Optimal Stopping and Worker Selection in Crowdsourcing: an Adaptive Sequential Probability Ratio Test Framework

Xiaoou Li[1], Yunxiao Chen[2], Xi Chen[3], Jingchen Liu[4], and Zhiliang Ying[4]

University of Minnesota[1], London School of Economics and Political Science[2], New York University[3], and Columbia University[4]

Abstract: In this paper, we aim at solving a class of multiple testing problems under the Bayesian sequential decision framework. Our motivating application comes from binary labeling tasks in crowdsourcing, where a requestor needs to simultaneously decide which worker to choose to provide a label and when to stop collecting labels under a certain budget constraint. We start with a binary hypothesis testing problem to determine the true label of a single object, and provide an optimal solution by casting it under the adaptive sequential probability ratio test (Ada-SPRT) framework. We characterize the structure of the optimal solution, i.e., the optimal adaptive sequential design, which minimizes the Bayes risk by making use of a log-likelihood ratio statistic. We also develop a dynamic programming algorithm that can efficiently compute the optimal solution. For the multiple testing problem, we further propose to adopt an empirical Bayes approach for estimating class priors and show that our method has an averaged loss which converges to the minimal Bayes risk under the true model. The experiments on both simulated and real data show the robustness of our method and its superiority in labeling accuracy comparing with several other recently proposed approaches.

Key words and phrases: Bayesian decision theory, Crowdsourcing, Empirical Bayes, Sequential analysis, Sequential probability ratio test

1. Introduction

Crowdsourcing – as an emerging technology for data-intensive tasks – leverages a "large group of people in the form of an open call" to achieve a cumulative result (Howe, 2006). Over the past ten years, crowdsourcing has become an efficient and economical approach to obtaining labels for tasks that are difficult for computers but easy for humans. For example, the requestor can post a large number of images on a popular crowdsourcing platform (e.g., Amazon Mechanical Turk) and ask a crowd of workers to tag each picture as a portrait or a landscape with a small amount of payment for each label. The crowdsourcing technique has helped address a wide range of challenges in scientific areas, e.g., enabling an understanding of the evolution of galaxies via the crowd classifying galaxy morphology (Galaxy Zoo (Raddick et al., 2010)), and aiding the diagnosis of malaria epidemics by asking crowd workers to identify malaria-infected red blood cells (MOLT (Mavandadi et al., 2012) and MalariaSpot (Luengo-Oroz et al., 2012)). Interested readers may refer to Doan et al. (2011), Slivkins and Vaughan (2013), and Marcus and Parameswaran (2015) for more comprehensive reviews of crowdsourcing techniques and their applications.

Despite its efficiency and immediate availability, the labels generated by non-expert crowd workers are quite noisy. For example, as reported in Yalavarthi et al. (2017) and Ke et al. (2018), "even considering answers from workers with high-accuracy statistics in Amazon Mechanical Turk, we find that the average crowd error rate can be up to 25%". As a remedy, most requestors resort to repetitive labeling for each object (e.g., an image), i.e., collecting multiple labels from different workers for a single object. Then, the requestor aggregates the collected labels to infer its true label. Generally, more labels for an object will lead to higher accuracy for the inferred label. However,

each label comes with a fixed amount of cost: the requestor has to pay a pre-specified monetary cost for each obtained label, regardless of its correctness. Therefore, when using crowdsourcing service for large-scale labeling tasks, a requestor usually faces two challenges:

1. The requestor needs to carefully balance between the labeling accuracy and the cost of collecting labels. That is, for each object, the requestor needs to decide when to stop collecting the next label based on the current information.

2. The crowd workers have different levels of quality/reliability. The requestor needs to adaptively choose the next worker to label the object based on the current information.

To address these challenges, we cast the problem into a general multiple testing problem under a sequential analysis framework. In particular, we study the most popular crowdsourcing task, binary labeling tasks, e.g., categorization of an image as a portrait or a landscape or a website as pornography or not. We assume that there are K objects and for each object, we are interested in testing whether its true label (denoted by $\theta_k \in \{0, 1\}$) belongs to class zero or one. More specifically, this problem can be formulated into K hypothesis testing problems:

$$H_{k0} : \theta_k = 0 \quad \text{against} \quad H_{k1} : \theta_k = 1, \quad \text{for } \ k = 1, 2, ..., K. \tag{1.1}$$

Since the true classes of objects might be highly unbalanced, it is natural to assume that there is a prior π_1 (and $\pi_0 = 1 - \pi_1$) such that

$$\pi_0 = \mathbb{P}(\theta_k = 0) \quad \text{and} \quad \pi_1 = \mathbb{P}(\theta_k = 1), \quad \text{for } \ k = 1, \ldots, K. \tag{1.2}$$

The parameter π_1 models the unbalancedness between two classes, which is usually unknown. To study this problem, let us first assume that π_1 is known and consider the following hypothesis testing problem:

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta = 1. \tag{1.3}$$

To solve this problem, we propose an adaptive sequential probability ratio test (Ada-SPRT) under the Bayesian sequential analysis framework. We first formulate a risk function which is the expected probability of making the wrong decision plus the expected labeling cost. There are three components that we need to optimize over in this Ada-SPRT procedure:

1. Stopping time: The first question is when to stop collecting more data (i.e., labels) under a certain budget constraint (e.g., given a pre-specified maximum number of labels that can be collected). Early stopping is important for cost-effective crowdsourcing since a requestor should stop if the labels collected so far have already reached good consensus in order to avoid unnecessary cost.

2. Adaptive experimental selection rule: We assume that there are M possible experiments (corresponding to heterogeneous workers), where different experiments lead to different distributions for generating data (i.e., labels) under the true θ. The key question is how to select the next experiment given existing data.

3. Decision rule: Upon stopping, we need to decide whether H_0 or H_1 is true.

It is worth noting that our Ada-SPRT can be viewed as an extension of the classical SPRT by Wald (1945) and Wald and Wolfowitz (1948), which only optimizes the stopping time and decision rule, without considering experiment selection.

In the sequential analysis literature, Chernoff (1959) and many follow-up works have provided asymptotically optimal solutions for various sequential design problems (see Section 2 for more details). However, the classical asymptotic regime is not suitable for our problem for two reasons:

1. In crowdsourcing applications, a requestor usually has a limited budget (e.g., at most 10 labels for each object), which translates into an upper bound on stopping time, i.e., a truncation length. Under this constraint, the sample size cannot go to infinity and thus theory from asymptotically optimal experimental design will not hold anymore.

2. There is a class prior distribution π_1 in (1.2), which needs to be estimated and plays an important role in our problem given a small truncation length. However, for the classical asymptotically optimal results, the effect of the prior probability distribution is usually ignored as the expected sample size goes to infinity.

To address these challenges and to solve the general multiple testing problem in (1.1), we propose an *empirical Bayes approach* with a *dynamic programming algorithm* to solve the single hypothesis testing problem in (1.3) with a pre-specified truncation length T. For a single truncated test, the sequential decision problem can be formulated into a Markov decision process (MDP) problem, where the state space is characterized by a log-likelihood ratio statistic and the current sample size. To solve this MDP, we first provide a few structural results:

1. The *optimal stopping time* is a boundary hitting time based on the *log-likelihood ratio* statistic. The upper boundary curve is non-increasing with respect to

(w.r.t.) the sample size $n = 1, \ldots, T$ and the lower boundary curve is non-decreasing w.r.t. the sample size n.

2. The *optimal decision* for the true label is according to whether the *log-likelihood ratio* hits the upper or lower boundary.

3. The *experiment/worker selection rule* is determined by the current *log-likelihood ratio* and the sample size.

With these structural results, we develop a dynamic programming algorithm for solving the MDP. We also characterize the relationship between the simpler non-truncated test (i.e., the truncation length $T = \infty$) and the truncated test and show that one can treat the non-truncated test as a limiting version of the truncated test as T goes to infinity.

With the Ada-SPRT for solving (1.3) in place, we solve the multiple testing problem in (1.1) using an empirical Bayes approach that estimates the class prior π_1. We prove that as long as the class prior estimate is consistent, the averaged loss will converge to the minimal Bayes risk under the true model. We further demonstrate its superior performance and robustness against different setups of true prior distribution (e.g., unbalanced class setting) using empirical studies.

Finally, we highlight that although our paper is motivated from a crowdsourcing application, the proposed empirical Bayes with the Ada-SPRT approach is a general method for solving the multiple testing problem in (1.1). The proposed method can be applied to a wide class of problems. For example, computerized mastery testing (Lewis and Sheehan, 1990; Chang, 2004, 2005; Bartroff et al., 2008), which is based on item response theory models (e.g. Embretson and Reise, 2000) and aims to classify examinees into "mastery" and "non-mastery" categories, has become an important

testing mode in educational assessment. The Ada-SPRT could be extended to provide an optimal adaptive mastery test design (in terms of Bayes risk).

