Statistica Sinica **31** (2021), 519-546
doi:https://doi.org/10.5705/ss.202018.0300

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]

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

Abstract: In this study, we solve a class of multiple testing problems under a Bayesian sequential decision framework. Our work is motivated by binary labeling tasks in crowdsourcing, where a requestor needs to simultaneously choose a worker to provide a label and decide when to stop collecting labels, under a certain budget constraint. We begin by using a binary hypothesis testing problem to determine the true label of a single object, and provide an optimal solution by casting it under an adaptive sequential probability ratio test framework. Then, we characterize the structure of the optimal solution, that is, the optimal adaptive sequential design, which minimizes the Bayes risk using a log-likelihood ratio statistic. We also develop a dynamic programming algorithm to efficiently compute the optimal solution. For the multiple testing problem, we propose an empirical Bayes approach for estimating the class priors, and show that the average loss of our method converges to the minimal Bayes risk under the true model. Experiments on both simulated and real data show the robustness of our method, as well as its superiority over existing methods in terms of its labeling accuracy.

Key words and phrases: Bayesian decision theory, crowdsourcing, empirical Bayes, sequential analysis, sequential probability ratio test.

1. Introduction

Crowdsourcing, 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 10 years, crowdsourcing has become an efficient and economical way to obtain 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 then ask a crowd of workers to tag each picture as a portrait or a landscape,

Corresponding author: Xiaoou Li, School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA. E-mail: lixx1766@umn.edu.

offering a small payment for each label. This technique has helped address a wide range of challenges in scientific areas, such as understanding of the evolution of galaxies using a crowd classifying galaxy morphology (Galaxy Zoo (Raddick et al. (2010))), and diagnosing malaria epidemics by asking crowd workers to identify malaria-infected red blood cells (MOLT (Mavandadi et al. (2012)) and MalariaSpot (Luengo-Oroz, Arranz and Frean (2012))); see Doan, Ramakrishnan and Halevy (2011), Slivkins and Vaughan (2013), and Marcus and Parameswaran (2015) for comprehensive reviews of crowdsourcing techniques and their applications.

Despite the efficiency of the technique and the immediate availability of the data, the labels generated by nonexpert crowd workers are quite noisy. For example, as reported in Yalavarthi, Ke and Khan (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); that is, they collect multiple labels from different workers for a single object. Then, the requestor aggregates the collected labels to infer the true label. In general, a greater number of labels yields a more accurate inferred label. However, each label incurs a fixed cost: the requestor has to pay a prespecified monetary cost for each label, regardless of its correctness. Therefore, when using a crowdsourcing service for large-scale labeling tasks, a requestor usually faces two challenges:

1. The requestor needs to carefully balance 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. Crowd workers have different levels of quality/reliability. Thus, 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 as a general multiple testing problem under a sequential analysis framework. In particular, we examine binary labeling tasks, which are used to categorize, for example, an image as a portrait or a landscape or a website as pornographic or not. We assume there are K objects. For each object, we test whether its true label (denoted by $\theta_k \in \{0, 1\}$) belongs to class zero or one. More specifically, this problem can be formulated as K hypothesis testing problems:

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

Because the true classes of the objects might be highly unbalanced, it is natural to assume 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 imbalance between two classes, which is usually unknown. To study this problem, we 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 a Bayesian sequential analysis framework. We first formulate a risk function, defined as the expected probability of making the wrong decision plus the expected labeling cost. Here, we need to optimize three components:

1. Stopping time: This indicates when to stop collecting additional data (i.e., labels) under a certain budget constraint (e.g., given a prespecified maximum number of labels that can be collected). Cost-effective crowdsourcing requires that the requestor stop when there is consensus on a label thus avoiding unnecessary costs.

2. Adaptive experimental selection rule: We assume there are M possible experiments (corresponding to heterogeneous workers), where different experiments lead to different distributions from which to generate 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.

Note that our Ada-SPRT can be viewed as an extension of the classical SPRT by Wald (1945) and Wald and Wolfowitz (1948), which optimizes the stopping time and decision rule, but does not consider experiment selection.

In the sequential analysis literature, Chernoff (1959), among many others, provided asymptotically optimal solutions for various sequential design problems (see Section 2). 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 the stopping time (i.e., a truncation length). Under this constraint, the sample size cannot go to infinity and, thus, the theory related to the asymptotically optimal experimental design no longer holds.

