A RATE FUNCTION APPROACH TO THE COMPUTERIZED ADAPTIVE TESTING FOR COGNITIVE DIAGNOSIS

Jingchen Liu, Zhiliang Ying, and Stephanie Zhang

columbia university

June 16, 2013

Correspondence should be sent to Jingchen Liu

E-Mail: jcliu@stat.columbia.edu
Phone: 212.851.2146
Fax: 212.851.2164
Website: http://stat.columbia.edu/~jcliu

A RATE FUNCTION APPROACH TO COMPUTERIZED ADAPTIVE TESTING FOR COGNITIVE DIAGNOSIS

Abstract

Computerized adaptive testing (CAT) is a sequential experiment design scheme that tailors the selection of experiments to each subject. Such a scheme measures subjects' attributes (unknown parameters) more accurately than the regular pre-fixed design. In this paper, we consider CAT for diagnostic classification models, for which attribute estimation corresponds to a classification problem. After a review of existing methods, we propose an alternative criterion based on the asymptotic decay rate of the misclassification probabilities. The new criterion is then developed into new CAT algorithms, which are shown to achieve the asymptotically optimal misclassification rate. Simulation studies are conducted to compare the new approach with existing methods, demonstrating its effectiveness, even for moderate length tests.

Key words: Computerized adaptive testing, cognitive diagnosis, large deviation, classification

1. Introduction

Cognitive diagnosis has recently gained prominence in educational assessment, psychiatric evaluation, and many other disciplines. Various modeling approaches have been discussed in the literature both intensively and extensively (e.g. K. K. Tatsuoka, 1983). A short list of such developments includes the rule space method (K. K. Tatsuoka, 1985, 2009), the reparameterized unified/fusion model (RUM) (DiBello, Stout, & Roussos, 1995; Hartz, 2002; Templin, He, Roussos, & Stout, 2003), the conjunctive (noncompensatory) DINA and NIDA models (Junker & Sijtsma, 2001; de la Torre & Douglas, 2004), the compensatory DINO and NIDO models (Templin & Henson, 2006), the attribute hierarchy method (Leighton, Gierl, & Hunka, 2004). C. Tatsuoka (2002) discussed a model usually referred to as the conjunctive DINA model that has both conjunctive and disjunctive components in its attribute specifications to allow for multiple strategies and a discussion of identifiability is also provided. See also Rupp, Templin, and Henson (2010) for more approaches to cognitive diagnosis.

Another important development in educational measurement is computerized adaptive testing (CAT), that is, a testing mode in which item selection is sequential and individualized to each subject. In particular, subsequent items are selected based on the subject's (e.g. examinee's) responses to prior items. CAT was originally proposed by Lord (1971) for item response theory (IRT) models as a method through which items are tailored to each examinee to 'best fit' his or her ability level θ. More capable examinees avoid receiving problems that are too simple and less capable examinees avoid receiving problems that are too difficult. Such individualized testing schemes perform better than traditional exams

with a prefixed set of items because the optimal selection of testing problems is examinee dependent. It also leads to greater efficiency and precision than that which can be found in traditional tests. For the traditional CAT under IRT settings, items are typically chosen to maximize the Fisher information (MFI) (Lord, 1980; Thissen & Mislevy, 2000) or to minimize the expected posterior variance (MEPV) (van der Linden, 1998; Owen, 1975).

It is natural to consider incorporating the CAT design into cognitive diagnosis. The sequential nature of CAT could conceivably bring major benefits to cognitive diagnosis. First, diagnostic classification models are multidimensional with simultaneous consideration of different attributes. To fully delineate each dimension with sufficient accuracy would certainly demand a large number of items to cover all attributes. Thus, the ability to reduce test length in CAT can be attractive. Second, it is often desirable to provide feedback learning or remedial training after diagnosis; see Junker (2007). The CAT design can serve as a natural framework under which a feedback online learning system may be incorporated.

