incentive-compatibility does not change upon any shift of the mechanism by a constant value or any scaling by a positive constant value. As a result, for the purposes of these proofs, we can assume without loss of generality that $\mu_{\min} = 0$.

A.1 Proof of Lemma 4: The Workhorse Lemma

First we consider the case of $G = N$. In the set $\{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_G\}$, for some $(\eta, \gamma) \in \{0, \dots, G-1\}^2$ such that $\eta + \gamma + 1 \leq G$, suppose there are η elements with a value 1, γ elements with a value -1 , and $(G - 1 - \eta - \gamma)$ elements with a value 0. Let us assume for now that $i = \eta + \gamma + 1$, $y_1 = 1, \dots, y_\eta = 1, y_{\eta+1} = -1, \dots, y_{\eta+\gamma} = -1, y_{\eta+\gamma+2} = 0, \dots, y_G = 0$.

Suppose the worker has confidences $(p_1, \dots, p_{\eta+\gamma}) \in (T, 1]^{\eta+\gamma}$ for the first $(\eta + \gamma)$ questions, a confidence of $q \in (0, 1]$ for the next question, and confidences smaller than T for the remaining $(G - \eta - \gamma - 1)$ questions. The mechanism must incentivize the worker to answer the first $(\eta + \gamma)$ questions and skip the last $(G - \eta - \gamma - 1)$ questions; for question $(\eta + \gamma + 1)$, it must incentivize the worker to answer if $q > T$ and skip if $q < T$. Supposing the worker indeed attempts the first $(\eta + \gamma)$ questions and skips the last $(G - \eta - \gamma - 1)$ questions, let $\mathbf{x} = \{x_1, \dots, x_{\eta+\gamma}\} \in \{-1, 1\}^{\eta+\gamma}$ denote the evaluation of the worker’s answers to the first $(\eta + \gamma)$ questions. Define quantities $\{r_j\}_{j \in [\eta+\gamma]}$ as $r_j = 1 - p_j$ for $j \in \{1, \dots, \eta\}$, and $r_j = p_j$ for $j \in \{\eta + 1, \eta + \gamma\}$. The requirement

of incentive compatibility necessitates

$$\begin{aligned}
 & q \sum_{\mathbf{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \dots, x_\eta, -x_{\eta+1}, \dots, -x_{\eta+\gamma}, 1, 0, \dots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right) \\
 & + (1-q) \sum_{\mathbf{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \dots, x_\eta, -x_{\eta+1}, \dots, -x_{\eta+\gamma}, -1, 0, \dots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right) \\
 & \stackrel{q < T}{\stackrel{q > T}{\sum}} \sum_{\mathbf{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \dots, x_\eta, -x_{\eta+1}, \dots, -x_{\eta+\gamma}, 0, 0, \dots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right).
 \end{aligned}$$

The left hand side of this expression is the expected payment if the worker chooses to answer question $(\eta + \gamma + 1)$, while the right hand side is the expected payment if she chooses to skip it. For any real-valued variable q , and for any real-valued constants a , b and c ,

$$aq \stackrel{q < c}{\stackrel{q > c}{\sum}} b \Rightarrow ac = b.$$

As a result,

$$\begin{aligned}
 & T \sum_{\mathbf{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \dots, x_\eta, -x_{\eta+1}, \dots, -x_{\eta+\gamma}, 1, 0, \dots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right) \\
 & + (1-T) \sum_{\mathbf{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \dots, x_\eta, -x_{\eta+1}, \dots, -x_{\eta+\gamma}, -1, 0, \dots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right) \\
 & - \sum_{\mathbf{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \dots, x_\eta, -x_{\eta+1}, \dots, -x_{\eta+\gamma-1}, 0, 0, \dots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right) = 0.
 \end{aligned} \tag{8}$$

The left hand side of (8) represents a polynomial in $(\eta + \gamma)$ variables $\{r_j\}_{j=1}^{\eta+\gamma}$ which evaluates to zero for all values of the variables within a $(\eta + \gamma)$ -dimensional solid Euclidean ball. Thus, the coefficients of the monomials in this polynomial must be zero. In particular, the constant term must be zero. The constant term appears when $x_j = 1 \forall j$ in the summations in (8). Setting the constant term to zero gives

$$\begin{aligned}
 & Tf(x_1 = 1, \dots, x_\eta = 1, -x_{\eta+1} = -1, \dots, -x_{\eta+\gamma} = -1, 1, 0, \dots, 0) \\
 & + (1-T)f(x_1 = 1, \dots, x_\eta = 1, -x_{\eta+1} = -1, \dots, -x_{\eta+\gamma} = -1, -1, 0, \dots, 0) \\
 & - f(x_1 = 1, \dots, x_\eta = 1, -x_{\eta+1} = -1, \dots, -x_{\eta+\gamma} = -1, 0, 0, \dots, 0) = 0,
 \end{aligned}$$

as desired. Since the arguments above hold for any permutation of the G questions, this completes the proof for the case of $G = N$.

Now consider the case $G < N$. Let $g : \{-1, 0, 1\}^N \rightarrow \mathbb{R}_+$ represent the expected payment given an evaluation of all the N answers, when the identities of the gold standard questions are

unknown. Here, the expectation is with respect to the (uniformly random) choice of the G gold standard questions. If $(x_1, \dots, x_N) \in \{-1, 0, 1\}^N$ are the evaluations of the worker's answers to the N questions then the expected payment is

$$g(x_1, \dots, x_N) = \frac{1}{\binom{N}{G}} \sum_{(i_1, \dots, i_G) \subseteq \{1, \dots, N\}} f(x_{i_1}, \dots, x_{i_G}). \quad (9)$$

Notice that when $G = N$, the functions f and g are identical.

In the set $\{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_G\}$, for some $(\eta, \gamma) \in \{0, \dots, G-1\}^2$ with $\eta + \gamma < G$, suppose there are η elements with a value 1, γ elements with a value -1 , and $(G-1-\eta-\gamma)$ elements with a value 0. Let us assume for now that $i = \eta + \gamma + 1$, $y_1 = 1, \dots, y_\eta = 1, y_{\eta+1} = -1, \dots, y_{\eta+\gamma} = -1, y_{\eta+\gamma+2} = 0, \dots, y_G = 0$.

Suppose the worker has confidences $\{p_1, \dots, p_{\eta+\gamma}\} \in (T, 1]^{\eta+\gamma}$ for the first $(\eta + \gamma)$ of the N questions, a confidence of $q \in (0, 1]$ for the next question, and confidences smaller than T for the remaining $(N - \eta - \gamma - 1)$ questions. The mechanism must incentivize the worker to answer the first $(\eta + \gamma)$ questions and skip the last $(N - \eta - \gamma - 1)$ questions; for the $(\eta + \gamma + 1)^{\text{th}}$ question, the mechanism must incentivize the worker to answer if $q > T$ and skip if $q < T$. Supposing the worker indeed attempts the first $(\eta + \gamma)$ questions and skips the last $(N - \eta - \gamma - 1)$ questions, let $\mathbf{x} = \{x_1, \dots, x_{\eta+\gamma}\} \in \{-1, 1\}^{\eta+\gamma}$ denote the the evaluation of the worker's answers to the first $(\eta + \gamma)$ questions. Define quantities $\{r_j\}_{j \in [\eta+\gamma]}$ as $r_j = 1 - p_j$ for $j \in \{1, \dots, \eta\}$, and $r_j = p_j$ for $j \in \{\eta + 1, \eta + \gamma\}$. The requirement of incentive compatibility necessitates

$$\begin{aligned} & q \sum_{\mathbf{x} \in \{-1, 1\}^{\eta+\gamma}} \left(g(x_1, \dots, x_\eta, -x_{\eta+1}, \dots, -x_{\eta+\gamma}, 1, 0, \dots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right) \\ & + (1-q) \sum_{\mathbf{x} \in \{-1, 1\}^{\eta+\gamma}} \left(g(x_1, \dots, x_\eta, -x_{\eta+1}, \dots, -x_{\eta+\gamma}, -1, 0, \dots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right) \\ & \stackrel{q < T}{\stackrel{q > T}{\sum}} \sum_{\mathbf{x} \in \{-1, 1\}^{\eta+\gamma}} \left(g(x_1, \dots, x_\eta, -x_{\eta+1}, \dots, -x_{\eta+\gamma}, 0, 0, \dots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right). \end{aligned} \quad (10)$$

Again, applying the fact that for any real-valued variable q and for any real-valued constants a, b and c , $aq \stackrel{q < c}{\stackrel{q > c}{\leq}} b \Rightarrow ac = b$, we get that

$$\begin{aligned} & Tg(x_1 = 1, \dots, x_\eta = 1, -x_{\eta+1} = -1, \dots, -x_{\eta+\gamma} = -1, 1, 0, \dots, 0) \\ & + (1-T)g(x_1 = 1, \dots, x_\eta = 1, -x_{\eta+1} = -1, \dots, -x_{\eta+\gamma} = -1, -1, 0, \dots, 0) \\ & - g(x_1 = 1, \dots, x_\eta = 1, -x_{\eta+1} = -1, \dots, -x_{\eta+\gamma} = -1, 0, 0, \dots, 0) = 0. \end{aligned} \quad (11)$$

The proof now proceeds via induction on the quantity $(G - \eta - \gamma - 1)$, i.e., on the number of skipped questions in $\{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_G\}$. We begin with the case of $(G - \eta - \gamma - 1) = G - 1$ which implies $\eta = \gamma = 0$. In this case (11) simplifies to

$$Tg(1, 0, \dots, 0) + (1-T)g(-1, 0, \dots, 0) = g(0, 0, \dots, 0).$$

Applying the expansion of function g in terms of function f from (9) gives

$$\begin{aligned} T(c_1 f(1, 0, \dots, 0) + c_2 f(0, 0, \dots, 0)) + (1 - T)(c_1 f(-1, 0, \dots, 0) + c_2 f(0, 0, \dots, 0)) \\ = (c_1 f(0, 0, \dots, 0) + c_2 f(0, 0, \dots, 0)) \end{aligned}$$

for constants $c_1 > 0$ and $c_2 > 0$ that respectively denote the probabilities that the first question is picked and not picked in the set of G gold standard questions. Canceling out the common terms on both sides of the equation, we get the desired result

$$Tf(1, 0, \dots, 0) + (1 - T)f(-1, 0, \dots, 0) = f(0, 0, \dots, 0).$$

Next, we consider the case when $(G - \eta - \gamma - 1)$ questions are skipped in the gold standard, and assume that the result is true when more than $(G - \eta - \gamma - 1)$ questions are skipped in the gold standard. In (11), the functions g decompose into a sum of the constituent f functions. These constituent functions f are of two types: the first where all of the first $(\eta + \gamma + 1)$ questions are included in the gold standard, and the second where one or more of the first $(\eta + \gamma + 1)$ questions are not included in the gold standard. The second case corresponds to situations where there are more than $(G - \eta - \gamma - 1)$ questions skipped in the gold standard and hence satisfies our induction hypothesis. The terms corresponding to these functions thus cancel out in the expansion of (11). The remainder comprises only evaluations of function f for arguments in which the first $(\eta + \gamma + 1)$ questions are included in the gold standard: since the last $(N - \eta - \gamma - 1)$ questions are skipped by the worker, the remainder evaluates to

$$\begin{aligned} Tc_3 f(y_1, \dots, y_{\eta+\gamma}, 1, 0, \dots, 0) + (1 - T)c_3 f(y_1, \dots, y_{\eta+\gamma}, -1, 0, \dots, 0) \\ = c_3 f(y_1, \dots, y_{\eta+\gamma}, 0, 0, \dots, 0) \end{aligned}$$

for some constant $c_3 > 0$. Dividing throughout by c_3 gives the desired result.