The rest of the paper is organized as follows. In Section 2, we discuss related works in the literature of crowdsourcing, sequential analysis, and empirical Bayes. In Section 3, we present the crowdsourcing model and the Bayesian decision framework, along with our Bayes risk function. In Section 4, we provide the optimal adaptive sequential design, and develop numerical algorithms for optimal worker selection, stopping time and decision for both truncated and non-truncated tests. In Section 5, we extend the algorithm to the multiple testing problem and present an empirical Bayes approach which estimates class priors. In Section 6, we demonstrate the performance of the proposed algorithm on real crowdsourcing datasets, followed by discussions in Section 7. The proofs of the theoretical results and simulated experiments are provided as supplementary material.

2. Related Works

Crowdsourcing, as the most popular paradigm for effectively collecting labels at low cost, has received a great deal of attention from researchers in statistics and machine learning communities. Many works in this field are solving a static problem, i.e., inferring true labels and workers' quality parameters based on a static set of labels (see, e.g., Raykar et al. (2010); Karger et al. (2013); Liu et al. (2012); Gao and Zhou (2013); Ertekin et al. (2014); Zhang et al. (2016); Khetan and Oh (2016); Shah et al. (2016); Ok et al. (2016)). Most of these works adopt the Dawid-Skene model (Dawid and Skene, 1979), which is also known as the two-coin model for binary labeling tasks, for modeling workers' quality. In the rest of the model, we shall assume the Dawid-Skene model.

We also note that some recent works (e.g., Shah et al. (2016); Khetan and Oh (2016)) study more general models such as generalized Dawid-Skene model and permutation model. For the adaptive worker selection problem, there are relatively fewer results in the existing literature. Karger et al. (2013) proposed to assign workers according to a random bipartite graph. Chen et al. (2015) considered the fixed budget problem and formulated the problem into a Bayesian MDP. They studied two greedy policies to approximately solve the MDP: (1) the knowledge gradient (KG) policy, which chooses the best experiment/action that maximizes the expected reward for the next stage; (2) the optimistic knowledge gradient (Opt-KG) policy, which chooses the best action that maximizes the maximum of the reward for collecting a positive label and that for a negative label. We will compare with these two greedy policies in our experiments (see Section 6 for details). Further, Ertekin et al. (2014) proposed a confidence-score based algorithm called CrowdSense for budget allocation. Khetan and Oh (2016) investigated the sample complexity (minimum expected number of labels) for the classification error less than a small threshold with high probability.

Note that instead of pre-fixing a total sampling budget as in some works (e.g., Chen et al. (2015)), our goal is to simultaneously conduct worker selection and make the optimal decision on stopping time. To achieve this goal, we formulate the problem into a Bayesian sequential testing problem and propose an adaptive sequential probability ratio test (Ada-SPRT) framework. Sequential testing, starting with the seminal works of Wald (1945) and Wald and Wolfowitz (1948) for testing two simple hypotheses, is one of the most classical and well-studied problems in sequential analysis. We refer readers to the survey article (Lai, 2001) and books (Siegmund, 1985; Tartakovsky et al., 2014) for a comprehensive review. Sequential tests have received a wide range of

applications in areas such as industrial quality control, design of clinical trials, finance, educational testing, etc (Lai, 2001; Bartroff and Lai, 2008; Bartroff et al., 2008, 2013; Tartakovsky et al., 2014). The problem of sequential adaptive experiment selection was initially treated in Chernoff (1959), which considers a Bayes risk that is defined similarly to that in Wald and Wolfowitz (1948). Another related work is Robbins and Siegmund (1974), which presents Monte Carlo and theoretical analysis on several adaptive treatment selection rules in clinical trial, with the aim to reduce the expected number of observations made on the inferior treatment. The current work provides theoretical results in the sequential hypothesis testing framework that simultaneously considers the optimality of stopping, decision, and experiment selection.

The current work is also related to the multi-armed bandit problem (Robbins, 1952), which has been studied in many areas, such as clinical trials (Press, 2009), online advertising (Chakrabarti et al., 2009; Babaioff et al., 2009), and portfolio design (Hoffman et al., 2011). We refer the readers to the work by Lai (1987); Auer et al. (2002a,b); Li et al. (2010) and the survey paper by Bubeck and Cesa-Bianchi (2012). In a typical stochastic multi-armed bandit problem, there are n alternative arms, where each arm is associated with an unknown reward distribution. Upon pulling a particular arm, the reward is an *i.i.d.* sample from the underlying reward distribution. One needs to sequentially decide which arm to pull next and then collect the random reward. The goal is to maximize the expected collected total reward over a finite time horizon. Similar to the stochastic bandit setting, our problem also involves sequential decision-making on the worker selection. However, there are two unique challenges in our problem, which prevents the direct application of existing bandit algorithms from machine learning literature. First, bandit literature usually assumes an intermediate reward after each

action and the goal is to maximize the total rewards (or discounted rewards over time). In our problem, each answer from a worker provides some information. However, due to the noise of an answer, its usefulness only vaguely relates to the final testing error. Therefore, there is no clear intermediate "reward" associated with a new sample. Second, instead of fixing the length of the time horizon, we optimize over the random stopping time. In crowdsourcing applications, the optimal stopping provides a flexible trade-off between learning accuracy and cost, that can be controlled by the relative cost parameter c in our objective function; see equation (3.5) below.

Another related subject is the design problem for A/B testing (Bhat et al., 2019; Johari et al., 2015), which is a form of randomized controlled trials aiming at comparing the treatment effects of different groups. It is possible to apply the proposed adaptive sequential method to designing an A/B testing strategy. Comparing with the existing works in this area, on one hand, our method incorporates an early stopping strategy, and provides an optimal design under the Dawid-Skene model. On the other hand, our method does not take the covariate information into account (e.g., comparing with Bhat et al. (2019)). In crowdsourcing applications, personal information of workers is usually quite sensitive and might not be readily available. However, contextual information may be available and useful in other applications and adding contextual information into the current sequential framework can be an interesting future direction.

Empirical Bayes method has recently gained much prominence, in both theory and applications (e.g. Jiang and Zhang, 2009, 2010; Koenker and Mizera, 2014; Brown and Greenshtein, 2009; Efron, 2013). We refer to Zhang (2003), Efron (2013), and the references therein for a comprehensive review. In particular, Karunamuni (1988) combines the empirical Bayes method and sequential analysis, and then provide theoretical

analysis for the asymptotic behavior of a specific stopping rule. The current work extends this idea to an optimal design of experiment selection and early stopping. To the authors' best knowledge, this is the first result encompassing empirical Bayes method, sequential analysis and experiment selection simultaneously.

To highlight our contribution, we compare the proposed method on adaptive sequential testing with the existing ones, which, in general, fall into one of the three classes: 1) sequential hypothesis testing with an adaptive sequential design in an asymptotic regime; 2) sequential hypothesis testing without an adaptive design in a non-asymptotic regime; 3) sequential hypothesis testing with an adaptive design in a non-asymptotic regime. The major differences between our work and the existing methods are summarized below.

1) The hypothesis testing with a sequential design in an asymptotic regime was first studied in Chernoff (1959), followed by a large body of literature including Albert (1961); Tsitovich (1985); Naghshvar and Javidi (2013a,b); Bessler (1960); Nitinawarat and Veeravalli (2015). This line of research focuses on the behavior of sequential designs when their expected sample sizes grow large. Asymptotically optimal properties for different procedures have been derived. Motivated from the crowdsourcing application, we consider a different regime where the sample size is not allowed to go to infinity (fixed the cost c and with a maximum test length constraint T). Thus, methods and techniques for the asymptotic regime are not applicable to our problem.

2) The sequential hypothesis testing in a non-asymptotic regime was first considered by Wald and Wolfowitz (1948), followed by a vast literature including Wald (1947), Wald and Wolfowitz (1950), Sobel and Wald (1949), Arrow et al. (1949),

Bussgang and Middleton (1955), Irle and Schmitz (1984), Nikiforov (1975), Bertsekas and Shreve (1978), and Shiryaev (1978). Under the non-asymptotic regime, SPRT is shown to be optimal from a non-Bayesian point of view (Wald and Wolfowitz, 1948). The optimal truncated and non-truncated Bayesian sequential tests have been developed in Arrow et al. (1949).

Theorem 1 of this paper extends results for the optimal Bayesian sequential test in Arrow et al. (1949) by incorporating an adaptive design (i.e., enabling the adaptive selection of the next experiment based on the current information).