A major difference between classical IRT models and diagnostic classification models (DCM) is that the parameter space of the latter is usually discrete, for the purpose of diagnosis. Thus, standard CAT methods developed for IRT (such as the MFI or MEPV) do not apply. Several alternative methods have already been developed in the literature. The parameter spaces of most DCM's admit a partially ordered structure (von Davier, 2005). Under such a setting, C. Tatsuoka and Ferguson (2003) developed a general theorem on the asymptotically optimal sequential selection of items for finite partially ordered parameter spaces. In particular, the asymptotically optimal design maximizes the convergence rate of the parameters'

posterior distribution to the true parameter. Xu, Chang, and Douglas (2003) investigated two methods based on different ideas. One method is based on the Shannon entropy of the posterior distribution. The other method is based on the Kullback-Leibler (KL) information that describes the global information of a set of items for parameter estimation. The concept of global information was introduced by Chang and Ying (1996). Cheng (2009) further extended the KL information method by taking into account the posterior distribution and the distances between the alternative parameters and the current estimate when computing the global information. Arising from such extensions are two new methods known as the posterior-weighted KL algorithm (PWKL) and the hybrid KL algorithm (HWKL). See C. Tatsuoka (2002) for a real-data application of CAT.

A key component in the study of CAT lies in evaluating the efficiency of a set of items, i.e., what makes a good selection of exam problems for a particular examinee. Efficiency is typically expressed in terms of the accuracy of the resulting estimator. In classic IRT, items are selected to carry maximal information about the underlying parameter θ. This is reflected by the MFI or the MEPV (Lord, 1980; Thissen & Mislevy, 2000; van der Linden, 1998; Owen, 1975). On the other hand, for diagnostic classification models, parameter spaces are usually discrete, and the task of parameter estimation is equivalent to a classification problem. In this paper, we address the problem of CAT for cognitive diagnosis (CD-CAT) by focusing on the misclassification probability. The misclassification probability, though conceptually a natural criterion, is typically not in a closed form. Thus, it is not a feasible criterion. Nonetheless, under very mild conditions, we show that this probability decays

exponentially fast as the number of items (m) increases, that is, $P(\hat{\alpha} \neq \alpha_0) \approx e^{-m \times I}$, where $\hat{\alpha}$ is an estimator of the true parameter α_0. We use the exponential decay rate, denoted by I, as a criterion. That is, a set of items is said to be *asymptotically optimal* if it maximizes the rate I (instead of directly minimizing the misclassification probability). The rate I is usually easy to compute and often in a closed form. Therefore, the proposed method is computationally efficient. In Section 3.3, we derive the specific form of the rate function for the Bernoulli response distribution that is popular in cognitive diagnosis. Based on the rate function I, we propose CD-CAT procedures that correspond to this idea. Simulation studies are conducted to compare among the existing methods for CD-CAT.

This paper is organized as follows. Section 2 provides a general introduction to the problem of CD-CAT, including an overview of existing methods. In Section 3 we introduce the idea of asymptotically optimal design and the corresponding CD-CAT procedures. We examine the connection of our new approach to previously developed methods in Section 4. Further discussion is provided in Section 5. Finally, Section 6 contains simulation studies.

2. Computerized adaptive testing for cognitive diagnosis

2.1. Problem setting

Let the random variable X be the outcome of an experiment e. The distribution of X depends on the experiment e and an underlying parameter α. We use $\alpha_0 \in \mathcal{A}$ to denote the true parameter and $e^1, ..., e^m \in \mathcal{E}$ to denote different experiments. In the context of cognitive diagnosis, α corresponds to the "attribute profile" or "knowledge state" of a subject, e is

an item or an exam problem, and X is the subject's response to the item e. Suppose that independent outcomes $X^1, ..., X^m$ of experiments $e^1, ..., e^m$ are collected. For each $e \in \mathcal{E}$ and $\alpha \in \mathcal{A}$, let $f(x|e, \alpha)$ be the probability density (mass) function of X. Throughout this paper, we suppose that the parameter space α takes finitely many values in $\mathcal{A} = \{1, ..., J\}$ and that there are κ types of experiments $\mathcal{E} = \{1, ..., \kappa\}$.