Finally, the arguments above hold for any permutation of the first G questions, thus completing the proof.

A.2 Proof of Theorem 7: Working of Algorithm 2

We first state three properties that the constants $\{\alpha_l\}_{l=-L}^L$ defined in Algorithm 2 must satisfy. We will use these properties subsequently in the proof of Theorem 7.

Lemma 14 For every $l \in \{0, \dots, L - 1\}$

$$T_{l+1}\alpha_{l+1} + (1 - T_{l+1})\alpha_{-(l+1)} = T_{l+1}\alpha_l + (1 - T_{l+1})\alpha_{-l}, \quad (12a)$$

and

$$S_{l+1}\alpha_{l+1} + (1 - S_{l+1})\alpha_{-(l+1)} = \alpha_0 = 1. \quad (12b)$$

Lemma 15 $\alpha_L > \alpha_{L-1} > \dots > \alpha_{-L} = 0$.

Lemma 16 For any $m \in \{1, \dots, L\}$, any $p > T_m$ and any $z < m$,

$$p\alpha_m + (1 - p)\alpha_{-m} > p\alpha_z + (1 - p)\alpha_{-z}, \quad (13a)$$

and for any $m \in \{0, \dots, L - 1\}$, any $p < T_{m+1}$ and any $z > m$,

$$p\alpha_m + (1 - p)\alpha_{-m} > p\alpha_z + (1 - p)\alpha_{-z}. \quad (13b)$$

The proof of these results are available at the end of this subsection. Assuming these lemmas hold, we will now complete the proof of Theorem 7.

The choice of $\alpha_{-L} = 0$ made in Algorithm 2 ensures that the payment is zero whenever any answer in the gold standard evaluates to $-L$. This choice ensures that the no-free-lunch axiom is satisfied. One can easily verify that the payment lies in the interval $[0, \mu_{\max}]$. It remains to prove that the proposed mechanism is incentive-compatible.

Define $E = (\epsilon_1, \dots, \epsilon_G) \in \{-1, 1\}^G$ and $E_{\setminus 1} = (\epsilon_2, \dots, \epsilon_G)$. Suppose the worker has confidences p_1, \dots, p_N for her N answers. For some $(s(1), \dots, s(N)) \in \{0, \dots, L\}^N$ suppose $p_i \in (T_{s(i)}, T_{s(i)+1}) \forall i \in \{1, \dots, N\}$, i.e., $s(1), \dots, s(N)$ are the correct confidence-levels for her answers. Consider any other set of confidence-levels $s'(1), \dots, s'(N)$. When the mechanism of Algorithm 2 is employed, the expected payment (from the point of view of the worker) on selecting confidence-levels $s(1), \dots, s(N)$ is

$$\mathbb{E}[\text{Pay}] = \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1, \dots, j_G) \\ \subseteq \{1, \dots, N\}}} \sum_{E \in \{-1, 1\}^G} \prod_{i=1}^G \alpha_{\epsilon_i s(j_i)} (p_{j_i})^{\frac{1+\epsilon_i}{2}} (1-p_{j_i})^{\frac{1-\epsilon_i}{2}} \quad (14a)$$

$$= \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1, \dots, j_G) \\ \subseteq \{1, \dots, N\}}} \sum_{E_{\setminus 1} \in \{-1, 1\}^{G-1}} (p_{j_1} \alpha_{s(j_1)} + (1-p_{j_1}) \alpha_{-s(j_1)}) \prod_{i=2}^G \alpha_{\epsilon_i s(j_i)} (p_{j_i})^{\frac{1+\epsilon_i}{2}} (1-p_{j_i})^{\frac{1-\epsilon_i}{2}} \quad (14b)$$

⋮

$$= \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1, \dots, j_G) \\ \subseteq \{1, \dots, N\}}} \prod_{i=1}^G (p_{j_i} \alpha_{s(j_i)} + (1-p_{j_i}) \alpha_{-s(j_i)}) \quad (14c)$$

$$> \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1, \dots, j_G) \\ \subseteq \{1, \dots, N\}}} \prod_{i=1}^G (p_{j_i} \alpha_{s'(j_i)} + (1-p_{j_i}) \alpha_{-s'(j_i)}) \quad (14d)$$

which is the expected payment under any other set of confidence-levels $s'(1), \dots, s'(N)$. The last inequality is a consequence of Lemma 16.

An argument similar to the above also proves that for any $m \in \{1, \dots, L\}$, if allowed to choose between only skipping and confidence-level m , the worker is incentivized to choose confidence-level m over skip if her confidence is greater S_m , and choose skip over level m if her confidence is smaller than S_m . Finally, from Lemma 15 we have $\alpha_L > \dots > \alpha_{-L} = 0$. It follows that the expected payment (14c) is strictly increasing in each of the values p_1, \dots, p_N . Thus the worker is incentivized to report the answer that she thinks is most likely to be correct.

A.2.1 PROOF OF LEMMA 14

Algorithm 2 states that $\alpha_{-l} = \frac{1-\alpha_l S_l}{1-S_l}$ for all $l \in [L]$. A simple rearrangement of the terms in this expression gives (12b).

Towards the goal of proving (12a), we will first prove an intermediate result:

$$\alpha_l > 1 > \alpha_{-l} \forall l \in \{L, \dots, 1\}. \quad (15)$$

The proof proceeds via an induction on $l \in \{L, \dots, 2\}$. The case of $l = 1$ will be proved separately. The induction hypothesis involves two claims: $\alpha_l > 1 > \alpha_{-l}$ and $T_l \alpha_l + (1 - T_l) \alpha_{-l} > 1$. The base case is $l = L$ for which we know that $\alpha_L = \frac{1}{S_L} > 1 > 0 = \alpha_{-L}$ and $T_L \alpha_L + (1 - T_L) \alpha_{-L} = \frac{T_L}{S_L} > 1$. Now suppose that the induction hypothesis is true for $(l+1)$. Rearranging the terms in the expression defining α_l in Algorithm 2 and noting that $1 > T_{l+1} > S_l$, we get

$$\begin{aligned} \alpha_l &= \frac{(1 - S_l) (T_{l+1} \alpha_{l+1} + (1 - T_{l+1}) \alpha_{-(l+1)}) - (1 - T_{l+1})}{(1 - S_l) - (1 - T_{l+1})} \\ &> \frac{(1 - S_l) - (1 - T_{l+1})}{(1 - S_l) - (1 - T_{l+1})} \\ &= 1. \end{aligned} \tag{16}$$

From (12b) we see that the value 1 is a convex combination of α_l and α_{-l} . Since $\alpha_l > 1$ and $S_l \in (0, 1)$, it must be that $\alpha_{-l} < 1$. Furthermore, since $T_l > S_l$ we get

$$\begin{aligned} T_l \alpha_l + (1 - T_l) \alpha_{-l} &> S_l \alpha_l + (1 - S_l) \alpha_{-l} \\ &= 1. \end{aligned}$$

This proves the induction hypothesis. Let us now consider $l = 1$. If $L = 1$ then we have $\alpha_L = \frac{1}{S_L} > 1 > 0 = \alpha_{-L}$ and we are done. If $L > 1$ then we have already proved that $\alpha_2 > 1 > \alpha_{-2}$ and $T_2 \alpha_2 + (1 - T_2) \alpha_{-2} > 1$. An argument identical to (16) onwards proves that $\alpha_1 > 1 > \alpha_{-1}$.

Now that we have proved $\alpha_l > \alpha_{-l} \forall l \in [L]$, we can rewrite the expression defining α_{-l} in Algorithm 2 as

$$S_l = \frac{1 - \alpha_{-l}}{\alpha_l - \alpha_{-l}}.$$

Substituting this expression for S_l in the definition of α_l in Algorithm 2 and making some simple rearrangements gives the desired result (12a).

A.2.2 PROOF OF LEMMA 15

We have already shown (15) in the proof of Lemma 14 above that $\alpha_l > 1 > \alpha_{-l} \forall l \in [L]$.

Next we will show that $\alpha_{l+1} > \alpha_l$ and $\alpha_{-(l+1)} < \alpha_{-l} \forall l \geq 0$. First consider $l = 0$, for which Algorithm 2 sets $\alpha_0 = 1$, and we have already proved that $\alpha_1 > 1 > \alpha_{-1}$.