3) The study of general stochastic control problem under the non-asymptotic regime dates back to Bertsekas and Shreve (1978), Bellman (1957), and Shiryaev (1978). Recent works, including Bai and Gupta (2016) and Naghshvar and Javidi (2010), establish theoretical properties of the optimal procedure for specific hypothesis testing problems with experiment design, under the non-asymptotic regime. In particular, Naghshvar and Javidi (2010) consider the problem of a single non-truncated sequential test with $M \geq 2$ hypotheses (among which only one holds true). In this paper, we study a multiple testing problem and develop an empirical Bayes approach. For each single test with $M = 2$ hypotheses, we provide a refined result on the continuation region either with or without a maximum test length constraint.

3. Model and Problem Setup

In this section, we first introduce the problem setup with full generality, followed by a specific application to crowdsourcing.

3.1 Problem Setup

For ease of exposition, we first consider the case where there is only one object with a known prior probability, and then extend the result to the case where there are K objects. For a single object with true label $\theta \in \{0, 1\}$, we are interested in the hypothesis testing problem in (1.3). Let $X_1, X_2, \ldots$ be the observed responses. The selection of the n-th *experiment* depends on all the previous responses. In particular, let $I = \{1, \ldots, M\}$ be the *experiment pool* and $\delta_n \in I$ be the selected n-th experiment, we have $\delta_n = j_n(X_1, ..., X_{n-1})$, where function $j_n(\cdot)$ is the experiment selection rule that needs to be learned. We use J to denote the sequence of experiment selection rule $\{j_n : n = 1, 2, \ldots\}$.

Given θ and δ_n, we denote the probability mass or density function of $X_n \in \mathbb{R}^d$ by f_{θ,δ_n}. We make the assumption that there exists at least one experiment $\delta \in I$ such that the Kullback-Leibler divergence is bounded away from zero and infinity, i.e.,

$$0 < \mathbb{E}\left[\log \frac{f_{0,\delta}(X)}{f_{1,\delta}(X)} \Big| \theta = 0\right] < \infty \text{ and } 0 < \mathbb{E}\left[\log \frac{f_{1,\delta}(X)}{f_{0,\delta}(X)} \Big| \theta = 1\right] < \infty.$$

Here, X is a generic notation for an observation with the probability mass or density function $f_{\theta,\delta}(x)$. Under this assumption, the model is identifiable and the standard SPRT has a finite expected sample size. We would like to point out that our results are applicable to both continuous and discrete observations.

We further consider a random sample size denoted by N, that is, the test stops once sufficient observations have been collected. We consider the case that there is a deterministic upper bound, or truncation length T, on the stopping time, that is, $N \leq T$. Given all the responses and the stopping rule, one is able to decide whether

to continue collecting at least one more response or to stop the test. Upon stopping, one is able to decide whether H_0 or H_1 to choose. We denote by D the decision rule, where $D = 1$ represents H_1 is chosen while $D = 0$ means H_0 is chosen.

The test procedure that has an experiment selection rule J, a stopping rule N, and a decision rule D is called as an *adaptive sequential design*. Our goal is to search for the optimal $J^\dagger$, $N^\dagger$ and $D^\dagger$ to minimize the composite risk of making a wrong test decision and the expected total labeling cost as defined below.

To define the risk, we adopt the *Bayesian decision framework*. In particular, we introduce the class prior

$$\pi_0 = \mathbb{P}(\theta = 0) \quad \text{and} \quad \pi_1 = \mathbb{P}(\theta = 1), \tag{3.4}$$

with $\pi_0 + \pi_1 = 1$. We assume that π_1 is known for the single hypothesis testing problem since it is impossible to estimate π_1 when there is only one object. Let $c \in [0, 1]$ be the *relative cost* of collecting one response/label. The Bayes risk of an adaptive sequential test with experiment selection rule J, stopping time N and decision rule D is defined by Wald and Wolfowitz (1948) as the expected probability of making wrong decision plus expected labeling cost,

$$\begin{aligned} \mathbf{R}(J, N, D) =& \pi_0 \mathbb{P}(D = 1|\theta = 0) + \pi_1 \mathbb{P}(D = 0|\theta = 1) \\ &+ c\{\pi_0 \mathbb{E}(N|\theta = 0) + \pi_1 \mathbb{E}(N|\theta = 1)\}. \end{aligned} \tag{3.5}$$

We note that the *relative cost* c, which is used to balance the trade-off between the labeling accuracy and labeling cost, needs to be set between zero and one. Since $\mathbb{P}(D = 1|\theta = 0) \leq 1$ and $\mathbb{P}(D = 0|\theta = 1) \leq 1$, to minimize the Bayes risk in (3.5),

one will not collect any label when $c > 1$. In practice, the requestor usually chooses c depending on the nature of labeling tasks (e.g., smaller c for more challenging data to collect more labels) and the availability of the budget (e.g., a large c for very limited amount of budget). We will demonstrate the effect of c in our experiments in Section 6.

We denote by $\mathcal{A}^T$ the set of all adaptive sequential designs (J, N, D) such that the stopping time $N \leq T$. We call the test procedure $(J^\dagger, N^\dagger, D^\dagger)$ an *optimal test* among a class of adaptive sequential testing procedure $\mathcal{A}^T$ (depending on the truncation length T) if

$$\mathbf{R}(J^\dagger, N^\dagger, D^\dagger) = \min_{(J,N,D)\in\mathcal{A}^T} \mathbf{R}(J, N, D). \tag{3.6}$$

Now, for K objects with true label $\theta_k \in \{0, 1\}$ for $1 \leq k \leq K$, we consider K hypothesis testing problems. Given $\theta_1, ..., \theta_K$, we assume that the responses $(X_{kj}; j = 1, 2, ..)$ obtained for the k's object are independent across different k. Let $D = \{D_k\}_{k=1}^K$ be the set of decisions and $N = \{N_k\}_{k=1}^K$ be the set of stopping times. The performance of the method is evaluated by the following *averaged loss* defined over K objects:

$$L_K = \frac{1}{K} \sum_{k=1}^{K} \left[\mathbf{1}_{\{D_k \neq \theta_k\}} + cN_k \right]. \tag{3.7}$$

Our goal is to provide a *consistent procedure* such that L_K converges to the minimal Bayes risk under the true model (i.e., $\min_{(J,N,D)\in\mathcal{A}^T} \mathbf{R}(J, N, D)$) in probability as K goes to infinity.

3.2 Applications to Crowdsourcing

Here, we briefly illustrate how this general sequential testing framework is connected to our motivating crowdsourcing application. We assume that there are M workers (i.e., experiments) and we denote the set of workers by $I = \{1, \ldots, M\}$, which is our experiment pool.

For an object with the true label $\theta \in \{0, 1\}$, let $\widehat{\theta^i}$ be the label provided by worker i, $i \in I$. The quality of worker i is characterized by two quantities:

$$\tau_{00}^i = \mathbb{P}(\widehat{\theta^i} = 0 | \theta = 0) \quad \text{and} \quad \tau_{11}^i = \mathbb{P}(\widehat{\theta^i} = 1 | \theta = 1). \tag{3.8}$$

In other words, τ_{00}^i is the probability that worker i will provide the correct label to an object when the true label is zero and τ_{11}^i is that when the true label is one. This model is widely used in modeling crowd worker quality and is usually referred to as "two-coin model" or Dawid-Skene model (Dawid and Skene, 1979; Raykar et al., 2010; Zhang et al., 2016). For ease of presentation, we assume that τ_{00}^i and τ_{11}^i are given and will discuss, in Section 5, how to estimate these parameters in an online fashion as the labeling process goes on.

The observed responses X_n for $n = 1, 2, \ldots$ are the labels from the selected n-th worker δ_n according to the worker selection rule $j_n(X_1, ..., X_{n-1})$. Under the two-coin model in (3.8), each response takes the binary value, with the following probability mass function:

$$\begin{aligned} f_{\theta,\delta_n}(1) &= \mathbb{P}(X_n = 1 | \delta_n, \theta) = \tau_{11}^{\delta_n} \, \mathbf{1}_{\{\theta=1\}} + (1 - \tau_{00}^{\delta_n}) \, \mathbf{1}_{\{\theta=0\}}, \\ f_{\theta,\delta_n}(0) &= \mathbb{P}(X_n = 0 | \delta_n, \theta) = (1 - \tau_{11}^{\delta_n}) \, \mathbf{1}_{\{\theta=1\}} + \tau_{00}^{\delta_n} \, \mathbf{1}_{\{\theta=0\}}, \end{aligned} \tag{3.9}$$

where $\mathbf{1}_{\{.\}}$ denotes the indicator function.