Let superscripts indicate independent outcomes, e.g., X^i is the outcome of the i-th experiment $e^i \in \mathcal{E}$. The experiments are possibly repeated, i.e., $e^i = e^j$ (i.e. i.i.d. outcomes can be collected from the same experiment). Suppose that the prior distribution of the parameter α is $\pi(\alpha)$. Given the observed responses $X^1, ..., X^m$ the posterior distribution of α is

$$\pi(\alpha|X^i, e^i, \text{ for } i = 1, ..., m) \propto \pi(\alpha) \prod_{i=1}^{m} f(X^i|e^i, \alpha). \tag{1}$$

To simplify the notation, we let $\mathbf{X}_m = (X^1, ..., X^m)$ and $\mathbf{e}_m = (e^1, ..., e^m)$, e.g.,

$$\pi(\alpha|\mathbf{X}_m, \mathbf{e}_m) = \pi(\alpha|X^i, e^i, \text{ for } i = 1, ..., m). \tag{2}$$

Thus, a natural estimator of α is the posterior mode

$$\hat{\alpha}(\mathbf{X}_m, \mathbf{e}_m) = \arg \sup_{\alpha \in \mathcal{A}} \pi(\alpha|\mathbf{X}_m, \mathbf{e}_m). \tag{3}$$

For any two distributions P and Q, with densities p and q, the Kullback-Leibler (KL) divergence/information is

$$D_{KL}(P||Q) = E_P[\log(p(X)/q(X))] \tag{4}$$

For two parameter values α_0 and α_1, let $D_e(\alpha_0, \alpha_1) = E_{e,\alpha_0}\big\{\log[f(X|e,\alpha_0)/f(X|e,\alpha_1)]\big\}$ be the KL divergence between the outcome distributions of experiment e, where the notation E_{e,α_0} indicates that X follows distribution $f(x|e,\alpha_0)$. We say that an experiment e *separates* parameter values α_0 and α_1 if $D_e(\alpha_0,\alpha_1) > 0$. Note that $D_e(\alpha_0,\alpha_1) = 0$ if the two distributions $f(x|e,\alpha_0)$ and $f(x|e,\alpha_1)$ are identical and statistically indistinguishable. Thus, if $D_e(\alpha_0,\alpha_1) > 0$ and independently and identically distributed (i.i.d.) outcomes can be generated from the experiment e, the parameters α_0 and α_1 are eventually distinguishable. An experiment with a large value of $D_e(\alpha_0,\alpha_1)$ is powerful in differentiating α_0 and α_1. To simplify the discussion, we assume that for each pair of distinct parameters $\alpha_0 \neq \alpha_1$ there exists an experiment $e \in \mathcal{E}$ that separates α_0 and α_1; otherwise, we simply merge the inseparable parameters and reduce the parameter space. This is discussed in the supplemental material (Appendix B); see also Chiu, Douglas, and Li (2009); K. Tatsuoka (1991); C. Tatsuoka (1996) for discussions of parameter identifiability for specific models. To illustrate these ideas, we provide two stylized examples that are frequently considered.

Example 1. (Partially ordered sets, C. Tatsuoka and Ferguson (2003)) Consider that the parameter space $\mathcal{A}$ is a partially ordered set with a binary relation "$\leq$". The set of experiment is identical to the parameter space, i.e. $\mathcal{A} = \mathcal{E}$. The outcome distribution of experiment e and parameter α is given by $f(x|e,\alpha) = f(x)$ if $e \leq \alpha_0$, and $f(x|e,\alpha) = g(x)$ otherwise.

The following example can be viewed as an extension of Example 1 where the distributions f and g are experiment-dependent.

Example 2. (DINA model, Junker and Sijtsma (2001)) Consider a parameter space $\alpha =$

$(a_1, ..., a_k) \in \mathcal{A} = \{0,1\}^k$ and $e = (\varepsilon_1, ..., \varepsilon_k) \in \{0,1\}^k$. Under the context of educational testing, each a_i indicates if a subject possesses a certain skill. Each experiment corresponds to one exam problem and ε_i indicates if this problem requires skill i. A subject is capable of solving an exam problem if and only if he or she possesses all the required skills, i.e., $e \leq \alpha$, defined as $\varepsilon_i \leq a_i$ for all $i = 1, ..., k$. The outcome in this context is typically binary: $X = 1$ for the correct solution to the exam problem and $X = 0$ for the incorrect solution. We let $\xi = \mathbf{1}(e \leq \alpha)$ be the ideal response. The outcome follows a Bernoulli distribution

$$P(X = 1|e, \alpha) = \begin{cases} 1 - s_e & \text{if } \xi = 1 \\ g_e & \text{otherwise} \end{cases}.$$

The parameter s_e is known as the slipping parameter and g_e is the guessing parameter. Both the slipping and the guessing parameters are experiment specific. The general form of DINA model allows heterogeneous slipping and guessing parameters for different exam problems with identical skill requirements. Thus, in addition to the attribute requirements, the model also specifies the slipping and the guessing parameters for each exam problem.