Now consider some $l > 0$. Observe that since $S_l \alpha_l + (1 - S_l) \alpha_{-l} = 1$ (Lemma 14), $S_{l+1} > S_l$ and $\alpha_l > \alpha_{-l}$, it must be that

$$S_{l+1} \alpha_l + (1 - S_{l+1}) \alpha_{-l} > 1. \tag{17}$$

From Lemma 14, we also have

$$S_{l+1} \alpha_{l+1} + (1 - S_{l+1}) \alpha_{-(l+1)} = 1. \tag{18}$$

Subtracting (17) from (18) we get

$$S_{l+1} (\alpha_{l+1} - \alpha_l) + (1 - S_{l+1}) (\alpha_{-(l+1)} - \alpha_{-l}) < 0. \tag{19}$$

From Lemma 14 we also have

$$T_{l+1}\alpha_{l+1} + (1 - T_{l+1})\alpha_{-(l+1)} = T_{l+1}\alpha_l + (1 - T_{l+1})\alpha_{-l} \quad (20)$$

$$\Rightarrow T_{l+1}(\alpha_{l+1} - \alpha_l) + (1 - T_{l+1})(\alpha_{-(l+1)} - \alpha_{-l}) = 0. \quad (21)$$

Subtracting (19) from (21) gives

$$(T_{l+1} - S_{l+1})[(\alpha_{l+1} - \alpha_l) + (\alpha_{-l} - \alpha_{-(l+1)})] > 0. \quad (22)$$

Since $T_{l+1} > S_{l+1}$ by definition, it must be that

$$\alpha_{l+1} - \alpha_l > \alpha_{-(l+1)} - \alpha_{-l}. \quad (23)$$

Now, rearranging the terms in (20) gives

$$(\alpha_{l+1} - \alpha_l)T_{l+1} = -(\alpha_{-(l+1)} - \alpha_{-l})(1 - T_{l+1}). \quad (24)$$

Since $T_{l+1} \in (0, 1)$, it follows that the terms $(\alpha_{l+1} - \alpha_l)$ and $(\alpha_{-(l+1)} - \alpha_{-l})$ have opposite signs. Using (23) we conclude that $\alpha_{l+1} - \alpha_l > 0$ and $\alpha_{-(l+1)} - \alpha_{-l} < 0$.

A.2.3 PROOF OF LEMMA 16

Let us first prove (13a). First consider the case $z = m - 1$. From Lemma 14 we know that

$$T_m\alpha_{m-1} + (1 - T_m)\alpha_{-(m-1)} = T_m\alpha_m + (1 - T_m)\alpha_{-m},$$

and hence

$$\begin{aligned} 0 &= T_m(\alpha_m - \alpha_{m-1}) + T_m(\alpha_{-(m-1)} - \alpha_{-m}) - (\alpha_{-(m-1)} - \alpha_{-m}) \\ &< p(\alpha_m - \alpha_{m-1}) + p(\alpha_{-(m-1)} - \alpha_{-m}) - (\alpha_{-(m-1)} - \alpha_{-m}), \end{aligned} \quad (25)$$

where (25) is a consequence of $p > T_m$ and Lemma 15. A simple rearrangement of the terms in (25) gives (13a). Now, for any $z < m$, recursively apply this result to get

$$\begin{aligned} p\alpha_m + (1 - p)\alpha_{-m} &> p\alpha_{m-1} + (1 - p)\alpha_{-(m-1)} \\ &> p\alpha_{m-2} + (1 - p)\alpha_{-(m-2)} \\ &\vdots \\ &> p\alpha_z + (1 - p)\alpha_{-z}. \end{aligned}$$

Let us now prove (13b). We first consider the case $z = m + 1$. From Lemma 14 we know that

$$T_{m+1}\alpha_m + (1 - T_{m+1})\alpha_{-m} = T_{m+1}\alpha_{m+1} + (1 - T_{m+1})\alpha_{-(m+1)},$$

and hence

$$\begin{aligned} 0 &= T_{m+1}(\alpha_{m+1} - \alpha_m) + T_{m+1}(\alpha_{-m} - \alpha_{-(m+1)}) - (\alpha_{-m} - \alpha_{-(m+1)}) \\ &> p(\alpha_{m+1} - \alpha_m) + p(\alpha_{-m} - \alpha_{-(m+1)}) - (\alpha_{-m} - \alpha_{-(m+1)}), \end{aligned} \quad (26)$$

where (26) is a consequence of $p < T_{m+1}$ and Lemma 15. A simple rearrangement of the terms in (26) gives (13b). For any $z > m$, applying this result recursively gives

$$\begin{aligned} p\alpha_m + (1-p)\alpha_{-m} &> p\alpha_{m+1} + (1-p)\alpha_{-(m+1)} \\ &> p\alpha_{m+2} + (1-p)\alpha_{-(m+2)} \\ &\vdots \\ &> p\alpha_z + (1-p)\alpha_{-z}. \end{aligned}$$

A.3 Proof of Theorem 9: Uniqueness of Algorithm 2

We will first define one additional piece of notation. Let $g : \{-L, \dots, L\}^N \rightarrow \mathbb{R}_+$ denote the expected payment given an evaluation of the N answers, where the expectation is with respect to the (uniformly random) choice of the G gold standard questions. If $(x_1, \dots, x_N) \in \{-L, \dots, L\}^N$ are the evaluations of the worker's answers to the N questions then the expected payment is

$$g(x_1, \dots, x_N) = \frac{1}{\binom{N}{G}} \sum_{(i_1, \dots, i_G) \subseteq \{1, \dots, N\}} f(x_{i_1}, \dots, x_{i_G}). \quad (27)$$

Notice that when $G = N$, the functions f and g are identical.

The proof of uniqueness is based on a certain condition necessitated by incentive-compatibility stated in the form of Lemma 17 below. Note that this lemma does *not* require the generalized-no-free-lunch axiom, and may be of independent interest.

Lemma 17 *Any incentive-compatible mechanism must satisfy, for every question $i \in \{1, \dots, G\}$, every*

$$(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_G) \in \{-L, \dots, L\}^{G-1}, \text{ and every } m \in \{1, \dots, L\},$$

$$\begin{aligned} &T_m f(y_1, \dots, y_{i-1}, m, y_{i+1}, \dots, y_G) + (1 - T_m) f(y_1, \dots, y_{i-1}, -m, y_{i+1}, \dots, y_G) \\ &= T_m f(y_1, \dots, y_{i-1}, m-1, y_{i+1}, \dots, y_G) + (1 - T_m) f(y_1, \dots, y_{i-1}, -(m-1), y_{i+1}, \dots, y_G) \end{aligned} \quad (28a)$$

and

$$\begin{aligned} &S_m f(y_1, \dots, y_{i-1}, m, y_{i+1}, \dots, y_G) + (1 - S_m) f(y_1, \dots, y_{i-1}, -m, y_{i+1}, \dots, y_G) \\ &= f(y_1, \dots, y_{i-1}, 0, y_{i+1}, \dots, y_G). \end{aligned} \quad (28b)$$

Note that (28a) and (28b) coincide when $m = 1$, since $T_1 = S_1$ by definition.

We first prove that any incentive compatible mechanism that satisfies the no-free-lunch axiom must give a zero payment when one or more questions are selected with a confidence L and turn out to be incorrect. Let us assume for now that in the G questions in the gold standard, the first question is answered incorrectly with a confidence of L , the next $(G - 1 - S)$ questions are answered by the worker and have arbitrary evaluations, and the remaining S questions are skipped. The proof proceeds by an induction on S . If $S = G - 1$, the only attempted question is the first question and this is incorrect with confidence L . The no-free-lunch axiom necessitates a zero payment in this

case, thus satisfying the base case of our induction hypothesis. Now we prove the hypothesis for some S under the assumption that the hypothesis is true for every $S' > S$. From Lemma 4 with $m = 1$, we have

$$\begin{aligned}
 & T_1 f(-L, y_2, \dots, y_{G-S-1}, 1, 0, \dots, 0) + (1 - T_1) f(-L, y_2, \dots, y_{G-S-1}, -1, 0, \dots, 0) \\
 &= T_1 f(-L, y_2, \dots, y_{G-S-1}, 0, 0, \dots, 0) + (1 - T_1) f(-L, y_2, \dots, y_{G-S-1}, 0, 0, \dots, 0) \\
 &= f(-L, y_2, \dots, y_{G-S-1}, 0, 0, \dots, 0) \\
 &= 0,
 \end{aligned} \tag{29}$$

where the final equation (29) is a consequence of our induction hypothesis given the fact that $f(-L, y_2, \dots, y_{G-S-1}, 0, 0, \dots, 0)$ corresponds to the case when the last $(S + 1)$ questions are skipped and the first question is answered incorrectly with confidence L . Now, since the payment f must be non-negative and since $T \in (0, 1)$, it must be that

$$f(-L, y_2, \dots, y_{G-S-1}, 1, 0, \dots, 0) = 0 \tag{30a}$$

and

$$f(-L, y_2, \dots, y_{G-S-1}, -1, 0, \dots, 0) = 0. \tag{30b}$$

Repeatedly applying the same argument to $m = 2, \dots, L$ gives that for every value of m , it must be that $f(-L, y_2, \dots, y_{G-S-1}, m, 0, \dots, 0) = f(-L, y_2, \dots, y_{G-S-1}, -m, 0, \dots, 0) = 0$. This completes the proof of our induction hypothesis. Observe that each of the aforementioned arguments hold for any permutation of the G questions, thus proving the necessity of zero payment when any one or more answers are incorrect.

We will now prove that when no answers in the gold standard are incorrect with confidence L , the payment must be of the form described in Algorithm 1. Let κ denote the payment when all G questions in the gold standard are skipped, i.e.,

$$\kappa = f(0, \dots, 0).$$

Now consider any $S \in \{0, \dots, G - 1\}$ and any $(y_1, \dots, y_{G-S-1}, m) \in \{-L, \dots, L\}^{G-S}$. The payments $\{f(y_1, \dots, y_{G-S-1}, m, 0, \dots, 0)\}_{m=-L}^L$ must satisfy the $(2L - 1)$ linear constraints arising out of Lemma 17 and must also satisfy $f(y_1, \dots, y_{G-S-1}, -L, 0, \dots, 0) = 0$. This set of conditions comprises a total of $2L$ linearly independent constraints on the set of $(2L + 1)$ values $\{f(y_1, \dots, y_{G-S-1}, m, 0, \dots, 0)\}_{m=-L}^L$. The only set of solutions that meet these constraints are

$$f(y_1, \dots, y_{G-S-1}, m, 0, \dots, 0) = \alpha_m f(y_1, \dots, y_{G-S-1}, 0, 0, \dots, 0),$$

where the constants $\{\alpha_m\}_{m=-L}^L$ are as specified in Algorithm 2. Applying this argument G times, starting from $S = 0$ to $S = G - 1$, gives

$$f(y_1, \dots, y_G) = \kappa \prod_{j=1}^G \alpha_{y_j}.$$

Finally, the budget requirement necessitates $\mu_{\max} = \kappa (\alpha_L)^G$, which mandates the value of κ to be $\mu_{\max} \left(\frac{1}{\alpha_L}\right)^G$. This is precisely the mechanism described in Algorithm 2.