4. Optimal Adaptive Sequential Probability Ratio Test

In this section, we explore the structure of the optimal adaptive sequential design for the single hypothesis testing problem in (1.3) and provide a dynamic programming algorithm to numerically solve an optimal adaptive sequential design problem.

4.1 Structure of Optimal Adaptive Sequential Designs

We consider the class of truncated adaptive sequential tests with the constraint that the sample size N is no greater than a pre-fixed truncation length T. The optimization problem (3.6) is challenging because both the experiment selection and the stopping rule lie in infinite-dimensional function spaces. Our approach is to make dimension reduction by exploring the relationship between optimal adaptive sequential design and a *log-likelihood ratio statistic*.

In particular, under the optimal selection rule $J^\dagger = \{j_1^\dagger, j_2^\dagger, \dots, \}$, the n-th selected experiment (for $n \leq N^\dagger$) is

$$\delta_n^\dagger = j_n^\dagger(X_1, \dots, X_{n-1}).$$

The corresponding log-likelihood ratio statistic is defined by

$$l_n^\dagger = \log\left(\frac{\prod_{i=1}^n f_{1,\delta_i^\dagger}(X_i)}{\prod_{i=1}^n f_{0,\delta_i^\dagger}(X_i)}\right), \quad \text{for } n = 1, 2, \dots, \tag{4.10}$$

where $f_{1,\delta_i^\dagger}(\cdot)$ and $f_{0,\delta_i^\dagger}(\cdot)$ are the probability density/mass functions when $\theta = 1$ and $\theta = 0$ for the experiment $\delta_n^\dagger$, respectively. The next theorem characterizes the structure of the optimal adaptive sequential design.

Theorem 1. *Let* $(J^\dagger, N^\dagger, D^\dagger)$ *be an optimal adaptive truncated sequential design as defined in* (3.6)*. Then* $(J^\dagger, N^\dagger, D^\dagger)$ *has the following properties.*

(i) The stopping time $N^\dagger$ *is described through the hitting boundary of the log-likelihood ratio and the current sample size. In particular, there exist two sequences of real values* $A^\dagger(n)$ *and* $B^\dagger(n)$ *for* $1 \leq n \leq T$ *such that*

$$\log \frac{\pi_0}{\pi_1} = A^\dagger(T) \leq A^\dagger(T-1) \leq ... \leq A^\dagger(1) \leq \log \frac{\pi_0(1-c)}{\pi_1 c}, \tag{4.11}$$

$$\log \frac{\pi_0 c}{\pi_1(1-c)} \leq B^\dagger(1) \leq B^\dagger(2) \leq ... \leq B^\dagger(T) = \log \frac{\pi_0}{\pi_1}, \tag{4.12}$$

and the optimal stopping for the truncated test is determined by

$$N^\dagger = \inf\{n : l_n^\dagger \geq A^\dagger(n) \text{ or } l_n^\dagger \leq B^\dagger(n)\}. \tag{4.13}$$

(ii) If $N^\dagger < T$*, then the decision rule is*

$$D^\dagger = 1 \text{ if } l_{N^\dagger}^\dagger \geq A^\dagger(N^\dagger) \quad \text{and} \quad D^\dagger = 0 \text{ if } l_{N^\dagger}^\dagger \leq B^\dagger(N^\dagger).$$

If $N^\dagger = T$ *where* $A^\dagger(T) = B^\dagger(T)$*, then*

$$D^\dagger = 1 \text{ if } l_T^\dagger \geq A^\dagger(T) \quad \text{and} \quad D^\dagger = 0 \text{ if } l_T^\dagger < B^\dagger(T).$$

(iii) There exists an experiment selection function $j^\dagger : \mathbb{R} \times \{1, 2, ...\} \to I$ *such that for* $n = 1, 2, ..., T$*,*

$$\delta_n^\dagger = j^\dagger(l_{n-1}^\dagger, n),$$

where $\delta_n^\dagger$ is the n-th selected experiment under the optimal selection rule $J^\dagger$.

Remark 1. According to Corollary 8.5.1 in Bertsekas and Shreve (1978), the optimal sequential adaptive design $(J^\dagger, N^\dagger, D^\dagger)$ always exists (not necessarily unique) for the truncated test. We also note that the existence of the optimal design for non-truncated problems when $T = \infty$ (see Proposition 1 in the later section) is guaranteed by Corollary 9.17.1 in Bertsekas and Shreve (1978).

The proof of Theorem 1 is provided in the supplementary material. The statements (i) and (ii) are extensions of the seminal work of SPRT (Wald and Wolfowitz, 1948) to the case of adaptive experiment selection. In contrast to the classical SPRT where the hitting boundaries are flat, the hitting boundaries of the truncated adaptive test include a non-increasing curve (i.e., the upper boundary $A^\dagger(T) \leq A^\dagger(T-1) \leq ... \leq A^\dagger(1)$) and a non-decreasing curve (i.e., the lower boundary $B^\dagger(1) \leq B^\dagger(2) \leq ... \leq B^\dagger(T)$). Note that since $A^\dagger(T)$ and $B^\dagger(T)$ take the same value $\log \frac{\pi_0}{\pi_1}$, the optimal stopping time $N^\dagger$ defined in (4.13) automatically satisfies the constraint $N^\dagger \leq T$. The experiment selection rule depends on both the log-likelihood ratio statistic in (4.10) and the current sample size.

4.2 Dynamic Programming Algorithm

Given the structure of the optimal adaptive sequential design, we present a dynamic programming algorithm for finding the optimal experiment selection rule and the hitting boundaries.

To describe the algorithm, we first introduce some necessary notation. Let $G(l, n)$ be the conditional risk associated with the log-likelihood ratio l and the current sample size $n \in \{1, \ldots, T\}$. When the sample size n reaches the truncation length T, the

testing procedure has to stop. For each l, we have

$$G(l,T) = \min\{\pi(\theta=0|l), \pi(\theta=1|l)\} + Tc, \tag{4.14}$$

where $\pi(\theta=0|l)$ and $\pi(\theta=1|l)$ are the posterior probabilities under the current log-likelihood ratio l and $\min\{\pi(\theta=0|l), \pi(\theta=1|l)\}$ is the Bayes risk of making the wrong decision. The term Tc is the cost of collecting T responses. By the standard Bayesian decision theory (see e.g., Tartakovsky et al. (2014), §3.2.2.),

$$\pi(\theta=0|l) = \frac{\pi_0}{\pi_0 + \pi_1 e^l} \quad \text{and} \quad \pi(\theta=1|l) = \frac{\pi_1 e^l}{\pi_0 + \pi_1 e^l}.$$

Given the definition of $G(l,n)$, for any current sample size $n < T$ and log-likelihood ratio l, the optimal selection rule $j^\dagger(l, n+1)$ should choose the $(n+1)$-th experiment $\delta_{n+1} \in I$ to minimize the next stage expected conditional risk, i.e.,

$$j^\dagger(l, n+1) = \arg\min_{\delta \in I} \mathbb{E}_{l,\delta} G\left(l + \log\frac{f_{1,\delta}(X)}{f_{0,\delta}(X)},\ n+1\right), \tag{4.15}$$

where the expectation is taken with respect to the next response X when the next selected experiment is $\delta \in I$.

As an illustration, we present an example of computing $\mathbb{E}_{l,\delta} G\left(l + \log\frac{f_{1,\delta}(X)}{f_{0,\delta}(X)},\ n+1\right)$ when $n = T-1$ (corresponding to the first step in the dynamic programming algorithm). In particular, we consider the two-coin model in (3.9) and the case for the i-th experiment. That is, $\delta = i$. Then, we have

$$\log\frac{f_{1,i}(X)}{f_{0,i}(X)} = X\log\left(\frac{\tau_{11}^i}{1-\tau_{00}^i}\right) + (1-X)\log\left(\frac{1-\tau_{11}^i}{\tau_{00}^i}\right).$$

To compute the conditional expectation of interest, we also need

$$\mathbb{P}_{l,i}(X=1)=\pi(\theta=0|l)f_{0,i}(1)+\pi(\theta=1|l)f_{1,i}(1)=\frac{\pi_0}{\pi_0+\pi_1e^l}(1-\tau_{00}^i)+\frac{\pi_1e^l}{\pi_0+\pi_1e^l}\tau_{11}^i.$$

Combining the above two equations and (4.14), we have

$$\begin{aligned}&\mathbb{E}_{l,\delta}G\left(l+\log\frac{f_{1,\delta}(X)}{f_{0,\delta}(X)},\ n+1\right)\\=&\mathbb{P}_{l,i}(X=1)G\left(l+\log\left(\frac{\tau_{11}^i}{1-\tau_{00}^i}\right),T\right)+(1-\mathbb{P}_{l,i}(X=1))G\left(l+\log\left(\frac{1-\tau_{11}^i}{\tau_{00}^i}\right),T\right).\end{aligned}$$