2. In (1.2), we need to estimate a class prior distribution π_1, which plays an important role in our problem, given the small truncation length. However, classical asymptotically optimal results usually ignore the effect of the prior probability distribution 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* and a *dynamic programming algorithm* to solve the single hypothesis testing problem in (1.3), using a prespecified truncation length T. For a single truncated test, the sequential decision problem can be formulated as 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 nonincreasing with respect to (w.r.t.) the sample size $n = 1, \ldots, T$, and the lower boundary curve is nondecreasing w.r.t. the sample size n.

2. The *optimal decision* for the true label is determined by 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.

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

Using the Ada-SPRT to solve (1.3), we then 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 converges to the minimal Bayes risk under the true model. We demonstrate the robustness of the proposed method, as well as its superior performance for different setups of the true prior distribution (e.g., unbalanced class setting) using empirical studies.

Finally, note that although our study is motivated by a crowdsourcing application, the proposed empirical Bayes and Ada-SPRT approach is a general method that can be used to solve the multiple testing problem given 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, Finkelman and Lai (2008)) has become an important part of educational assessment. These tests are based on item response theory models (e.g., Embretson and Reise (2000)) and are used to classify examinees into "mastery" and "nonmastery" categories. The Ada-SPRT can be extended to provide an optimal adaptive mastery test design (in terms of the Bayes risk).

The rest of the paper is organized as follows. In Section 2, we discuss related works on crowdsourcing, sequential analysis, and the empirical Bayes method. 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 times and decisions 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 to estimate class priors. In Section 6, we demonstrate the performance of the proposed algorithm using real crowdsourcing data sets. Section 7 concludes the paper. The proofs of the theoretical results and additional simulated experiments are provided in the online Supplementary Material.

2. Related Works

Crowdsourcing is a popular and effective method of collecting labels at low cost and, as a result, has received much attention from researchers in the fields of statistics and machine learning. Many works in this field attempt to solve static problems, such as inferring true labels and workers' quality parameters, based on a static set of labels (see, e.g., Raykar et al. (2010); Karger, Oh and Shah (2013); Liu, Peng and Ihler (2012); Gao and Zhou (2013); Ertekin, Rudin and Hirsh (2014); Zhang et al. (2016); Khetan and Oh (2016); Shah, Balakrishnan and Wainwright (2016); Ok et al. (2016)). To monitor worker quality, most of these works adopt the Dawid–Skene model (Dawid and Skene (1979)), also known as the two-coin model for binary labeling tasks. As such, we assume the same model. Several recent works (e.g., Shah, Balakrishnan and Wainwright (2016); Khetan and Oh (2016)) have examined more general models, such as the generalized Dawid–Skene and permutation models. However, relatively few studies have investigated the adaptive worker selection problem. Karger, Oh and Shah (2013) proposed assigning workers based on a random bipartite graph. Chen, Lin and Zhou (2015) considered a fixed-budget problem, which they formulated as a Bayesian MDP. They investigated two greedy policies that provide approximate solutions to the MDP: (1) the knowledge gradient (KG) policy, which chooses the best experiment/action that maximizes the expected reward for the next stage; and (2) the optimistic knowledge gradient (Opt-KG) policy, where the best action maximizes the reward for collecting a positive or a negative label. We compare these greedy policies in our experiments (see Section 6). In addition, Ertekin, Rudin and Hirsh (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 a classification error less than a small threshold, with high probability.

Note that instead of pre-fixing a total budget, as in some works (e.g., Chen, Lin and Zhou (2015)), our goal is to simultaneously select a worker and decide on an optimal stopping time. To achieve this goal, we formulate the problem as a Bayesian sequential testing problem and propose a framework (Ada-SPRT). Sequential testing has been well researched, beginning with the seminal works of Wald (1945) and Wald and Wolfowitz (1948) for testing two simple hypotheses, see Lai (2001) for a survey, and Siegmund (1985) and Tartakovsky, Nikiforov and Basseville (2014) for a comprehensive review. Sequential tests are widely applied in areas such as industrial quality control, the design of clinical trials, finance, educational testing, among others (Lai (2001); Bartroff and Lai (2008); Bartroff, Finkelman and Lai (2008); Bartroff, Lai and Shih (2013); Tartakovsky, Nikiforov and Basseville (2014)). The problem of sequential adaptive experiment selection was initially treated in Chernoff (1959), who considers a Bayes risk defined similarly to that in Wald and Wolfowitz (1948). Another related work is that of Robbins and Siegmund (1974), who present Monte Carlo and theoretical analyses of several adaptive treatment selection rules for clinical trials. Their aim is to reduce the expected number of observations made on the inferior treatment. The current work provides theoretical results in a sequential hypothesis testing framework that simultaneously considers optimal stopping times, worker decisions, and experiments.