A.4 Proof of Lemma 17: Necessary condition for any incentive-compatible mechanism

First consider the case of $G = N$. For every $j \in \{1, \dots, i-1, i+1, \dots, G\}$, define

$$r_j = \begin{cases} 1 - p_j & \text{if } y_j \geq 0 \\ p_j & \text{if } y_j < 0. \end{cases}$$

Define $E_{\setminus i} = \{\epsilon_1, \dots, \epsilon_{i-1}, \epsilon_{i+1}, \dots, \epsilon_G\}$. For any $l \in \{-L, \dots, L\}$ let $\Lambda_l \in \mathbb{R}_+$ denote the expected payment (from the worker's point of view) when her answer to the i^{th} question evaluates to l :

$$\Lambda_l = \sum_{E_{\setminus i} \in \{-1, 1\}^{G-1}} \left(f(y_1 \epsilon_1, \dots, y_{i-1} \epsilon_{i-1}, l, y_{i+1} \epsilon_{i+1}, \dots, y_G \epsilon_G) \prod_{j \in [G] \setminus \{i\}} r_j^{\frac{1-\epsilon_j}{2}} (1-r_j)^{\frac{1+\epsilon_j}{2}} \right). \quad (31)$$

Consider a worker who has confidences $\{p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_G\} \in (0, 1)^{G-1}$ for questions $\{1, \dots, i-1, i+1, \dots, G\}$ respectively, and for question i suppose she has a confidence of $q \in (T_{m-1}, T_{m+1})$. For question i , we must incentivize the worker to select confidence-level m if $q > T_m$, and to select $(m-1)$ if $q < T_m$. This necessitates

$$q\Lambda_m + (1-q)\Lambda_{-m} \underset{q > T_m}{\lesseqgtr} \underset{q < T_m}{\gtrless} q\Lambda_{m-1} + (1-q)\Lambda_{-(m-1)}. \quad (32)$$

Also, for question i , the requirement of level m having a higher incentive as compared to skipping when $q > S_m$ and vice versa when $q < S_m$ necessitates

$$q\Lambda_m + (1-q)\Lambda_{-m} \underset{q > S_m}{\lesseqgtr} \underset{q < S_m}{\gtrless} \Lambda_0. \quad (33)$$

Now, note that for any real-valued variable q , and for any real-valued constants a , b and c ,

$$aq \underset{q > c}{\lesseqgtr} \underset{q < c}{\gtrless} b \quad \Rightarrow \quad ac = b.$$

Applying this fact to (32) and (33) gives

$$(T_m \Lambda_m + (1 - T_m) \Lambda_{-m}) - (T_m \Lambda_{m-1} + (1 - T_m) \Lambda_{-(m-1)}) = 0, \quad (34a)$$

$$(S_m \Lambda_m + (1 - S_m) \Lambda_{-m}) - \Lambda_0 = 0. \quad (34b)$$

From the definition of Λ_l in (31), we see that the left hand sides of (34a) and (34b) are both polynomials in $(G-1)$ variables $\{r_j\}_{j \in [G] \setminus \{i\}}$ and take a value of zero for all values of the variables in a $(G-1)$ -dimensional solid ball. Thus, each of the coefficients (of the monomials) in both polynomials must be zero, and in particular, the constant terms must also be zero. Observe that in both these polynomials, the constant term arises only when $\epsilon_j = 1 \forall j \in [G] \setminus \{i\}$ (which makes the

exponent of r_j to be 0 and that of $(1 - r_j)$ to be 1). Thus, setting the constant term to zero in the two polynomials results in

$$\begin{aligned} & T_m f(y_1, \dots, y_{i-1}, m, y_{i+1}, \dots, y_G) + (1 - T_m) f(y_1, \dots, y_{i-1}, -m, y_{i+1}, \dots, y_G) \\ &= T_m f(y_1, \dots, y_{i-1}, m - 1, y_{i+1}, \dots, y_G) + (1 - T_m) f(y_1, \dots, y_{i-1}, -(m - 1), y_{i+1}, \dots, y_G) \end{aligned} \quad (35a)$$

and

$$\begin{aligned} & S_m f(y_1, \dots, y_{i-1}, m, y_{i+1}, \dots, y_G) + (1 - S_m) f(y_1, \dots, y_{i-1}, -m, y_{i+1}, \dots, y_G) \\ &= f(y_1, \dots, y_{i-1}, 0, y_{i+1}, \dots, y_G) \end{aligned} \quad (35b)$$

thus proving the claim for the case of $G = N$.

Now consider the case when $G < N$. In order to simplify notation, let us assume $i = 1$ without loss of generality (since the arguments presented hold for any permutation of the questions). Suppose a worker's answers to questions $\{2, \dots, G\}$ evaluate to $(y_2, \dots, y_G) \in \{-L, \dots, L\}^{G-1}$, and further suppose that the worker skips the remaining $(N - G)$ questions. By going through arguments identical to those for $G = N$, but with f replaced by g , we get the necessity of

$$\begin{aligned} & T_m g(m, y_2, \dots, y_G, 0, \dots, 0) + (1 - T_m) g(-m, y_2, \dots, y_G, 0, \dots, 0) \\ &= T_m g(m - 1, y_2, \dots, y_G, 0, \dots, 0) + (1 - T_m) g(-(m - 1), y_2, \dots, y_G, 0, \dots, 0) \end{aligned} \quad (36a)$$

and

$$S_m g(m, y_2, \dots, y_G, 0, \dots, 0) + (1 - S_m) g(-m, y_2, \dots, y_G, 0, \dots, 0) = g(0, y_2, \dots, y_G, 0, \dots, 0). \quad (36b)$$

We now use this result in terms of function g to get an equivalent result in terms of function f . For some $S \in \{0, \dots, G - 1\}$, suppose $y_{G-S+1} = 0, \dots, y_G = 0$. The remaining proof proceeds via an induction on S . We begin with $S = G - 1$. In this case, (36a) and (36b) simplify to

$$\begin{aligned} & T_m g(m, 0, \dots, 0) + (1 - T_m) g(-m, 0, 0, \dots, 0) \\ &= T_m g(m - 1, 0, \dots, 0) + (1 - T_m) g(-(m - 1), 0, \dots, 0) \end{aligned} \quad (37a)$$

and

$$S_m g(m, 0, \dots, 0) + (1 - S_m) g(-m, 0, \dots, 0) = g(0, 0, \dots, 0). \quad (37b)$$

Applying the definition of function g from (27) leads to

$$\begin{aligned} & T_m (c_1 f(m, 0, \dots, 0) + c_2 f(0, 0, \dots, 0)) + (1 - T_m) (c_1 f(-m, 0, \dots, 0) + c_2 f(0, 0, \dots, 0)) \\ &= T_m (c_1 f(m - 1, 0, \dots, 0) + c_2 f(0, 0, \dots, 0)) \\ &\quad + (1 - T_m) (c_1 f(-(m - 1), 0, \dots, 0) + c_2 f(0, 0, \dots, 0)), \end{aligned} \quad (38a)$$

and

$$\begin{aligned} & S_m (c_1 f(m, 0, \dots, 0) + c_2 f(0, 0, \dots, 0)) + (1 - S_m) (c_1 f(-m, 0, \dots, 0) + c_2 f(0, 0, \dots, 0)) \\ &= (c_1 f(0, 0, \dots, 0) + c_2 f(0, 0, \dots, 0)) \end{aligned} \quad (38b)$$

for constants $c_1 > 0$ and $c_2 > 0$ that respectively denote the probabilities that the first question is picked and not picked in the set of G gold standard questions. Cancelling out the common terms on both sides of the equation, we get the desired results

$$\begin{aligned} T_m f(m, 0, \dots, 0) + (1 - T_m) f(-m, 0, \dots, 0) \\ = T_m f(m - 1, 0, \dots, 0) + (1 - T_m) f(-(m - 1), 0, \dots, 0) \end{aligned}$$

and

$$S_m f(m, 0, \dots, 0) + (1 - S_m) f(-m, 0, \dots, 0) = f(0, 0, \dots, 0).$$

Next, we consider the case of a general $S \in \{0, \dots, G - 2\}$ and assume that the result is true when $y_{G-S} = 0, \dots, y_G = 0$. In (36a) and (36b), the functions g decompose into a sum of the constituent f functions. These constituent functions f are of two types: the first where all of the first $(G - S)$ questions are included in the gold standard, and the second where one or more of the first $(G - S)$ questions are not included in the gold standard. The second case corresponds to situations where there are more than S questions skipped in the gold standard, i.e., when $y_{G-S} = 0, \dots, y_G = 0$, and hence satisfies our induction hypothesis. The terms corresponding to these functions thus cancel out in the expansion of (36a) and (36b). The remainder comprises only evaluations of function f for arguments in which the first $(G - S)$ questions are included in the gold standard: since the last $(N - G + S)$ questions are skipped by the worker, the remainder evaluates to

$$\begin{aligned} T_m c_3 f(y_1, \dots, y_{i-1}, m, y_{i+1}, \dots, y_G) + (1 - T_m) c_3 f(y_1, \dots, y_{i-1}, -m, y_{i+1}, \dots, y_G) \\ = T_m c_3 f(y_1, \dots, y_{i-1}, m - 1, y_{i+1}, \dots, y_G) \\ + (1 - T_m) c_3 f(y_1, \dots, y_{i-1}, -(m - 1), y_{i+1}, \dots, y_G), \end{aligned}$$

and

$$\begin{aligned} S_m c_3 f(y_1, \dots, y_{i-1}, m, y_{i+1}, \dots, y_G) + (1 - S_m) c_3 f(y_1, \dots, y_{i-1}, -m, y_{i+1}, \dots, y_G) \\ = c_3 f(y_1, \dots, y_{i-1}, 0, y_{i+1}, \dots, y_G), \end{aligned}$$

for some constant $c_3 > 0$. Dividing throughout by c_3 gives the desired result.