Now, we are ready to provide the recursive equation for $G(l,n)$, which is known as the Bellman equation in Markov decision process (see, e.g., Puterman (2005); Bertsekas and Shreve (1978)). In particular, under the current sample size n and log-likelihood ratio l, the action for the next stage has two possible choices:

1) Stopping the testing procedure: the corresponding Bayes risk will be

$$\min\{\pi(\theta=0|l),\pi(\theta=1|l)\}+nc;$$

2) Collecting the next response from the experiment $j^\dagger(l,n+1)$ and the expected conditional risk becomes

$$\mathbb{E}_{l,j^\dagger(l,n+1)}G\Big(l+\log\frac{f_{1,j^\dagger(l,n+1)}(X)}{f_{0,j^\dagger(l,n+1)}(X)},\ n+1\Big).$$

Combining these two cases, one should choose the best possible action (either stop or continue) that leads to the minimum risk, resulting in the following recursive equation

for $G(l,n)$,

$$\begin{aligned} &G(l,n) \\ =&\min\Big\{\mathbb{E}_{l,j^\dagger(l,n+1)}G\Big(l+\log\frac{f_{1,j^\dagger(l,n+1)}(X)}{f_{0,j^\dagger(l,n+1)}(X)},n+1\Big),\min\Big\{\frac{\pi_0}{\pi_0+\pi_1e^l},\frac{\pi_1e^l}{\pi_0+\pi_1e^l}\Big\}+nc\Big\}. \end{aligned}$$

Finally, let $C(n)$ be the set of log-likelihood ratio at which one should stop when the current sample size is n.

The upper hitting boundary $A^\dagger(n)$ and lower hitting boundary $B^\dagger(n)$ should then be the supremum and infimum of the log-likelihood ratio in $C(n)$. Given all the previous discussions, we present a dynamic programming algorithm for the truncated test in Algorithm 1.

Remark 2. To implement Algorithm 1 and to solve for function $G(l,n), n=1,...,T$, discretization of l and interpolation for $G(\cdot,n), n=1,...,T$ is necessary. That is, we approximate $G(\cdot,n), n=1,...,T$ with piecewise linear functions corresponding to the discretization over l. To justify this approximation, we notice that $G(l,n)$ is the minimum of finitely many continuous functions for $n=1,...,T$. Therefore, $G(l,n)$ is a continuous function in l for $n=1,...,T$.

Remark 3. The computational complexity for the dynamic programming (DP) grows at the order of T times the discretization size of the likelihood ratio, where T is the truncation length. It is worth noting that the computation of the DP is done offline — *before* collecting any data and running the test. Given the computational power nowadays, the offline computation is usually not considered as a computational burden.

Algorithm 1 Dynamic Programming for truncated Ada-SPRT

Inputs:
T, c, π_0, π_1, $\{f_{1,\delta}(\cdot)\}_{\delta\in I}$, $\{f_{0,\delta}(\cdot)\}_{\delta\in I}$
Initialize:
$G(l,T) \leftarrow \min\left(\frac{\pi_0}{\pi_0+\pi_1 e^l}, \frac{\pi_1 e^l}{\pi_0+\pi_1 e^l}\right) + Tc$, for each l.
for $n = T-1$ to 0 **do**
$j^\dagger(l,n+1) \leftarrow \arg\min_{\delta\in I} \mathbb{E}_{l,\delta} G\left(l + \log\frac{f_{1,\delta}(X)}{f_{0,\delta}(X)},\ n+1\right)$

$$G(l,n) \leftarrow \min\left\{\mathbb{E}_{l,j^\dagger(l,n+1)} G\left(l + \log\frac{f_{1,j^\dagger(l,n+1)}(X)}{f_{0,j^\dagger(l,n+1)}(X)}, n+1\right), \min\left\{\frac{\pi_0}{\pi_0+\pi_1 e^l}, \frac{\pi_1 e^l}{\pi_0+\pi_1 e^l}\right\} + nc\right\}.$$

$$C(n) \leftarrow \left\{l : \min\left\{\frac{\pi_0}{\pi_0+\pi_1 e^l}, \frac{\pi_1 e^l}{\pi_0+\pi_1 e^l}\right\} + nc \geq \mathbb{E}_{l,j^\dagger(l,n+1)} G\left(l + \log\frac{f_{1,j^\dagger(l,n+1)}(X)}{f_{0,j^\dagger(l,n+1)}(X)},\ n+1\right)\right\}.$$

$A^\dagger(n) \leftarrow \arg\sup\{l : l \in C(n)\}$.
$B^\dagger(n) \leftarrow \arg\inf\{l : l \in C(n)\}$.
end for
Outputs:
$j^\dagger$, $A^\dagger(n), B^\dagger(n)$ for $n = 1, ..., T$.

4.3 Non-truncated Test

In this subsection, we investigate the relationship between the non-truncated ($T = \infty$) and the truncated test ($T < \infty$). The structure of the optimal adaptive sequential design for a truncated test is simpler than that for the truncated test. In particular, we extend the result in Shiryaev (1978, Chapter 4.1, Lemma 1 and Theorem 1), by adding the experiment selection component, and prove the following proposition on the structure of an optimal adaptive sequential design (J^*, N^*, D^*). Let $\mathcal{A}^*$ be the set of all the adaptive sequential designs such that both $\mathbb{E}(N|\theta = 0)$ and $\mathbb{E}(N|\theta = 1)$ are finite. We note that the assumptions $\mathbb{E}(N|\theta = 0) < \infty$ and $\mathbb{E}(N|\theta = 1) < \infty$ are commonly made in sequential analysis, e.g., Wald and Wolfowitz (1948).

Proposition 1. *Let* (J^*, N^*, D^*) *be an optimal adaptive sequential design for a non-truncated test such that.*

$$\mathbf{R}(J^*, N^*, D^*) = \min_{(J,N,D)\in\mathcal{A}^*} \mathbf{R}(J, N, D). \tag{4.16}$$

Then (J^*, N^*, D^*) *has the following properties.*

(i) The optimal stopping time N^* *is a boundary hitting time. That is, there exist real values* A^* *and* B^* *such that* $B^* \leq A^*$ *and*

$$N^* = \inf\{n : l_n^* \geq A^* \text{ or } l_n^* \leq B^*\}.$$

(ii) The optimal decision rule D^* *chooses between* H_0 *and* H_1 *according to whether the log-likelihood ratio statistic hits the upper or the lower boundary, i.e.,*

$$D^* = 1 \text{ if } l_{N^*}^* \geq A^* \quad \text{and} \quad D^* = 0 \text{ if } l_{N^*}^* \leq B^*.$$

(iii) Each j_n^* *in the optimal experiment selection rule* J^* *can be expressed as a single experiment selection function* $j^* : \mathbb{R} \to I$ *such that for any* $n = 1, 2, ..., N^*$,

$$\delta_n^* = j^*(l_{n-1}^*).$$

The proof of the Proposition 1 is provided in the supplementary material.

Remark 4. It was shown in Wald and Wolfowitz (1948) that if the stopping time is defined by the first passage time toward two flat boundaries, then the expected sample size is minimized under each hypothesis when the error probabilities are controlled.

With adaptive experiment selection, such an optimal solution usually does not exist. The main reason is that the best experiment selection rules are different under the null and alternative hypotheses, because the Kullback-Leibler information is not a symmetric function. Thus, an informative experiment for one hypothesis may contain little information about the other. Consequently, the expected sample sizes under both hypotheses may not be minimized simultaneously.

In contrast to the truncated case in Theorem 1, the boundaries for non-truncated tests are flat. Moreover, the selection function j^* is independent of the current sample size $n-1$ and depends on previous responses $X_1, \ldots, X_{n-1}$ only through the log-likelihood ratio statistic l^*_{n-1}.

The next theorem shows that in terms of the minimum Bayes risk, the non-truncated test is a limiting version of the truncated test as $T \to \infty$.

Theorem 2. *Let $\mathcal{A}^T$ denote the set of all adaptive sequential designs (J, N, D) such that $N \leq T$, and $\mathcal{A}^*$ the set of all sequential adaptive designs that have finite expected sample size. Then,*

$$\lim_{T\to\infty} \min_{(J,N,D)\in\mathcal{A}^T} \mathbf{R}(J, N, D) = \min_{(J,N,D)\in\mathcal{A}^*} \mathbf{R}(J, N, D).$$

The proof of Theorem 2 is provided in the supplementary material.