The current work is also related to the multi-armed bandit problem (Robbins (1952)), which has been studied in areas such as clinical trials (Press (2009)), online advertising (Chakrabarti et al. (2009); Babaioff, Sharma and Slivkins (2009)), and portfolio design (Hoffman, Brochu and de Freitas (2011)); see Lai (1987), Auer, Cesa-Bianchi and Fischer (2002); Auer et al. (2002), 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 independent and identically distributed (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 total reward over a finite time horizon. Similarly, our problem requires sequential decisions on the worker selection. However, there are two unique challenges in our problem, which prevent a direct application of existing bandit algorithms from the machine learning literature. First, previous works usually assume an intermediate reward after each action, and have as their goal to maximize the total reward (or discounted reward over time). In our problem, each answer from a worker provides some information. However, the noise of these answers means their usefulness is only vaguely related 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 time provides a flexible trade-off between learning accuracy and cost, and can be controlled using the relative cost parameter c in our objective function; see equation (3.2).

Another related subject is the design problem for A/B testing (Bhat et al. (2019); Johari, Pekelis and Walsh (2015)), which is a form of randomized controlled trial that compares treatment effects across groups. It is possible to apply the proposed adaptive sequential method to design an A/B testing strategy. In contrast to existing works in this area, our method incorporates an early stopping strategy and provides an optimal design under the Dawid–Skene model. However, we do not consider covariate information (as, for example, Bhat et al. (2019) do). In crowdsourcing applications, workers' personal information is usually quite sensitive, and might not be readily available. However, contextual information may be available and useful in other applications. Adding such information is left to future research.

The empirical Bayes method has recently gained prominence, both theoretically and in practice (e.g., Jiang and Zhang (2009, 2010); Koenker and Mizera (2014); Brown and Greenshtein (2009); Efron (2013)); see Zhang (2003), Efron (2013), and the references therein for a comprehensive review. In particular, Karunamuni (1988) combines the empirical Bayes method and a sequential test, and then provides a theoretical analysis of 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 best of our knowledge, this is the first result to include the empirical Bayes method, sequential analysis, and experiment selection simultaneously.

We examine our contribution to the literature by comparing the proposed method with existing methods on adaptive sequential testing, which, in general, fall into one of 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 nonasymptotic regime; and 3) sequential hypothesis testing with an adaptive design in a nonasymptotic regime. The major differences between our work and existing methods are summarized below.

1) Hypothesis testing using a sequential design in an asymptotic regime was first studied in Chernoff (1959), followed by Albert (1961), Tsitovich (1985), Naghshvar and Javidi (2013a,b), Bessler (1960), and Nitinawarat and Veeravalli (2015), among others. 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 by the crowdsourcing application, we consider a different regime, where the sample size is not allowed to go to infinity (i.e., we set a fixed cost c and a maximum test length constraint T). Thus, methods and techniques for the asymptotic regime are not applicable to our problem.

2) Sequential hypothesis testing in a nonasymptotic regime was first considered by Wald (1947), followed by Wald and Wolfowitz (1948), Wald and Wolfowitz (1950), Sobel and Wald (1949), Arrow, Blackwell and Girshick (1949), Bussgang and Middleton (1955), Irle and Schmitz (1984), Nikiforov (1975), Bertsekas and Shreve (1978), and Shiryaev (1978). Under a nonasymptotic regime, SPRT is shown to be optimal, from a nonBayesian point of view (Wald and Wolfowitz (1948)). Optimal truncated and nontruncated Bayesian sequential tests have been developed by Arrow, Blackwell and Girshick (1949).

   Our Theorem 1 extends the results for the optimal Bayesian sequential test in Arrow, Blackwell and Girshick (1949) by incorporating an adaptive design, in which we adaptively select the next experiment based on current information.

3) The study of general stochastic control problems under nonasymptotic regimes 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 nonasymptotic regimes. In particular, Naghshvar and Javidi (2010) consider the problem of a single nontruncated sequential test with $M \geq 2$ hypotheses (among which, only one holds). In this study, we examine 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 general problem setup, 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. Then, we extend the result to the case with K objects. For a single object with true label $\theta \in \{0, 1\}$, we investigate the hypothesis testing problem in (1.3). Let $X_1, X_2, \ldots$ be the observed responses. The selection of the nth *experiment* depends on all previous responses. In particular, let $I = \{1, \ldots, M\}$ be the *experiment pool*, and let $\delta_n \in I$ be the selected nth experiment. Then, we have $\delta_n = j_n(X_1, \ldots, X_{n-1})$, where the function $j_n(\cdot)$ is the experiment selection rule that needs to be learned. We use J to denote the sequence of experiment selection rules $\{j_n : n = 1, 2, \ldots\}$.