Finally, the arguments above hold for any permutation of the first G questions, thus completing the proof.

A.5 Necessity of $T_l > S_l$ for the Problem to be Well Defined

We now show that the restriction $T_l > S_l$ was necessary when defining the thresholds in Section 4.

Proposition 18 *Incentive-compatibility necessitates $T_l > S_l \forall l \in \{2, \dots, L\}$, even in the absence of the generalized-no-free-lunch axiom.*

First observe that the proof of Lemma 17 did not employ the generalized-no-free-lunch axiom, neither did it assume $T_l > S_l$. We will thus use the result of Lemma 17 to prove our claim.

Suppose the confidence of the worker for all but the first question is lower than T_1 and that the worker decides to skip all these questions. Suppose the worker attempts the first question. In order to ensure that the worker selects the answer that she believes is most likely to be true, it must be that

$$f(l, 0, \dots, 0) > f(-l, 0, \dots, 0) \quad \forall l \in [L].$$

We now call upon Lemma 17 where we set $i = 1$, $m = l$, $y_2 = \dots, y_G = 0$. Using the fact that $T_l > T_{l-1} \forall l \in \{2, \dots, L\}$, we get

$$\begin{aligned} & T_l f(l, 0, \dots, 0) + (1 - T_l) f(-l, 0, \dots, 0) \\ &= T_l f(l - 1, 0, \dots, 0) + (1 - T_l) f(-(l - 1), 0, \dots, 0) \\ &> T_{l-1} f(l - 1, 0, \dots, 0) + (1 - T_{l-1}) f(-(l - 1), 0, \dots, 0) \\ &= T_{l-1} f(l - 2, 0, \dots, 0) + (1 - T_{l-1}) f(-(l - 2), 0, \dots, 0) \\ &> T_{l-2} f(l - 2, 0, \dots, 0) + (1 - T_{l-2}) f(-(l - 2), 0, \dots, 0) \\ &\vdots \\ &> T_1 f(1, 0, \dots, 0) + (1 - T_1) f(-1, 0, \dots, 0) \\ &= f(0, \dots, 0) \\ &= S_l f(l, 0, \dots, 0) + (1 - S_l) f(-l, 0, \dots, 0). \end{aligned}$$

Since $f(l, 0, \dots, 0) > f(-l, 0, \dots, 0)$, we have the claimed result.

A.6 A Stronger No-free-lunch Axiom: Impossibility Results

In this section, we prove the various claims regarding the strong no-free-lunch axiom studied in Section 5.

A.6.1 PROOF OF PROPOSITION 10

If the worker skips all questions, then the expected payment is zero under the strong-no-free-lunch axiom. On the other hand, in order to incentivize knowledgeable workers to select answers whenever their confidences are greater than T , there must exist some situation in which the payment is strictly larger than zero. Suppose the payment is strictly positive when questions $\{1, \dots, z\}$ are answered correctly, questions $\{z + 1, \dots, z'\}$ are answered incorrectly, and the remaining questions are skipped. If the confidence of the unknowledgeable worker is in the interval $(0, T)$ for every question, then attempting to answer questions $\{1, \dots, z'\}$ and skipping the rest fetches her a payment that is strictly positive in expectation. Thus, this unknowledgeable worker is incentivized to answer at least one question.

A.6.2 PROOF OF PROPOSITION 11

Consider a (knowledgeable) worker who has a confidence of $p \in (T, 1]$ for the first question, $q \in (0, 1)$ for the second question, and confidences in the interval $(0, T)$ for the remaining questions. Suppose the worker attempts to answer the first question (and selects the answer she believes is most likely to be correct) and skips the last $(N - 2)$ questions as desired. Now, in order to incentivize her to answer the second question if $q > T$ and skip the second question if $q < T$, the payment

mechanism must satisfy

$$pqg(1, 1, 0, \dots, 0) + (1-p)qg(-1, 1, 0, \dots, 0) + p(1-q)g(1, -1, 0, \dots, 0) \\ + (1-p)(1-q)g(-1, -1, 0, \dots, 0) \stackrel{q < T}{\lesssim} pg(1, 0, 0, \dots, 0) + (1-p)g(-1, 0, 0, \dots, 0).$$

For any real-valued variable q , and for any real-valued constants a , b and c ,

$$aq \stackrel{q < c}{\lesssim} b \quad \Rightarrow \quad ac = b.$$

As a result,

$$pTg(1, 1, 0, \dots, 0) + (1-p)Tg(-1, 1, 0, \dots, 0) + p(1-T)g(1, -1, 0, \dots, 0) \\ + (1-p)(1-T)g(-1, -1, 0, \dots, 0) - pg(1, 0, 0, \dots, 0) - (1-p)g(-1, 0, 0, \dots, 0) = 0.$$

The left hand side of this equation is a polynomial in variable p and takes a value of zero for all values of p in a one-dimensional box $(T, 1]$. It follows that the monomials of this polynomial must be zero, and in particular the constant term must be zero:

$$Tg(-1, 1, 0, \dots, 0) + (1-T)g(-1, -1, 0, \dots, 0) - g(-1, 0, 0, \dots, 0) = 0.$$

The strong-no-free-lunch axiom implies $f(-1, -1, 0, \dots, 0) = f(-1, 0, \dots, 0) = f(0, \dots, 0) = 0$, and hence $g(-1, -1, 0, \dots, 0) = g(-1, 0, 0, \dots, 0) = 0$. Since $T \in (0, 1)$, we have

$$0 = g(-1, 1, 0, \dots, 0) \\ = c_1 f(-1, 1, 0, \dots, 0) + c_2 f(-1, 0, \dots, 0) + c_2 f(1, 0, \dots, 0), \quad (39)$$

for some constants $c_1 > 0$ and $c_2 > 0$ that represent the probability that the first two questions are included in the gold standard, and the probability that the first (or, second) but not the second (or, first) questions are included in the gold standard. Since f is a non-negative function, it must be that

$$f(1, 0, \dots, 0) = 0.$$

Now suppose a (knowledgeable) worker has a confidence of $p \in (T, 1]$ for the first question and confidences lower than T for the remaining $(N-1)$ questions. Suppose the worker chooses to skip the last $(N-1)$ questions as desired. In order to incentivize the worker to answer the first question, the mechanism must satisfy for all $p \in (T, 1]$,

$$0 < pg(1, 0, \dots, 0) + (1-p)g(-1, 0, \dots, 0) - g(0, 0, \dots, 0) \\ = pc_3 f(1, 0, \dots, 0) + pc_4 f(0, 0, \dots, 0) + (1-p)c_3 f(-1, 0, \dots, 0) \\ + (1-p)c_4 f(0, 0, \dots, 0) - f(0, 0, \dots, 0) \\ = 0,$$

where $c_3 > 0$ and $c_4 > 0$ are some constants. The final equation is a result of the strong-no-free-lunch axiom and the fact that $f(1, 0, \dots, 0) = 0$ as proved above. This yields a contradiction, and hence no incentive-compatible mechanism f can satisfy the strong-no-free-lunch axiom when $G < N$ even when allowed to address only knowledgeable workers.

Finally, as a sanity check, note that if $G = N$ then $c_2 = 0$ in (39). The proof above thus doesn't hold when $G = N$.

A.6.3 PROOF OF PROPOSITION 12

We will first show that the mechanism works as desired.

First consider the case when the worker is unknowledgeable and her confidences are of the form $T > p_{(1)} \geq p_{(2)} \geq p_{(3)} \geq \dots \geq p_{(G)}$. If she answers only the first question, then her expected payment is

$$\kappa \frac{p_{(1)}}{T}.$$

Let us now see her expected payment if she doesn't follow this answer pattern. The strong-no-free-lunch axiom implies that if the worker doesn't answer any question then her expected payment is zero. Suppose the worker chooses to answer questions $\{i_1, \dots, i_z\}$. In that case, her expected payment is

$$\kappa \frac{p_{i_1} \dots p_{i_z}}{T^z} = \kappa \frac{p_{i_1}}{T} \dots \frac{p_{i_z}}{T} \tag{40}$$

$$\leq \kappa \left(\frac{p_{(1)}}{T} \right)^z \tag{41}$$

$$\leq \kappa \frac{p_{(1)}}{T}, \tag{42}$$

where (42) uses the fact that $p_{(1)} < T$. The inequality in (42) becomes an equality only when $z = 1$. Now when $z = 1$, the inequality in (41) becomes an equality only when $i_1 = (1)$. Thus the unknowledgeable worker is incentivized to answer only one question – the one that she has the highest confidence in.

Now consider a knowledgeable worker and suppose her confidences are of the form $p_{(1)} \geq \dots \geq p_{(m)} > T > p_{(m+1)} \geq \dots \geq p_{(G)}$ for some $m \geq 1$. If the worker answers questions $(1), \dots, (m)$ as desired, her expected payment is

$$\kappa \frac{p_{(1)}}{T} \dots \frac{p_{(m)}}{T}.$$

Now let us see what happens if the worker does not follow this answer pattern. The strong-no-free-lunch axiom implies that if the worker doesn't answer any question then her expected payment is zero. Now, if she answers some other set of questions, say questions $\{i_1, \dots, i_z\}$ with $p_{(1)} \leq p_{i_1} < \dots < p_{i_y} \leq p_{(m)} < p_{i_{y+1}} < \dots < p_{i_z} \leq p_{(G)}$. The expected payment in that case is

$$\begin{aligned} \kappa \frac{p_{i_1} \dots p_{i_z}}{T^z} &= \kappa \frac{p_{i_1}}{T} \dots \frac{p_{i_z}}{T} \\ &\leq \kappa \frac{p_{i_1}}{T} \dots \frac{p_{i_y}}{T} \end{aligned} \tag{43}$$

$$\leq \kappa \frac{p_{(1)}}{T} \dots \frac{p_{(m)}}{T} \tag{44}$$

where inequality (43) is a result of $\frac{p_{i_j}}{T} \leq 1 \ \forall j > y$ and holds with equality only when $y = z$. Inequality (44) is a result of $\frac{p_{(j)}}{T} \geq 1 \ \forall j \leq m$ and holds with equality only when $y = m$. Thus the expected payment is maximized when $i_1 = (1), \dots, i_z = (m)$ as desired. Finally, the payment strictly increases with an increase in the confidences, and hence the worker is incentivized to always consider the answer that she believes is most likely to be correct.