5. Multiple Hypotheses Testing and Empirical Bayes Approach

So far, we have discussed optimal Ada-SPRT for a single object. Now we are ready to address our target problem in (1.1), which contains K hypothesis testing problems. We assume that $\theta_k \in \{0, 1\}$ for $k = 1, \ldots, K$ are independently and identically dis-

tributed following the Bernoulli distribution with a parameter π_1. Given $\theta_1, ..., \theta_K$, we assume that $(X_{kj}; j = 1, 2, ..)$ are responses (assumed independent across different k) obtained for the k'th object, whose density function is $f_{\theta_k, \delta_{kj}}(\cdot)$, and δ_{kj} denotes the j'th experiment selected for the k'th object.

Let us recall the last paragraph in Section 3, where $D = \{D_k\}_{k=1}^K$ is the set of decisions and $N = \{N_k\}_{k=1}^K$ is the set of stopping times. The averaged loss L_K is defined in (3.7).

If the class prior π_1 is known, then, according to Theorem 1, the optimal design that minimizes $\mathbb{E}L_K$ is that we run Algorithm 1 independently for each object k to obtain the optimal experiment selection rule (denoted by $j^{(k)}$) and boundaries or sequence of boundaries for the truncated case (denoted by $A^{(k)}$ and $B^{(k)}$). Given $j^{(k)}$, $A^{(k)}$ and $B^{(k)}$, the requestor collects labels according to the selection rule $j^{(k)}$ for each object k and makes the decision D_k according to the hitting boundary. Although such a procedure is easy to implement, the class prior π_1 and $\pi_0 = 1 - \pi_1$ in (1.2) are unknown in many real-world applications. With multiple objects, one can estimate the class prior via the *empirical Bayes* approach described as follows.

For each k, we estimate π_1 by some estimator $\widehat{\pi}_1$ based on the collected responses for previous hypothesis $1, 2, \ldots, k-1$. In principle, any estimator can be applied to estimate π_1 and we adopt the maximum likelihood estimator. Then, for the k-th hypothesis, we use Algorithm 1 with the estimated parameters $\widehat{\pi}_1^{(k)}$, $\widehat{\pi}_0^{(k)} = 1 - \widehat{\pi}_1^{(k)}$ to solve for the experiment selection rule and stopping time for the k-th hypothesis. The algorithm is presented in Algorithm 2, where we initialize the estimate for π_1 to be 0.5 for simplicity.

As the number of hypotheses K grows large, and the estimate $\widehat{\pi}_1$ becomes very

Algorithm 2 Ada-SPRT for multiple objects using empirical Bayes method

Inputs:
c, $\{f_{1,\delta}(\cdot)\}_{\delta\in I}$, $\{f_{0,\delta}(\cdot)\}_{\delta\in I}$, T
Initialize:
$\widehat{\pi}_0^{(0)} = \widehat{\pi}_1^{(0)} = 0.5$
for k=1 to K **do**
    Run Algorithm 1 with Inputs c, $\widehat{\pi}_0^{(k-1)}, \widehat{\pi}_1^{(k-1)}$, $\{f_{1,\delta}(\cdot)\}_{\delta\in I}$, $\{f_{0,\delta}(\cdot)\}_{\delta\in I}$, and T. Obtain Outputs $A^{(k)}, B^{(k)}, j^{(k)}$.
    Collect responses according to the experiment selection rule $j^{(k)}$ and obtain the decision D_k according to the boundary hitting.
    Update $\widehat{\pi}_0^{(k)}$ and $\widehat{\pi}_1^{(k)}$ with the newly collected responses.
end for
Outputs:
Decision D_k and sample size N_k for each hypothesis $k = 1, \ldots, K$.

accurate, the resulting averaged loss L_K in (3.7) will converge to the minimal Bayes risk corresponding to the true π_1. We characterize this asymptotic result in the next theorem.

Theorem 3. *Assume that $c < \pi_1 < 1 - c$, and $\widehat{\pi}_1 \to \pi_1$ in probability as $K \to \infty$ and the sequential adaptive deign D_k and N_k are determined through the empirical Bayes procedure described in Algorithm 2, then*

$$L_K \to \min_{(J,N,D)\in\mathcal{A}} \mathbf{R}(J, N, D) \text{ in probability as } K \to \infty,$$

where $\mathbf{R}(J, N, D)$ is the minimal Bayes risk of a single object defined in (4.16)*. That is, the averaged loss L_K in* (3.7) *converges to the minimal Bayes risk under the true model.*

The proof of Theorem 3 is provided in the supplement. We also note that in Theorem 3, the assumption $c < \pi_1 < 1 - c$ is a necessary condition for the optimal test procedure to be non-trivial, without which the optimal test will always stop with no

sample.

Remark 5. For applications in crowdsourcing, the classification results are usually not accurate due to the limited number of labels. Thus, the average performance L_K (defined as in (3.7)) is a common choice of error metric in practice when there are many labeling tasks. It treats type I and type II errors symmetrically, which is adopted in many classification problems.

There are other error metrics considered in the literature. For example, false discovery rate (FDR), positive false discovery rate (pFDR), marginal false discovery rate (mFDR), and family-wise error rate (FWER) are used for measuring the accuracy of multiple testing procedures (Benjamini and Hochberg, 1995; Storey, 2003), and corresponding sequential procedures have been developed (Bartroff, 2018; Song and Fellouris, 2019). Because our proposed method outputs individualized decision and posterior probability for each labeling task, we may estimate local false discovery rate (Efron, 2007) and further control the global FDR by adjusting the stopping boundaries. Another example of error metric is the maximum loss (or sample size), rather than the averaged loss. This metric corresponds to the analysis of the worst case scenario. Heuristically, the maximum loss grows to infinity as the number of tasks increases, and its asymptotically order is determined by its tail probability through extreme value theory. Overall, it is worth further investigation on the optimal procedures for various choices of error metrics.

In the sequential analysis literature, the distributions $\{f_{1,\delta}(\cdot)\}_{\delta\in I}$ and $\{f_{0,\delta}(\cdot)\}_{\delta\in I}$ are typically assumed to be known. However, in real crowdsourcing applications, it is quite often that no prior knowledge on workers' quality parameters $\{\tau_{00}^i\}_{i\in I}$, $\{\tau_{11}^i\}_{i\in I}$ in (3.9) is available. Therefore, one cannot directly compute the likelihood ratio statistics

in terms of $\{f_{1,\delta}(\cdot)\}_{\delta\in I}$ and $\{f_{0,\delta}(\cdot)\}_{\delta\in I}$. To address this issue, we propose to estimate the workers' quality parameters using a regularized maximum likelihood estimate under the two-coin model in (3.9) after finishing the labeling process for each object k. In particular, after each for-loop in Algorithm 2 (i.e., the labeling process and the decision for the k-th object has finished), we have collected all the responses $\{Z_{ji}\}$, where each Z_{ji} is a binary label from worker $i \in I$ to the object $j \in \{1,\ldots,k\}$. A regularized minus log-likelihood is defined as follows,

$$
\begin{aligned}
&h_k\left(\pi_1, \{\tau_{00}^i\}_{i\in I}, \{\tau_{11}^i\}_{i\in I}\right) = \qquad (5.17)\\
&-\sum_{1\le j\le k} \log\left((1-\pi_1)\prod_i (\tau_{00}^i)^{1-Z_{ji}}(1-\tau_{00}^i)^{Z_{ji}} + \pi_1 \prod_i (1-\tau_{11}^i)^{1-Z_{ji}}(\tau_{11}^i)^{Z_{ji}}\right)\\
&+\sum_{i\in I}\left((\alpha-1)\log(\tau_{00}^i) + (\beta-1)\log(1-\tau_{00}^i) + (\alpha-1)\log(\tau_{11}^i) + (\beta-1)\log(1-\tau_{11}^i)\right).
\end{aligned}
$$

The regularization term comes from the Beta priors on τ_{00}^i and τ_{11}^i for each $i \in I$ with parameters α and β, which makes the estimation stable when a worker has only labeled a small number of objects. We minimize $h_k\left(\pi_1, \{\tau_{00}^i\}_{i\in I}, \{\tau_{11}^i\}_{i\in I}\right)$ at the end of k-th iteration in Algorithm 2 using the expectation maximization (EM) algorithm (Dempster et al., 1977), which simultaneously provides the estimate of class prior π_1 (i.e., $\widehat{\pi}_1^{(k)}$) and workers' quality parameters $\{\tau_{00}^i\}_{i\in I}, \{\tau_{11}^i\}_{i\in I}$ (see the details in Dawid and Skene (1979)). These estimates will be used to construct the optimal adaptive sequential designs for the next object $k+1$. After the decision for the $(k+1)$-th object has been made, we re-optimize $h_{k+1}\left(\pi_1, \{\tau_{00}^i\}_{i\in I}, \{\tau_{11}^i\}_{i\in I}\right)$ using all the previously collected responses. We also adopt the estimate from the k-th iteration as the starting point (so-called warm-start) so that the EM algorithm usually quickly converges in a few iterations.