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 assume there exists at least one experiment $\delta \in I$, such that the Kullback–Leibler divergence is bounded away from zero and infinity; that is,

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

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. Note 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. Here, we consider 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, we can decide whether to continue collecting at least one more response or to stop the test. Upon stopping, we then choose H_0 or H_1. We denote the decision rule by D, where $D = 1$ indicates that H_1 is chosen, and $D = 0$ means H_0 is chosen.

The test procedure with experiment selection rule J, stopping rule N, and decision rule D is called as an *adaptive sequential design.* Our goal is to determine the optimal $J^\dagger$, $N^\dagger$, and $D^\dagger$ that 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 a *Bayesian decision framework.* In particular, we introduce the class prior

$$\pi_0 = \mathbb{P}(\theta = 0) \quad \text{and} \quad \pi_1 = \mathbb{P}(\theta = 1), \tag{3.1}$$

with $\pi_0 + \pi_1 = 1$. We assume that π_1 is known for the single hypothesis testing problem, because it is not possible 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 a wrong decision plus the expected labeling cost:

$$\begin{aligned}\mathbf{R}(J, N, D) = {} & \pi_0 \mathbb{P}(D = 1|\theta = 0) + \pi_1 \mathbb{P}(D = 0|\theta = 1) \\ & + c\{\pi_0 \mathbb{E}(N|\theta = 0) + \pi_1 \mathbb{E}(N|\theta = 1)\}.\end{aligned} \tag{3.2}$$

Note that the *relative cost* c, which is used to balance the labeling accuracy and cost, needs to be set between zero and one. Because $\mathbb{P}(D = 1|\theta = 0) \leq 1$ and $\mathbb{P}(D = 0|\theta = 1) \leq 1$, we minimize the Bayes risk in (3.2) by stopping collecting labels when $c > 1$. In practice, the requestor usually chooses c depending on the nature of the labeling task (e.g., a smaller c for more challenging data) and the budget (e.g., a large c for a limited budget). We 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 procedures $\mathcal{A}^T$ (depending on the truncation length T) if

$$\mathbf{R}(J^\dagger, N^\dagger, D^\dagger) = \min_{(J,N,D)\in\mathcal{A}^T} \mathbf{R}(J, N, D). \tag{3.3}$$

Now, for K objects with true label $\theta_k \in \{0,1\}$, for $1 \le k \le K$, we consider K hypothesis testing problems. Given $\theta_1, \ldots, \theta_K$, we assume that the responses $(X_{kj}; j = 1, 2, \ldots)$ obtained for the kth object are independent across 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 using the following *averaged loss* over K objects:

$$L_K = \frac{1}{K}\sum_{k=1}^{K}\left[\mathbf{1}_{\{D_k\neq\theta_k\}} + cN_k\right]. \tag{3.4}$$

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 there are M workers (i.e., experiments), and 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, for $i \in I$. The quality of worker i is characterized by two quantities:

$$\tau_{00}^i = \mathbb{P}(\widehat{\theta}^i = 0|\theta = 0) \quad \text{and} \quad \tau_{11}^i = \mathbb{P}(\widehat{\theta}^i = 1|\theta = 1). \tag{3.5}$$

Here, 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 the "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 τ_{00}^i and τ_{11}^i are given; in Section 5, we discuss how to estimate these parameters online as the labeling process continues.

The observed responses X_n, for $n = 1, 2, \ldots$, are the labels provided by the selected nth worker δ_n, according to the worker selection rule $j_n(X_1, \ldots, X_{n-1})$.

Under the two-coin model in (3.5), each response takes a binary value, with the following probability mass function:

$$f_{\theta,\delta_n}(1) = \mathbb{P}(X_n = 1|\delta_n, \theta) = \tau_{11}^{\delta_n}\ \mathbf{1}_{\{\theta=1\}} + (1 - \tau_{00}^{\delta_n})\ \mathbf{1}_{\{\theta=0\}}, \tag{3.6}$$
$$f_{\theta,\delta_n}(0) = \mathbb{P}(X_n = 0|\delta_n, \theta) = (1 - \tau_{11}^{\delta_n})\ \mathbf{1}_{\{\theta=1\}} + \tau_{00}^{\delta_n}\ \mathbf{1}_{\{\theta=0\}},$$

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

4. Optimal Ada-SPRT

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 such a 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 (3.3) is challenging because both the experiment selection and the stopping rule lie in infinite-dimensional function spaces. Our approach reduces the number of dimensions by exploring the relationship between optimal adaptive sequential design and a *log-likelihood ratio statistic*.