We now show that this mechanism is unique.

The necessary conditions derived in Lemma 4, when restricted to the case of $G = N$ and $(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_G) \neq \{0\}^{N-1}$, is also applicable to the present setting. This is because the strong-no-free-lunch axiom assumed here is a stronger condition than the no-free-lunch axiom considered in Lemma 4, and moreover, $(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_G) \neq \{0\}^{N-1}$ avoids the use of unknowledgeable workers in the proof of Lemma 4. It follows that for every question $i \in \{1, \dots, G\}$ and every $(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_G) \in \{-1, 0, 1\}^{G-1} \setminus \{0\}^{G-1}$, it must be that

$$\begin{aligned} T f(y_1, \dots, y_{i-1}, 1, y_{i+1}, \dots, y_G) + (1 - T) f(y_1, \dots, y_{i-1}, -1, y_{i+1}, \dots, y_G) \\ = f(y_1, \dots, y_{i-1}, 0, y_{i+1}, \dots, y_G). \end{aligned} \quad (45)$$

We claim that the payment must be zero whenever the number of incorrect answers $W > 0$. The proof proceeds by induction on the number of correct answers C . First suppose $C = 0$ (and $W > 0$). Then all questions are either wrong or skipped, and hence by the strong-no-free-lunch axiom, the payment must be zero. Now suppose the payment is necessarily zero whenever $W > 0$ and the total number of correct answers is $(C - 1)$ or lower, for some $C \in [G - 1]$. Consider any evaluation $(y_1, \dots, y_G) \in \{-1, 0, 1\}^G$ in which the number of incorrect answers is more than zero and the number of correct answers is C . Suppose $y_i = 1$ for some $i \in [G]$, and $y_j = -1$ for some $j \in [G] \setminus \{i\}$. Then from the induction hypothesis, we have $f(y_1, \dots, y_{i-1}, -1, y_{i+1}, \dots, y_G) = f(y_1, \dots, y_{i-1}, 0, y_{i+1}, \dots, y_G) = 0$. Applying (45) and noting that $T \in (0, 1)$, we get that $f(y_1, \dots, y_{i-1}, 1, y_{i+1}, \dots, y_G) = 0$ as claimed. This result also allows us to simplify (45) to: For every question $i \in \{1, \dots, G\}$ and every $(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_G) \in \{-1, 0, 1\}^{G-1} \setminus \{0\}^{G-1}$,

$$f(y_1, \dots, y_{i-1}, 1, y_{i+1}, \dots, y_G) = \frac{1}{T} f(y_1, \dots, y_{i-1}, 0, y_{i+1}, \dots, y_G). \quad (46)$$

We now show that when $C > 0$ and $W = 0$, the payment must necessarily be of the form described in the statement of Proposition 12. The proof again proceeds via an induction on the number of correct answers C (≥ 1). Define a quantity $\kappa > 0$ as

$$\kappa = T f(1, 0, \dots, 0). \quad (47)$$

Now consider the payment $f(1, y_2, \dots, y_G)$ for some $(y_2, \dots, y_G) \in \{0, 1\}^{G-1} \setminus \{0\}^{G-1}$ with C correct answers. Applying (46) repeatedly (once for every i such that $y_i = 1$), we get

$$f(1, y_2, \dots, y_G) = \frac{\kappa}{T^C}.$$

Unlike other results in this paper, at this point we cannot claim the result to hold for all permutations of the questions. This is because we have defined the quantity κ in an asymmetric manner (47), in terms of the payment function when the *first* question is correct and the rest are skipped. In what follows, we will prove that the result claimed in the statement of Proposition 12 indeed holds for all permutations of the questions.

From (46) we have

$$\begin{aligned} f(0, 1, 0, \dots, 0) &= T f(1, 1, 0, \dots, 0) \\ &= f(1, 0, 0, \dots, 0) \\ &= \kappa. \end{aligned}$$

Thus the payment must be κ even if the second answer in the gold standard is correct and the rest are skipped. In fact, the argument holds when any one answer in the gold standard is correct and the rest are skipped. Thus the definition of κ is not restricted to the first question alone as originally defined in (47), but holds for all permutations of the questions. This allows the other arguments above to be applicable to any permutation of the questions. Finally, the budget constraint of μ_{\max} fixes the value of κ to that claimed, thereby completing the proof.

A.6.4 PROOF OF PROPOSITION 13

Proposition 12 proved that under the skip-based setting with the strong-no-free-lunch axiom, the payment must be zero when one or more answers are incorrect. This part of the proof of Proposition 12 holds even when $L > 1$. It follows that for any question, the penalty for an incorrect answer is the same for any confidence-level in $\{1, \dots, L\}$. Thus the worker is incentivized to always select that confidence-level for which the payment is the maximum when the answer is correct, irrespective of her own confidence about the question. This contradicts our requirements.

Appendix B. Details of Experiments

In this section, we provide further details about the experiments described earlier in Section 6.2. The experiments were carried out on the Amazon Mechanical Turk (`mturk.com`) online crowdsourcing platform in the time period June to October 2013. Figure 6 illustrates the interface shown to the workers for each of the experiments described in Section 6.2, while Figure 7 depicts the instructions given to the workers. The following are more details of each individual experiment. In the description, the notation κ is as defined in Algorithm 1 and Algorithm 2, namely, $\kappa = (\mu_{\max} - \mu_{\min})T^G$ for the skip-based setting and $\kappa = (\mu_{\max} - \mu_{\min})\left(\frac{1}{\alpha_L}\right)^G$ for the confidence-based setting.

B.1 Recognizing the Golden Gate Bridge

A set of 21 photographs of bridges were shown to the workers, and for each photograph, they had to identify if it depicted the Golden Gate Bridge or not. An example of this task is depicted in Figure 6a, and the instructions provided to the worker under the three mechanisms are depicted in Figure 7. The fixed amount offered to workers was $\mu_{\min} = 3$ cents for the task, and the bonus was based on 3 gold standard questions. We compared (a) the baseline mechanism with 5 cents for each correct answer in the gold standard, (b) the skip-based mechanism with $\kappa = 5.9$ and $\frac{1}{T} = 1.5$, and (c) the confidence-based mechanism with $\kappa = 5.9$ cents, $L = 2$, $\alpha_2 = 1.5$, $\alpha_1 = 1.4$, $\alpha_0 = 1$, $\alpha_{-1} = 0.5$, $\alpha_{-2} = 0$. The results of this experiment are presented in Figure 4a.

B.2 Transcribing Vehicles' License Plate Numbers from Photographs

This task presented the workers with 18 photographs of cars and asked them to transcribe the license plate numbers from each of them (source of photographs: <http://www.coolpl8z.com>). An example of this task is depicted in Figure 6b. The fixed amount offered to workers was $\mu_{\min} = 4$ cents for the task, and the bonus was based on 4 gold standard questions. We compared (a) the baseline mechanism with 10 cents for each correct answer in the gold standard, (b) the skip-based mechanism with $\kappa = 0.62$ and $\frac{1}{T} = 3$, and (c) the confidence-based mechanism with $\kappa = 3.1$ cents, $L = 2$, $\alpha_2 = 2$, $\alpha_1 = 1.95$, $\alpha_0 = 1$, $\alpha_{-1} = 0.5$, $\alpha_{-2} = 0$. The results of this experiment are presented in

Figure 4b. When evaluating, in the worker’s answers as well as in the true solutions, we converted all text to upper case, and removed all spaces and punctuations. We then declared a worker’s answer to be in error if it did not have an exact match with the true solution.

B.3 Classifying Breeds of Dogs

This task required workers to identify the breeds of dogs shown in 85 images (source of images: Khosla et al. (2011); Deng et al. (2009)). For each image, the worker was given ten breeds to choose from. An example of this task is depicted in Figure 6c. The fixed amount offered to workers was $\mu_{\min} = 5$ cents for the task, and the bonus was based on 7 gold standard questions. We compared (a) the baseline mechanism with 8 cents for each correct answer in the gold standard, (b) the skip-based mechanism with $\kappa = 0.78$ and $\frac{1}{T} = 2$, and (c) the confidence-based mechanism with $\kappa = 0.78$ cents, $L = 2$, $\alpha_2 = 2$, $\alpha_1 = 1.66$, $\alpha_0 = 1$, $\alpha_{-1} = 0.67$, $\alpha_{-2} = 0$. The results of this experiment are presented in Figure 4c.

B.4 Identifying Heads of Countries

Names of 20 personalities were provided and had to be classified as to whether they were ever the (a) President of the USA, (b) President of India, (c) Prime Minister of Canada, or (d) neither of these. An example of this task is depicted in Figure 6d. The fixed amount offered to workers was $\mu_{\min} = 2$ cents for the task, and the bonus was based on 4 gold standard questions. While the ground truth in most other multiple-choice experiments had approximately an equal representation from all classes, this experiment was heavily biased with one of the classes never being correct and another being correct for just 3 of the 20 questions. We compared (a) the baseline mechanism with 2.5 cents for each correct answer in the gold standard, (b) the skip-based mechanism with $\kappa = 0.25$ and $\frac{1}{T} = 3$, and (c) the confidence-based mechanism with $\kappa = 1.3$ cents, $L = 2$, $\alpha_2 = 2$, $\alpha_1 = 1.95$, $\alpha_0 = 1$, $\alpha_{-1} = 0.5$, $\alpha_{-2} = 0$. The results of this experiment are presented in Figure 4d.

B.5 Identifying Flags

This was a relatively long task, with 126 questions. Each question required the workers to identify if a displayed flag belonged to a place in (a) Africa, (b) Asia/Oceania, (c) Europe, or (d) neither of these. An example of this task is depicted in Figure 6e. The fixed amount offered to workers was $\mu_{\min} = 4$ cents for the task, and the bonus was based on 8 gold standard questions. We compared (a) the baseline mechanism with 4 cents for each correct answer in the gold standard, (b) the skip-based mechanism with $\kappa = 0.2$ and $\frac{1}{T} = 2$, and (c) the confidence-based mechanism with $\kappa = 0.2$ cents, $L = 2$, $\alpha_2 = 2$, $\alpha_1 = 1.66$, $\alpha_0 = 1$, $\alpha_{-1} = 0.67$, $\alpha_{-2} = 0$. The results of this experiment are presented in Figure 4e.