6. Experimental Results

In this section, we demonstrate the performance of the proposed Ada-SPRT method using two benchmark binary labeling crowdsourcing datasets. We also conduct extensive simulation studies, which are relegated to the supplementary material due to space constraints. A brief summary of the two real datasets is provided below.

1) Recognizing textual entailment (RTE dataset (Snow et al., 2008)): there are $K = 800$ objects and each object is a sentence pair. Each sentence pair is presented to 10 different workers to acquire binary choices of whether the second hypothesis sentence can be inferred from the first one. There are in total $M = 164$ different workers and total number of available labels is $8,000$. Since each object receives 10 labels, we use the truncated Ada-SPRT with the truncation length $T = 10$.

2) Labeling bird species (Bird dataset (Liu et al., 2012; Welinder et al., 2010)): there are $K = 108$ objects and each object is an image of a bird. Each image receive 39 binary labels (either indigo bunting or blue grosbeak) from all $M = 39$ workers and the total number of available labels is $4,212$. We use the truncated Ada-SPRT with the truncation length $T = 39$.

We note that the true labels are available for both datasets from domain experts so that we could evaluate the labeling accuracy of the decision D_k for each object $k \in \{1, \ldots, K\}$.

For both datasets, we use truncated Ada-SPRT algorithm with EM algorithm to estimate class prior and workers' quality parameters as described in Section 5. We set $\alpha = 4$ and $\beta = 2$ in the regularized likelihood function in (5.17). Since α and β reflect

Table 1: Performance comparison on real datasets in terms of mean and standard deviation of accuracy for different approaches. KG and Opt-KG correspond to the knowledge gradient or optimistic knowledge gradient worker selection policy with the same stopping time as that of Ada-SPRT. KG Avg and Opt-KG Avg correspond to the knowledge gradient or optimistic knowledge gradient worker selection policy with the average stopping time for all objects. The accuracies in bold are the best accuracies for each choice of c.

RTE (Accuracy)	$c = 2^{-6}$	$c = 2^{-8}$	$c = 2^{-10}$	$c = 2^{-12}$
Total queried labels	3438(60)	3949 (46)	4365 (28)	4660(28)
Ada-SPRT	**92.1%** (0.4%)	**92.6%** (0.3%)	**92.5%** (0.3%)	**92.6%** (0.2%)
KG	86.9% (1.3%)	87.4% (0.9%)	88.0% (1.3%)	88.9% (1.3%)
Opt-KG	82.5% (2.2%)	84.3% (2.7%)	85.2% (1.5%)	88.5% (1.7%)
KG Avg	86.1% (2.9%)	86.0% (2.4%)	87.7% (1.1%)	87.9% (1.3%)
Opt-KG Avg	82.2% (4.3%)	83.2% (3.2%)	86.7% (2.0%)	88.0% (2.2%)
Bird (Accuracy)	$c = 2^{-6}$	$c = 2^{-8}$	$c = 2^{-10}$	$c = 2^{-12}$
Total queried labels	1253 (37)	1392 (40)	1523 (47)	1672 (57)
Ada-SPRT	**85.7**% (4%)	**87.5**% (2%)	**87.4**% (2%)	**87.1**% (1%)
KG	74.6% (5.8%)	75.9% (3.6%)	77.6% (4.4%)	77.4% (3.2%)
Opt-KG	71.3% (5.5%)	74.4% (5.1%)	77.1% (4.5%)	78.1% (3.9%)
KG Avg	80.4% (3.5%)	78.8% (4.6%)	80.0% (2.9%)	80.8% (2.4%)
Opt-KG Avg	83.9% (2.5%)	84.7% (2.8%)	85.9% (1.8%)	85.0% (2.4%)

the prior belief of workers' accuracy, $\alpha = 4$ and $\beta = 2$ correspond to a prior accuracy of $\frac{\alpha}{\alpha+\beta} = \frac{4}{4+2} = 66.7\%$. Other settings of α and β lead to similar performance as long as $\alpha > \beta$ (i.e., a worker is believed to perform better than random guess).

Since different ordering of objects in Algorithm 2 leads to slightly different results, we report the average over 20 random orderings. In addition, the first quarter of the objects (i.e., the first 200 objects for RTE and the first 27 objects for Bird) will be used as a "calibration" set. In particular, for those objects, we use all the T responses (i.e., setting $N_k = T$) without selecting workers so that good initial estimates of the class

prior and workers' quality parameters can be obtained based on the "calibration" set. For the objects not in the "calibration" set, the averaged stopping times as c ranging from 2^{-6} to 2^{-12} are 2.4, 3.2, 3.9, and 4.4, respectively, for the RTE dataset. For the bird dataset, the averaged stopping times as c ranging from 2^{-6} to 2^{-12} are 2.5, 4.2, 5.8, and 7.6, respectively.

We compare Ada-SPRT with two state-of-the-art worker selection policies in Chen et al. (2015): (1) Knowledge gradient (KG) policy and (2) Optimistic knowledge gradient (Opt-KG) policy. We note that both KG and Opt-KG are myopic index policy only for worker selection but not for optimal stopping. To make a fair comparison, we consider different ways of adding stopping times for KG and Opt-KG: (1) using the same stopping time N_k from Ada-SPRT for each object k (2) using the average stopping time $\lceil \frac{1}{K} \sum_{i=1}^{K} N_k \rceil$ for all objects. Recall that N_k is the stopping time obtained by Ada-SPRT in Algorithm 2 for the k-th object. We vary the cost parameter c and report mean and standard deviation of total queried labels (i.e., $\sum_{i=1}^{K} N_k$) and labeling accuracies for different approaches.

The comparison results are provided in Table 1 for RTE and Bird datasets respectively. As can be seen from Table 1, Ada-SPRT greatly outperforms other approaches on both datasets. We also note that under the two-coin model, when using all the available labels, the labeling accuracy is 92.88% (with $8,000$ labels) for RTE dataset and 89.1% (for $4,212$ labels) for Bird dataset. Therefore, from Table 1, Ada-SPRT achieves on average $\frac{92.1}{92.88} = 99\%$ of the best possible labeling accuracy using only $\frac{3438}{8000} = 43\%$ of the total labels for RTE, and $\frac{85.7}{89.1} = 96\%$ of the best possible labeling accuracy using only $\frac{1253}{4212} = 30\%$ of the total labels for Bird.

7. Discussions

In this paper, we propose an adaptive sequential probability ratio test (Ada-SPRT) which finds the optimal experimental selection rule, stopping time, and decision rule for a single hypothesis testing problem. For multiple testing problems, we further propose an empirical Bayes approach which estimates the class prior. We demonstrate the effectiveness of our methods on real crowdsourcing applications.

There are several directions along which this work may be extended. First, we only consider simple versus simple hypothesis for binary labeling tasks. It is of great interest to extend the current framework to composite hypotheses. Second, although we mainly consider crowdsourcing applications with a brief mention of computerized mastery testing, our Ada-SPRT is a general framework for adaptive sequential test, for which we would like to explore more applications.

Acknowledgment

The authors thank the editors and two referees for their constructive comments. Xiaoou Li's research is partially supported by the NSF grant DMS-1712657. Yunxiao Chen's research is supported in part by NAEd/Spencer postdoctoral fellowship. Xi Chen's research is partially supported by the NSF grant IIS-1845444 and the Bloomberg Data Science Research Grant. Jingchen Liu's research is partially supported by NSF IIS-1633360 and SES-1826540. Zhiliang Ying's research is partially supported by NSF IIS-1633360 and SES-1826540, and NIH grant R01GM047845.

Supplementary Materials

In the Supplement, we present the proofs of all the theoretical results, including Proposition 1, Theorems 1, 2 and 3 and the supporting lemmas. We also present

simulated experiments in the supplement.

References

Albert, A. E. (1961). The sequential design of experiments for infinitely many states of nature. *The Annals of Mathematical Statistics*, 32(3):774–799.

Arrow, K. J., Blackwell, D., and Girshick, M. A. (1949). Bayes and minimax solutions of sequential decision problems. *Econometrica*, 17(3/4):213–244.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Babaioff, M., Sharma, Y., and Slivkins, A. (2009). Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the ACM Conference on Electronic Commerce*.

Bai, C. Z. and Gupta, V. (2016). An on-line sensor selection algorithm for SPRT with multiple sensors. *IEEE Transactions on Automatic Control*, 62(7):3532–3539.