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

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

The corresponding log-likelihood ratio statistic is defined by

$$l_n^\dagger = \log\left(\frac{\prod_{i=1}^n f_{1,\delta_i^\dagger}(X_i)}{\prod_{i=1}^n f_{0,\delta_i^\dagger}(X_i)}\right), \quad \text{for } n = 1, 2, \ldots, \tag{4.1}$$

where $f_{1,\delta_i^\dagger}(\cdot)$ and $f_{0,\delta_i^\dagger}(\cdot)$ are the probability density/mass functions, respectively, when $\theta = 1$ and $\theta = 0$ in experiment $\delta_n^\dagger$. 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* (3.3)*. Then,* $(J^\dagger, N^\dagger, D^\dagger)$ *has the following properties.*

(i) The stopping time $N^\dagger$ *is described using the hitting boundary of the log-likelihood ratio and current sample size. In particular, there exist two se-*

quences of real values $A^\dagger(n)$ and $B^\dagger(n)$, for $1 \le n \le T$, such that

$$\log \frac{\pi_0}{\pi_1} = A^\dagger(T) \le A^\dagger(T-1) \le \cdots \le A^\dagger(1) \le \log \frac{\pi_0(1-c)}{\pi_1 c}, \quad (4.2)$$

$$\log \frac{\pi_0 c}{\pi_1(1-c)} \le B^\dagger(1) \le B^\dagger(2) \le \cdots \le B^\dagger(T) = \log \frac{\pi_0}{\pi_1}, \quad (4.3)$$

and the optimal stopping for the truncated test is determined by

$$N^\dagger = \inf\{n : l_n^\dagger \ge A^\dagger(n) \text{ or } l_n^\dagger \le B^\dagger(n)\}. \quad (4.4)$$

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

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

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

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

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

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

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

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

The proof of Theorem 1 is provided in the Supplementary Material. Statements (i) and (ii) are extensions of the SPRT (Wald and Wolfowitz (1948)) to 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 nonincreasing curve (i.e., the upper boundary $A^\dagger(T) \le A^\dagger(T-1) \le \cdots \le A^\dagger(1)$) and a nondecreasing curve (i.e., the lower boundary $B^\dagger(1) \le B^\dagger(2) \le \cdots \le B^\dagger(T)$). Note that because $A^\dagger(T)$ and $B^\dagger(T)$ take the same value $\log \pi_0/\pi_1$, the optimal stopping time $N^\dagger$ defined in (4.4) automatically satisfies the constraint $N^\dagger \le T$. The experiment selection rule depends on both the log-likelihood ratio statistic in (4.1) 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

$$G(l,T) = \min\{\pi(\theta=0|l), \pi(\theta=1|l)\} + Tc, \tag{4.5}$$

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 a wrong decision. The term Tc is the cost of collecting T responses. From standard Bayesian decision theory (see, e.g., Tartakovsky, Nikiforov and Basseville (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 expected conditional risk of the next stage; that is,

$$j^\dagger(l,n+1) = \operatorname*{argmin}_{\delta\in I} \mathbb{E}_{l,\delta} G\left(l + \log\frac{f_{1,\delta}(X)}{f_{0,\delta}(X)},\ n+1\right), \tag{4.6}$$

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 compute $\mathbb{E}_{l,\delta}G(l+\log(f_{1,\delta}(X)/f_{0,\delta}(X)),\ n+1)$ when $n = T-1$ (corresponding to the first step in the dynamic programming algorithm). In particular, we consider the two-coin model in (3.6) for the ith 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

$$\mathbb{P}_{l,i}(X=1) = \pi(\theta=0|l)f_{0,i}(1) + \pi(\theta=1|l)f_{1,i}(1)$$

$$= \frac{\pi_0}{\pi_0 + \pi_1 e^l}(1 - \tau_{00}^i) + \frac{\pi_1 e^l}{\pi_0 + \pi_1 e^l}\tau_{11}^i.$$

Combining the above two equations and (4.5), we have

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

Now, we are ready to provide the recursive equation for $G(l,n)$, which is known as the Bellman equation in Markov decision processes (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 is

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

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 $G(l,n)$:

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

Finally, let $C(n)$ be the set of log-likelihood ratios 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, respectively, of the log-likelihood ratio in $C(n)$. Based on the discussion thus far, we present a dynamic programming algorithm for the truncated test in Algorithm 1.