B.6 Distinguishing Textures

This task required the workers to identify the textures shown in 24 grayscale images (source of images: Lazechnik et al. (2005, Data set 1: Textured surfaces)). For each image, the worker had to choose from 8 different options. Such a task has applications in computer vision, where it aids in recognition of objects or their surroundings. An example of this task is depicted in Figure 6f. The fixed amount offered to workers was $\mu_{\min} = 3$ cents for the task, and the bonus was based on 4 gold

standard questions. We compared (a) the baseline mechanism with 10 cents for each correct answer in the gold standard, (b) the skip-based mechanism with $\kappa = 3.1$ and $\frac{1}{T} = 2$, and (c) the confidence-based mechanism with $\kappa = 3.1$ cents, $L = 2$, $\alpha_2 = 2$, $\alpha_1 = 1.66$, $\alpha_0 = 1$, $\alpha_{-1} = 0.67$, $\alpha_{-2} = 0$. The results of this experiment are presented in Figure 4f.

B.7 Transcribing Text from an Image: Film Certificate

The task showed an image containing 11 (short) lines of blurry text which the workers had to decipher. We used text from a certain certificate which movies releasing in India are provided. We slightly modified its text in order to prevent workers from searching a part of it online and obtaining the entire text by searching the first few transcribed lines on the Internet. An example of this task is depicted in Figure 6g. The fixed amount offered to workers was $\mu_{\min} = 5$ cents for the task, and the bonus was based on 2 gold standard questions. We compared (a) the baseline mechanism with 20 cents for each correct answer in the gold standard, (b) the skip-based mechanism with $\kappa = 5.5$ and $\frac{1}{T} = 3$, and (c) the confidence-based mechanism with $\kappa = 12.5$ cents, $L = 2$, $\alpha_2 = 2$, $\alpha_1 = 1.95$, $\alpha_0 = 1$, $\alpha_{-1} = 0.5$, $\alpha_{-2} = 0$. The results of this experiment are presented in Figure 4g. When evaluating, in the worker’s answers as well as in the true solutions, we converted all text to upper case, and removed all spaces and punctuations. We then declared a worker’s answer to be in error if it did not have an exact match with the true solution.


B.8 Transcribing Text from an Image: Script of a Play

The task showed an image containing 12 (short) lines of blurry text which the workers had to decipher. We borrowed a paragraph from Shakespeare’s play ‘As You Like It.’ We slightly modified the text of the play in order to prevent workers from searching a part of it online and obtaining the entire text by searching the first few transcribed lines on the internet. An example of this task is depicted in Figure 6h. The fixed amount offered to workers was 5 cents for the task, and the bonus was based on 2 gold standard questions. We compared (a) the baseline mechanism with $\mu_{\min} = 20$ cents for each correct answer in the gold standard, (b) the skip-based mechanism with $\kappa = 5.5$ and $\frac{1}{T} = 3$, and (c) the confidence-based mechanism with $\kappa = 12.5$ cents, $L = 2$, $\alpha_2 = 2$, $\alpha_1 = 1.95$, $\alpha_0 = 1$, $\alpha_{-1} = 0.5$, $\alpha_{-2} = 0$. The results of this experiment are presented in Figure 4h. When evaluating, in the worker’s answers as well as in the true solutions, we converted all text to upper case, and removed all spaces and punctuations. We then declared a worker’s answer to be in error if it did not have an exact match with the true solution.

B.9 Transcribing Text from Audio Clips


The workers were given 10 audio clips which they had to transcribe to text. Each audio clip was 3 to 6 seconds long, and comprised of a short sentence, e.g., “my favorite topics of conversation are sports, politics, and movies.” Each of the clips were recorded in different accents using a text-to-speech converter. An example of this task is depicted in Figure 6i. The fixed amount offered to workers was $\mu_{\min} = 5$ cents for the task, and the bonus was based on 2 gold standard questions. We compared (a) the baseline mechanism with 20 cents for each correct answer in the gold standard, (b) the skip-based mechanism with $\kappa = 5.5$ and $\frac{1}{T} = 3$, and (c) the confidence-based mechanism with $\kappa = 12.5$ cents, $L = 2$, $\alpha_2 = 2$, $\alpha_1 = 1.95$, $\alpha_0 = 1$, $\alpha_{-1} = 0.5$, $\alpha_{-2} = 0$. The results of this experiment are presented in Figure 4i.

a Recognize the Golden Gate Bridge




Golden Gate
 NOT Golden Gate

b Transcribe the license plate number



Answer:

c Mark the breed of the dog




Afghan Hound
 Doberman
 French Bulldog
 Tibetan Terrier
 ⋮

d Identify heads of countries

Mohandas Gandhi


President of the USA
 President of India
 Prime Minister of Canada
 None of the above

e Mark the continent to which the flag belongs



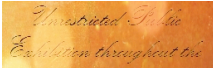
Africa
 Asia/Oceania
 Europe
 None of these

f Identify the texture



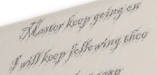
Granite
 Carpet
 Fur
 Glass
 Corduroy
 Wood
 None of these

g Transcribe text (playscript)




Line 1:
Line 2:

h Transcribe text (certificate)



Line 1:
Line 2:

i Transcribe the audio clip



Answer:

Figure 6: Various tasks on which the payment mechanisms were tested. The interfaces shown are that of the baseline mechanism, i.e., without the skipping or confidence choices.

a Baseline Mechanism

*** Instructions for BONUS (Read Carefully) ***

- There are three questions whose answers are known to us, based on which the bonus is calculated
- **BONUS (cents)** = 5 * number of questions out of these that you correctly answer

b Skip-based multiplicative mechanism

If you are not sure about any answer, then mark "I'm not sure"
 You **need to mark** at least something for **every question**, otherwise your work will be **rejected**

*** Instructions for BONUS (Read Carefully) ***

- You start with 5.9 cents of bonus for this HIT
- There are three questions whose answers are known to us, based on which the bonus is calculated
- For each of these questions you answer **CORRECTLY**, your bonus will **INCREASE BY 50%** (every 1 cent will become 1.5 cents)
- If you answer any of these questions **WRONG**, your bonus will become **ZERO**
- So for questions you are not sure of, mark the **"I'm not sure"** option: this does not affect the bonus

c Confidence-based multiplicative mechanism

For each answer, you also need to indicate how sure you are about that answer
 If you are not sure about any answer, then mark "I don't know"
 You **need to mark** at least something for **every question**, otherwise your work will be **rejected**

*** Instructions for BONUS (Read Carefully) ***

- If any answer marked **"absolutely sure"** is **wrong**, your bonus will become **ZERO** for this entire HIT (you do not get any bonus for this HIT)
- For every answer marked **"absolutely sure"** that is **correct**, your bonus will **INCREASE BY 50%** (every 1 cent will become 1.5 cents)
- For every answer marked **"moderately sure"** that is **wrong**, your bonus will be **HALVED** (every 1 cent will become half a cent)
- For every answer marked **"moderately sure"** that is **correct**, your bonus will be **INCREASE BY 40%** (every 1 cent will become 1.4 cents)
- Marking **"I don't know"** for any answer **does not change** your bonus

Figure 7: An example of the instructions displayed to the worker under the three mechanisms.

Appendix C. General Utility Functions

In this section, we consider a setting where the worker, instead of maximizing her expected payment, aims to maximize the expected value of some *utility function* of her payment. Consider any function $U : \mathbb{R}_+ \rightarrow \mathcal{I}$, where \mathcal{I} is any interval on the real number line. We will require the function U to be strictly increasing and to have an inverse. Examples of such functions include $U(x) = \log(1 + x)$ with $\mathcal{I} = \mathbb{R}_+$, $U(x) = \sqrt{x}$ with $\mathcal{I} = \mathbb{R}_+$, and $U(x) = 1 - e^{-x}$ with $\mathcal{I} = [0, 1]$. For any payment f made to the worker (based on the evaluation of her answers to the gold standard questions), her utility for this payment is $U(f)$. The worker aims to maximize the expected value of $U(f)$, where the expectation is with respect to her beliefs regarding correctness of her answers and the uniformly random distribution of the G gold standard questions among the set of N questions. The function U is assumed to be known to the worker as well as the system designer.

Consider the confidence-based setting of Section 4 (of which, the skip-based setting of Section 3 is a special case). Recall the notation $\{x_i\}_{i=1}^G$, $\{\alpha_j\}_{j=-L}^L$ and κ from Algorithm 2. Also recall the (generalized-)no-free-lunch axiom which mandates a zero payment if, in the gold standard, (all attempted questions are marked as the highest confidence L and) the answers to all the attempted questions are incorrect. The following proposition extends the results of the main text in the paper to this setting with utility functions.

Proposition 19 *For a worker who aims to maximize function U of the payment, the one and only mechanism that is incentive-compatible and satisfies the (generalized-)no-free-lunch axiom is*

$$\text{Payment}(x_1, \dots, x_G) = U^{-1} \left(\kappa \prod_{i=1}^G \alpha_{x_i} + U(\mu_{\min}) \right),$$

where the constants $\{\alpha_j\}_{j=-L}^L$ are as defined in Algorithm 2 and $\kappa = (U(\mu_{\max}) - U(\mu_{\min}))\alpha_L^{-G}$.

Note that for the problem to be well defined, the interval $[\mu_{\min}, \mu_{\max}]$ should be contained in the interval \mathcal{I} . The proof of Proposition 19 follows easily from the results proved earlier in the paper, and is provided below for completeness.

Proof of Proposition 19. We will first verify that the proposed payment is always non-negative and satisfies the (generalized-)no-free-lunch axiom. Recall from Theorem 7 that for every evaluation $\{x_1, \dots, x_G\}$ for which the (generalized-)no-free-lunch axiom mandates a zero payment, the value of $\kappa \prod_{i=1}^G \alpha_{x_i}$ is zero. It follows that the payment $U^{-1} \left(\kappa \prod_{i=1}^G \alpha_{x_i} + U(\mu_{\min}) \right) = U^{-1}(0 + U(\mu_{\min})) = \mu_{\min}$, where the final equation is a consequence of the invertibility of U . Further, recall that the value of $\kappa \prod_{i=1}^G \alpha_{x_i}$ in Algorithm 2 is never smaller than zero. Since the function U is increasing, so is U^{-1} , and hence the payment is always non-negative.