Bartroff, J. (2018). Multiple hypothesis tests controlling generalized error rates for sequential data. *Statistica Sinica*, 28(1):363–398.

Bartroff, J., Finkelman, M., and Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika*, 73(3):473–486.

Bartroff, J. and Lai, T. L. (2008). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Statistics in Medicine*, 27(10):1593–1611.

Bartroff, J., Lai, T. L., and Shih, M. C. (2013). *Sequential Experimentation in Clinical Trials*. New York: Springer.

Bellman, R. (1957). A markovian decision process. Technical report, DTIC Document.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57(1):289–300.

Bertsekas, D. P. and Shreve, S. E. (1978). *Stochastic optimal control: The discrete time case*, volume 23. Academic Press.

Bessler, S. A. (1960). *Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments*. Department of Statistics, Stanford University.

Bhat, N., Farias, V. F., Moallemi, C. C., and Sinha, D. (2019). Near optimal A/B testing. *Management Science (to appear)*.

Brown, L. D. and Greenshtein, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 37(4):1685–1704.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Bussgang, J. and Middleton, D. (1955). Optimum sequential detection of signals in noise. *IRE Transactions on Information Theory*, 1(3):5–18.

Chakrabarti, D., Kumar, R., Radlinski, F., and Upfal, E. (2009). Mortal multi-armed bandits. In *Proceedings of Advances in Neural Information Processing Systems*.

Chang, Y. I. (2004). Application of sequential probability ratio test to computerized criterion-referenced testing. *Sequential Analysis*, 23(1):45–61.

Chang, Y. I. (2005). Application of sequential interval estimation to adaptive mastery testing. *Psychometrika*, 70(4):685–713.

Chen, X., Lin, Q., and Zhou, D. (2015). Statistical decision making for optimal budget allocation in crowd labeling. *Journal of Machine Learning Research*, 16(1):1–46.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM

algorithm. *Journal of the Royal Statistical Society Series C*, 28(1):20–28.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38.

Doan, A., Ramakrishnan, R., and Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics*, 35(4):1351–1377.

Efron, B. (2013). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Embretson, S. E. and Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Ertekin, S., Rudin, C., and Hirsh, H. (2014). Approximating the crowd. *Data Mining and Knowledge Discovery*, 28(5–6):1189–1221.

Gao, C. and Zhou, D. (2013). Minimax optimal convergence rates for estimating ground truth from crowdsourced labels. *arXiv preprint arXiv:1310.5764*.

Hoffman, M. D., Brochu, E., and de Freitas, N. (2011). Portfolio allocation for Bayesian optimization. In *Proceedings of Uncertainty in Artificial Intelligence*.

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Irle, A. and Schmitz, N. (1984). On the optimality of the SPRT for processes with continuous time parameter. *Statistics: A Journal of Theoretical and Applied Statistics*, 15(1):91–104.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684.

Jiang, W. and Zhang, C.-H. (2010). *Empirical Bayes in-season prediction of baseball batting averages*, volume 6, pages 263–273. Institute of Mathematical Statistics.

Johari, R., Pekelis, L., and Walsh, D. J. (2015). Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922*.

Karger, D. R., Oh, S., and Shah, D. (2013). Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24.

Karunamuni, R. J. (1988). On empirical Bayes testing with sequential components. *The Annals of Statistics*, 16(3):1270–1282.

Ke, X., Teo, M., Khan, A., and Yalavarthi, V. K. (2018). A demonstration of perc: probabilistic entity resolution with crowd errors. *VLDB Endowment*, 11(12):1922–1925.

Khetan, A. and Oh, S. (2016). Achieving budget-optimality with adaptive schemes in crowdsourcing. In *Proceedings of Advances in Neural Information Processing Systems*.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association*, 109(506):674–685.

Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3):1091–1114.

Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, 11(2):303–408.

Lewis, C. and Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *ETS Research Report Series*, 14(2):367–86.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the International Conference on World Wide Web*.

Liu, Q., Peng, J., and Ihler, A. (2012). Variational inference for crowdsourcing. In *Proceedings of Advances in Neural Information Processing Systems*.

Luengo-Oroz, A. M., Arranz, A., and Frean, J. (2012). Crowdsourcing malaria parasite quantification: An online game for analyzing images of infected thick blood smears. *Journal of Medical Internet Research*, 14(6):e167.

Marcus, A. and Parameswaran, A. (2015). Crowdsourced data management: Industry and academic perspec-

tives. *Foundations and Trends in Databases*, 6(1-2):1–161.

Mavandadi, S., Dimitrov, S., Feng, S., Yu, F., Sikora, U., Yaglidere, O., Padmanabhan, S., Nielsen, K., and Ozcan, A. (2012). Distributed medical image analysis and diagnosis through crowd-sourced games: A malaria case study. *PLoS ONE*, 7(5):e37245.

Naghshvar, M. and Javidi, T. (2010). Active M-ary sequential hypothesis testing. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*.

Naghshvar, M. and Javidi, T. (2013a). Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738.

Naghshvar, M. and Javidi, T. (2013b). Sequentiality and adaptivity gains in active hypothesis testing. *IEEE Journal of Selected Topics in Signal Processing*, 7(5):768–782.

Nikiforov, I. V. (1975). Sequential analysis applied to autoregression processes. *Avtomatika i Telemekhanika*, 36(8):174–177.

Nitinawarat, S. and Veeravalli, V. V. (2015). Controlled sensing for sequential multihypothesis testing with controlled Markovian observations and non-uniform control cost. *Sequential Analysis*, 34(1):1–24.

Ok, J., Oh, S., Shin, J., and Yi, Y. (2016). Optimality of belief propagation for crowdsourced classification. In *Proceedings of the International Conference on Machine Learning*.

Press, W. H. (2009). Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proceedings of the National Academy of Sciences*, 106(52):22387–22392.

Puterman, M. L. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: John Wiley & Sons, Inc.

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Murray, P., Schawinski, K., Szalay, A. S., and Vandenberg, J. (2010). Galaxy Zoo: Exploring the motivations of citizen science volunteers. *Astronomy Education Review*, 9(1):1–6.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from

crowds. *Journal of Machine Learning Research*, 11(4):1297–1322.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Robbins, H. and Siegmund, D. O. (1974). Sequential tests involving two populations. *Journal of the American Statistical Association*, 69(345):132–139.

Shah, N. B., Balakrishnan, S., and Wainwright, M. J. (2016). A permutation-based model for crowd labeling: Optimal estimation and robustness. *arXiv preprint arXiv:1606.09632v1*.

Shiryaev, A. N. (1978). *Optimal stopping rules*. Springer Science & Business Media.

Siegmund, D. O. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer New York.

Slivkins, A. and Vaughan, J. W. (2013). Online decision making in crowdsourcing markets: Theoretical challenges. *SIGecom Exchanges*, 12(2):4–23.

Snow, R., Connor, B. O., Jurafsky, D., and Ng., A. Y. (2008). Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of Empirical Methods for Natural Language Processing*.

Sobel, M. and Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *The Annals of Mathematical Statistics*, 20(4):502–522.

Song, Y. and Fellouris, G. (2019). Sequential multiple testing with generalized error control: An asymptotic optimality theory. *Ann. Statist.*, 47(3):1776–1803.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035.

Tartakovsky, A., Nikiforov, I., and Basseville, M. (2014). *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Chapman and Hall/CRC.

Tsitovich, I. (1985). Sequential design of experiments for hypothesis testing. *Theory of Probability & Its Applications*, 29(4):814–817.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wald, A. (1947). *Sequensial Analysis*. Dover Publications.

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326–339.

Wald, A. and Wolfowitz, J. (1950). Bayes solutions of sequential decision problems. *The Annals of Mathematical Statistics*, 21(1):82–99.

Welinder, P., Branson, S., Belongie, S., and Perona, P. (2010). The multidimensional wisdom of crowds. In *Proceedings of Advances in Neural Information Processing Systems*.

Yalavarthi, V. K., Ke, X., and Khan, A. (2017). Select your questions wisely: For entity resolution with crowd errors. In *Proceedings of the International Conference on Information and Knowledge Management*.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods. *The Annals of Statistics*, 31(2):379–390.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. (2016). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(1):1–44.

Xiaoou Li, School of Statistics, University of Minnesota

E-mail: lixx1766@umn.edu

Yunxiao Chen, Department of Statistics, London School of Economics and Political Science

E-mail: y.chen186@lse.ac.uk

Xi Chen, Stern School of Business, New York University

E-mail: xichen@nyu.edu

Jingchen Liu, Department of Statistics, Columbia University

E-mail: jcliu@stat.columbia.edu

Zhiliang Ying, Department of Statistics, Columbia University

E-mail: zying@stat.columbia.edu