Algorithm 1 Dynamic Programming for truncated Ada-SPRT

```
Inputs:
    $T$, $c$, $\pi_0$, $\pi_1$, $\{f_{1,\delta}(\cdot)\}_{\delta\in I}$, $\{f_{0,\delta}(\cdot)\}_{\delta\in I}$
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
    $j^\dagger(l,n+1) \leftarrow \arg\min_{\delta\in I} \mathbb{E}_{l,\delta} G\left(l + \log\frac{f_{1,\delta}(X)}{f_{0,\delta}(X)},\ n+1\right)$
    $G(l,n) \leftarrow \min\Bigg\{\mathbb{E}_{l,j^\dagger(l,n+1)} G\left(l + \log\frac{f_{1,j^\dagger(l,n+1)}(X)}{f_{0,j^\dagger(l,n+1)}(X)}, n+1\right),$
        $\min\left\{\frac{\pi_0}{\pi_0+\pi_1 e^l}, \frac{\pi_1 e^l}{\pi_0+\pi_1 e^l}\right\} + nc\Bigg\}.$
    $C(n) \leftarrow \Bigg\{l : \min\left\{\frac{\pi_0}{\pi_0+\pi_1 e^l}, \frac{\pi_1 e^l}{\pi_0+\pi_1 e^l}\right\} + nc$
        $\geq \mathbb{E}_{l,j^\dagger(l,n+1)} G\left(l + \log\frac{f_{1,j^\dagger(l,n+1)}(X)}{f_{0,j^\dagger(l,n+1)}(X)},\ n+1\right)\Bigg\}.$
    $A^\dagger(n) \leftarrow \arg\sup\{l : l \in C(n)\}$.
    $B^\dagger(n) \leftarrow \arg\inf\{l : l \in C(n)\}$.
end for
Outputs:
    $j^\dagger$, $A^\dagger(n), B^\dagger(n)$ for $n = 1, \ldots, T$.
```

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

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

4.3. Nontruncated test

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

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

$$\mathbf{R}(J^*, N^*, D^*) = \min_{(J,N,D)\in\mathcal{A}^*} \mathbf{R}(J, N, D). \tag{4.7}$$

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

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

(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; that is,

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

(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 $n = 1, 2, \ldots, N^*$,

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

The proof of the Proposition 1 is provided in the Supplementary Material.

Remark 4. Wald and Wolfowitz (1948) show 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 the two hypotheses may not be minimized simultaneously.

In contrast to the truncated case in Theorem 1, the boundaries for nontruncated tests are flat. Moreover, the selection function j^* is independent of the current sample size $n-1$, and it 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 nontruncated 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 let $\mathcal{A}^*$ be the set of all sequential adaptive designs that have a 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 Hypothesis Testing and Empirical Bayes Approach

Thus 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 $\theta_k \in \{0, 1\}$, for $k = 1, \ldots, K$, are i.i.d. following a Bernoulli distribution with parameter π_1. Given $\theta_1, \ldots, \theta_K$, we assume that $(X_{kj}; j = 1, 2, ..)$ are responses (assumed independent across k) obtained for the kth object, which has a density function given by $f_{\theta_k,\delta_{kj}}(\cdot)$, and that δ_{kj} denotes the jth experiment selected for the kth object.

Recall that $D = \{D_k\}_{k=1}^K$ 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 (3.4).

If the class prior π_1 is known, then, according to Theorem 1, the optimal design that minimizes $\mathbb{E}L_K$ runs Algorithm 1 independently for each object k. In this way, we obtain the optimal experiment selection rule (denoted by $j^{(k)}$) and boundaries, or sequence of boundaries, for the truncated case (denoted by $A^{(k)}$