We will now prove that the proposed payment is incentive-compatible. To this end, observe that the utility of the proposed payment is

$$\begin{aligned} U(\text{Payment}) &= U \left(U^{-1} \left(\kappa \prod_{i=1}^G \alpha_{x_i} + U(\mu_{\min}) \right) \right) \\ &= \kappa \prod_{i=1}^G \alpha_{x_i} + U(\mu_{\min}). \end{aligned}$$

Noting that $U(0)$ is a constant independent of the worker’s answers, the result of Theorem 7 implies that the expectation of $U(\text{Payment})$ behaves exactly as required for incentive-compatibility.

We will now prove uniqueness of this mechanism. Replacing $f(\cdot)$ by $U(\text{Payment}(\cdot))$ in the proof of Theorem 9, we get that the function $U(\text{Payment})$ must be of the form

$$U(\text{Payment}(x_1, \dots, x_G)) = c_1 \prod_{i=1}^G \alpha_{x_i} + c_2,$$

for some constants c_1 and c_2 , where $\{\alpha_{x_j}\}_{j=-L}^L$ are as defined in Algorithm 2. In other words, the payment must be of the form

$$\text{Payment}(x_1, \dots, x_G) = U^{-1} \left(c_1 \prod_{i=1}^G \alpha_{x_i} + c_2 \right).$$

One can evaluate that the maximum value of this payment is $c_1 + c_2$. From our μ_{\max} -budget constraint, we then have $c_1 + c_2 = \mu_{\max}$. Furthermore, When the evaluations x_1, \dots, x_G are such that the (generalized-)no-free-lunch applies, we need $\text{Payment} = \mu_{\min}$. It follows that $c_2 = U(\mu_{\min})$, and consequently $c_1 = U(\mu_{\max}) - U(\mu_{\min})$, thereby completing the proof.

Bibliography

- Nima Anari, Gagan Goel, and Afshin Nikzad. Mechanism design for crowdsourcing: An optimal 1-1/e competitive budget-feasible mechanism for large markets. In *Foundations of Computer Science (FOCS)*, pages 266–275, 2014.
- Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.
- Jason Baldridge and Alexis Palmer. How well does active learning actually work?: Time-based evaluation of cost-reduction strategies for language documentation. In *Conference on Empirical Methods in Natural Language Processing*, pages 296–305, 2009.
- Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. Soyent: a word processor with a crowd inside. In *ACM symposium on User interface software and technology (UIST)*, pages 313–322, 2010.
- John Bohannon. Social science for pennies. *Science*, 334(6054):307–307, 2011.
- Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.
- Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, pages 477–505, 2007.
- Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*, 2005.
- Yang Cai, Constantinos Daskalakis, and Christos H Papadimitriou. Optimum statistical estimation with strategic data sources. In *Conference on Learning Theory (COLT)*, 2015.

- Izquierdo JM Cano, Yannis A Dimitriadis, Sánchez E Gómez, and Coronado J López. Learning from noisy information in fasart and fasback neuro-fuzzy systems. *Neural networks: the official journal of the International Neural Network Society*, 14(4-5):407–425, 2001.
- Andrew Carlson, Justin Betteridge, Richard C Wang, Estevam R Hruschka Jr, and Tom M Mitchell. Coupled semi-supervised learning for information extraction. In *ACM international conference on Web search and data mining*, pages 101–110, 2010.
- Jenny J Chen, Natala J Menezes, Adam D Bradley, and TA North. Opportunities for crowdsourcing research on Amazon mechanical turk. *Interfaces*, 5(3), 2011.
- Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *ACM international conference on Web search and data mining*, pages 193–202, 2013.
- Xi Chen, Sivakanth Gopi, Jieming Mao, and Jon Schneider. Competitive analysis of the top-K ranking problem. *arXiv preprint arXiv:1605.03933*, 2016.
- Fang Chu, Yizhou Wang, and Carlo Zaniolo. An adaptive learning approach for noisy data streams. In *IEEE International Conference on Data Mining (ICDM)*, pages 351–354, 2004.
- Vincent Conitzer. Prediction markets, mechanism design, and cooperative game theory. In *Uncertainty in Artificial Intelligence (UAI)*, pages 101–108, 2009.
- Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *International conference on World Wide Web (WWW)*, pages 319–330, 2013.
- Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics*, pages 20–28, 1979.
- Ofer Dekel, Felix Fischer, and Ariel D Procaccia. Incentive compatible regression learning. In *ACM-SIAM symposium on Discrete algorithms (SODA)*, pages 884–893, 2008.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.
- Double or Nothing. http://wikipedia.org/wiki/Double_or_nothing, 2014. Last accessed: July 31, 2014.
- Fang Fang, Maxwell Stinchcombe, and Andrew Whinston. Putting your money where your mouth is: A betting platform for better prediction. *Review of Network Economics*, 6(2), 2007.
- Michael J Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. CrowdDB: answering queries with crowdsourcing. In *ACM SIGMOD International Conference on Management of Data*, pages 61–72, 2011.
- Tilman Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

- Steve Hanneke and Liu Yang. Negative results for active learning with convex losses. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 321–325, 2010.
- Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.
- Chien-Ju Ho, Shahin Jabbari, and Jennifer W Vaughan. Adaptive task assignment for crowdsourced classification. In *International Conference on Machine Learning (ICML)*, pages 534–542, 2013.
- Panagiotis G Ipeirotis, Foster Provost, Victor S Sheng, and Jing Wang. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, 28(2):402–441, 2014.
- Srikanth Jagabathula, Lakshminarayanan Subramanian, and Ashwin Venkataraman. Reputation-based worker filtering in crowdsourcing. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2492–2500, 2014.
- W Paul Jones and Scott A Loe. Optimal number of questionnaire response categories more may not be better. *SAGE Open*, 3(2):2158244013489691, 2013.
- David R Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems (NIPS)*, 2011.
- Gabriella Kazai, Jaap Kamps, Marijn Koolen, and Natasa Milic-Frayling. Crowdsourcing for book search evaluation: impact of HIT design on comparative system ranking. In *ACM SIGIR conference on Research and development in Information Retrieval*, pages 205–214, 2011.
- Firas Khatib, Frank DiMaio, Seth Cooper, Maciej Kazmierczyk, Mirosław Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, Mariusz Jaskolski, and David Baker. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177, 2011.
- Ashish Khetan and Sewoong Oh. Reliable crowdsourcing under the generalized dawid-skene model. *arXiv preprint arXiv:1602.03481*, 2016.
- Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-fei Li. L.: Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, CVPR*, 2011.
- Nicolas Lambert and Yoav Shoham. Eliciting truthful answers to multiple-choice questions. In *ACM conference on Electronic commerce*, pages 109–118, 2009.
- ASID Lang and Joshua Rio-Ross. Using Amazon Mechanical Turk to transcribe historical handwritten documents. *The Code4Lib Journal*, 2011.
- Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005.

- John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, pages 21–26, 2010.
- Eric WM Lee, Chee Peng Lim, Richard KK Yuen, and SM Lo. A hybrid neural network model for noisy data regression. *IEEE Transactions on Systems, Man, and Cybernetics*, 34(2):951–960, 2004.
- Qiang Liu, Jian Peng, and Alexander T Ihler. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems (NIPS)*, pages 701–709, 2012.
- Philip M Long and Rocco A Servedio. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78(3):287–304, 2010.
- Naresh Manwani and PS Sastry. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics*, 43(3):1146–1151, 2013.
- David Mease, Abraham J Wyner, and Andreas Buja. Boosted classification trees and class probability/quantile estimation. *The Journal of Machine Learning Research*, 8:409–439, 2007.
- George A Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.
- Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.
- Dražen Prelec. A Bayesian truth serum for subjective data. *Science*, 306(5695):462–466, 2004.
- Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *The Journal of Machine Learning Research (JMLR)*, 11:1297–1322, 2010.
- Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.
- Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *arXiv preprint arXiv:1512.08949*, 2015.
- Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Journal of Machine Learning Research (JMLR)*, 2016a.
- Nihar B Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *International Conference on Machine Learning (ICML)*, 2016b.
- Nihar B Shah, Sivaraman Balakrishnan, and Martin J Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *arXiv preprint arXiv:1606.09632*, 2016c.
- Richard M Shiffrin and Robert M Nosofsky. Seven plus or minus two: A commentary on capacity limitations. *Psychological Review*, 101(2):357–61, 1994.

- Aditya Vempaty, Lav R Varshney, and Pramod K Varshney. Reliable crowdsourcing for multi-class labeling using coding theory. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):667–679, 2014.
- Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321(5895): 1465–1468, 2008.
- Jeroen Vuurens, Arjen P de Vries, and Carsten Eickhoff. How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In *ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, pages 21–26, 2011.
- Paul Wais, Shivaram Lingamneni, Duncan Cook, Jason Fennell, Benjamin Goldenberg, Daniel Lubarov, David Marin, and Hari Simons. Towards building a high-quality workforce with Mechanical Turk. *NIPS workshop on computational social science and the wisdom of crowds*, 2010.
- Fabian L Wauthier and Michael Jordan. Bayesian bias mitigation for crowdsourcing. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.
- Justin Wolfers and Eric Zitzewitz. Prediction markets. Technical report, National Bureau of Economic Research, 2004.
- Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. Task matching in crowdsourcing. In *IEEE International Conference on Cyber, Physical and Social Computing*, pages 409–412, 2011.
- Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- Dengyong Zhou, John Platt, Sumit Basu, and Yi Mao. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2204–2212, 2012.
- Dengyong Zhou, Qiang Liu, John C Platt, Christopher Meek, and Nihar B Shah. Regularized minimax conditional entropy for crowdsourcing. *arXiv preprint arXiv:1503.07240*, 2015.
- Yuan Zhou, Xi Chen, and Jian Li. Optimal PAC multiple arm identification with applications to crowdsourcing. In *International Conference on Machine Learning (ICML)*, pages 217–225, 2014.