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 $k = 1, \ldots, K$.
```

and $B^{(k)}$, respectively). Given $j^{(k)}$, $A^{(k)}$, and $B^{(k)}$, the requestor collects labels according to the selection rule $j^{(k)}$ for each object k, and then makes 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, we can estimate the class prior using the following *empirical Bayes* approach.

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

As the number of hypotheses K grows large and the estimate $\widehat{\pi}_1$ becomes more accurate, the resulting averaged loss L_K in (3.4) converges to the minimal Bayes risk corresponding to the true π_1. We characterize this asymptotic result in the next theorem.

Theorem 3. *Assume $c < \pi_1 < 1 - c$, $\widehat{\pi}_1 \to \pi_1$ in probability as $K \to \infty$, and the sequential adaptive design D_k and N_k are determined using 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 (4.7).

That is, the averaged loss L_K in (3.4) *converges to the minimal Bayes risk under the true model.*

The proof of Theorem 3 is provided in the Supplementary Material. Note that in Theorem 3, the assumption $c < \pi_1 < 1 - c$ is a necessary condition for the optimal test procedure to be nontrivial, because without it, the optimal test will always stop with no sample.

Remark 5. In crowdsourcing applications, the classification results are usually not accurate, owning to the limited number of labels. Thus, the average performance L_K (defined as in (3.4)) is a common choice of error metric in practice when there are many labeling tasks. The metric treats type-I and type-II errors symmetrically, as in many classification problems.

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

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 the case that no prior knowledge on workers' quality parameters $\{\tau_{00}^i\}_{i\in I}$, $\{\tau_{11}^i\}_{i\in I}$ in (3.6) 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 estimate the workers' quality parameters using a regularized maximum likelihood estimate under the two-coin model in (3.6) after finishing the labeling process for each object k. In particular, after each for-loop in Algorithm 2 (i.e., when the labeling process and the decision for the kth object are complete), we have collected all responses $\{Z_{ji}\}$, where each Z_{ji} is a binary label from worker $i \in I$ to object $j \in \{1, \ldots, k\}$. A regularized minus log-likelihood is defined as follows:

$$\begin{aligned} h_k\left(\pi_1, \{\tau_{00}^i\}_{i\in I}, \{\tau_{11}^i\}_{i\in I}\right) = -\sum_{1\leq j\leq k} \log\Bigg(& (1-\pi_1)\prod_i (\tau_{00}^i)^{1-Z_{ji}}(1-\tau_{00}^i)^{Z_{ji}} \\ & +\pi_1 \prod_i (1-\tau_{11}^i)^{1-Z_{ji}}(\tau_{11}^i)^{Z_{ji}}\Bigg) + \sum_{i\in I}\big((\alpha-1)\log(\tau_{00}^i) \\ & +(\beta-1)\log(1-\tau_{00}^i) + (\alpha-1)\log(\tau_{11}^i) \\ & +(\beta-1)\log(1-\tau_{11}^i)\big). \end{aligned} \tag{5.1}$$

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, that make the estimation stable when a worker has labeled only a small number of objects. We minimize $h_k(\pi_1, \{\tau_{00}^i\}_{i\in I}, \{\tau_{11}^i\}_{i\in I})$ at the end of the kth iteration in Algorithm 2 using the expectation maximization (EM) algorithm (Dempster, Laird and Rubin (1977)). This simultaneously estimates the class prior π_1 (i.e., $\widehat{\pi}_1^{(k)}$) and the workers' quality parameters, $\{\tau_{00}^i\}_{i\in I}$ and $\{\tau_{11}^i\}_{i\in I}$ (see Dawid and Skene (1979)). These estimates are 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 previously collected responses. We also adopt the estimate from the kth iteration as the starting point (so-called warm-start) so that the EM algorithm converges within a few iterations.

6. Experimental Results

In this section, we demonstrate the performance of the proposed Ada-SPRT using two benchmark binary labeling crowdsourcing data sets. We also conduct extensive simulation studies in the Supplementary Material. A brief summary of the two real data sets is provided below.

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

Table 1. Performance comparison on real data sets in terms of the mean and standard deviation of the accuracy of different approaches. KG and Opt-KG correspond to the knowledge gradient and optimistic knowledge gradient worker selection policy, respectively, with the same stopping time as that of Ada-SPRT. KG Avg and Opt-KG Avg correspond to the knowledge gradient and 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	3,438(60)	3,949 (46)	4,365 (28)	4,660(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	1,253 (37)	1,392 (40)	1,523 (47)	1,672 (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%)

2) Labeling bird species (Bird data set (Liu, Peng and Ihler (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 there are 4,212 labels. We use the truncated Ada-SPRT with a truncation length of $T = 39$.

Note that the true labels are available for both data sets from domain experts. As a result, we can evaluate the labeling accuracy of the decision D_k for each object $k \in \{1, \ldots, K\}$.

For both data sets, we use the truncated Ada-SPRT algorithm with an EM algorithm to estimate the class prior and the workers' quality parameters, as described in Section 5. We set $\alpha = 4$ and $\beta = 2$ in the regularized likelihood function in (5.1). Because α and β reflect the prior belief of workers' accuracy, $\alpha = 4$ and $\beta = 2$ correspond to a prior accuracy of $alpha/(\alpha + \beta) = 4/(4 + 2) =$ 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 a random guess).

Because different orderings of objects in Algorithm 2 lead 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 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 ranges from 2^{-6} to 2^{-12} are 2.4, 3.2, 3.9, and 4.4, respectively, for the RTE data set. For the bird data set, the averaged stopping times as c ranges 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, described in Chen, Lin and Zhou (2015): (1) the knowledge gradient (KG) policy; and (2) the optimistic knowledge gradient (Opt-KG) policy. Note that both are myopic index policies 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; and (2) using the average stopping time $\lceil K^{-1} \sum_{i=1}^{K} N_k \rceil$ for all objects. Recall that N_k is the stopping time obtained using Ada-SPRT in Algorithm 2 for the kth object. We vary the cost parameter c and report the mean and standard deviation of the total number of queried labels (i.e., $\sum_{i=1}^{K} N_k$) and labeling accuracies for the different approaches.

The comparison results are provided in Table 1 for the RTE and Bird data sets. As shown in Table 1, Ada-SPRT outperforms the other approaches on both data sets. Under the two-coin model, when using all available labels, the labeling accuracy is 92.88% (with 8,000 labels) for the RTE data set and 89.1% (for $4,212$ labels) for the Bird data set. Therefore, from Table 1, Ada-SPRT achieves, on average, $92.1/92.88 = 99\%$ of the best possible labeling accuracy using only $3{,}438/8{,}000 = 43\%$ of the total labels for RTE, and $85.7/89.1 = 96\%$ of the best possible labeling accuracy using only $1{,}253/4{,}212 = 30\%$ of the total labels for Bird.

7. Conclusion

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

There are several directions in which this work may be extended. First, we consider only simple versus simple hypotheses for the binary labeling tasks. It would be worth extending the current framework to include 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 tests. Thus, future research will identify and investigate additional applications.

Supplementary Material

In the online Supplementary Material, we present proofs of the theoretical results, including Proposition 1, Theorems 1, 2, and 3, and the supporting lemmas, as well as additional simulated experiments.

Acknowledgments

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

References

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

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

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

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

Babaioff, M., Sharma, Y. and Slivkins, A. (2009). Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, 79–88.

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

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

Bartroff, J., Finkelman, M. and Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika* **73**, 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**, 1593–1611.

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

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. (Statistical Methodology)* **57**, 289–300.

Bertsekas, D. P. and Shreve, S. E. (1978). *Stochastic Optimal Control: The Discrete Time Case*, volume 23. Academic Press, New York.

Bessler, S. A. (1960). *Theory and Applications of The Sequential Design of Experiments, k-Actions and Infinitely Many Experiments*. Technical Report, 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**, 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**, 1–122.

Bussgang, J. and Middleton, D. (1955). Optimum sequential detection of signals in noise. *IRE Transactions on Information Theory* **1**, 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 21*.

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

Chang, Y. I. (2005). Application of sequential interval estimation to adaptive mastery testing. *Psychometrika* **70**, 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**, 1–46.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics* **30**, 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. (Applied Statistics)* **28**, 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. (Statistical Methodology)* **39**, 1–38.

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

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

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

Prediction. Cambridge University Press, Cambridge.

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**, 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**, 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**, 91–104.

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

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

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**, 1–24.

Karunamuni, R. J. (1988). On empirical Bayes testing with sequential components. *The Annals of Statistics* **16**, 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**, 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**, 674–685.

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

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

Lewis, C. and Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *ETS Research Report Series* **14**, 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**, e167.

Marcus, A. and Parameswaran, A. (2015). Crowdsourced data management: Industry and academic perspectives. *Foundations and Trends in Databases* **6**, 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**, 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**, 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**, 768–782.

Nikiforov, I. V. (1975). Sequential analysis applied to autoregression processes. *Avtomatika i Telemekhanika* **36**, 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**, 1–24.

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**, 22387–22392.

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

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**, 1–6.

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**, 1297–1322.

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

Robbins, H. and Siegmund, D. O. (1974). Sequential tests involving two populations. *Journal of the American Statistical Association* **69**, 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**, 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**, 502–522.

Song, Y. and Fellouris, G. (2019). Sequential multiple testing with generalized error control: An asymptotic optimality theory. *The Annals of Statistics* **47**, 1776–1803.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics* **31**, 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**, 814–817.

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

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

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

Wald, A. and Wolfowitz, J. (1950). Bayes solutions of sequential decision problems. *The Annals of Mathematical Statistics* **21**, 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**, 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**, 1–44.

Xiaoou Li
School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.
E-mail: lixx1766@umn.edu
Yunxiao Chen
Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, UK.
E-mail: y.chen186@lse.ac.uk
Xi Chen
Stern School of Business, New York University, New York, NY 10003, USA.
E-mail: xichen@nyu.edu
Jingchen Liu
Department of Statistics, Columbia University, New York, NY 10027, USA.
E-mail: jcliu@stat.columbia.edu
Zhiliang Ying
Department of Statistics, Columbia University, New York, NY 10027, USA.
E-mail: zying@stat.columbia.edu

(Received July 2018; accepted May 2019)