In practice, there may not be completely identical items. For instance, one may design two exam problems requiring precisely the same skills. However, it is difficult to ensure the same slipping and the guessing parameters. Thus, we can only expect independent (but not identically distributed) outcomes. In the previous discussion, we assume that i.i.d. outcomes can be collected from the same experiment. This assumption is imposed simply to reduce the complexity of the theoretical development, and is not really required by the proposed

CAT procedures (Algorithm 1). More discussion on this issue is provided in Remark 2.

2.2. Existing methods for the CD-CAT

2.2.1. Asymptotically optimal design by Tatsuoka and Ferguson (2003)

Tatsuoka and Ferguson (2003) proposes a general theorem on the asymptotically optimal selection of experiments when the parameter space is a finite and partially ordered set. It is observed that the posterior probability of the true parameter α_0 converges to one exponentially fast, that is, $1 - \pi(\alpha_0|\mathbf{X}_m, \mathbf{e}_m) \approx e^{-m\times H}$ as $m \to \infty$. The authors propose the selection of experiments (items) that maximize the asymptotic convergence rate H.

In particular, the asymptotically optimal selection of experiments can be represented by the KL divergence in the following way. Let h_e be the proportion of experiment e among the m experiments. For each alternative $\alpha_1 \neq \alpha_0$, define $D_{\mathbf{h}}(\alpha_0, \alpha_1) = \sum_{e\in\mathcal{E}} h_e D_e(\alpha_0, \alpha_1)$, where $\mathbf{h} = (h_1, \ldots, h_\kappa)$ and $\sum_j h_j = 1$. Then, the asymptotically optimal selection solves the optimization problem $\mathbf{h}^* = \arg\max_{\mathbf{h}}[\min_{\alpha_1\neq\alpha_0} D_{\mathbf{h}}(\alpha_0, \alpha_1)]$. The authors show that several procedures achieve the asymptotic optimal proportion $\mathbf{h}^*$ under their setting.

2.2.2. The KL divergence based algorithms

There are several CD-CAT methods based on the Kullback-Leibler divergence. The basic idea is to choose experiments such that the distribution of the outcome X associated with the true parameter α_0 looks most dissimilar to the distributions associated with the alternative parameters. The initial idea was proposed by Chang and Ying (1996), who define the global

information by summing the KL information over all possible alternatives, i.e.,

$$KL_e(\alpha_0) = \sum_{\alpha \neq \alpha_0} D_e(\alpha_0, \alpha). \tag{5}$$

If an experiment e has a large value of $KL_e(\alpha_0)$, then the outcome distributions associated with α_0 and the alternative parameters are very different. Thus, e is powerful in differentiating the true parameter α_0 from other parameters. For a sequential algorithm, let $\hat{\alpha}_m$ be the estimate of α based on the first m outcomes. The next experiment is chosen to maximize $KL_e(\hat{\alpha}_m)$ (Xu et al., 2003).

This idea is further extended by Cheng (2009), who proposes the weighting of each $D_e(\hat{\alpha}_m, \alpha_1)$ in (5) by the posterior probability conditional on $\mathbf{X}_m$, that is, each successive experiment maximizes $PWKL_e(\hat{\alpha}_m) = \sum_{\alpha \neq \alpha_0} D_e(\hat{\alpha}_m, \alpha)\pi(\alpha|\mathbf{X}_m, \mathbf{e}_m)$. An α with a higher value of $\pi(\alpha|\mathbf{X}_m, \mathbf{e}_m)$ is more difficult to differentiate from the posterior mode $\hat{\alpha}_m$. Thus, it carries more weight when choosing subsequent items. This method is known as the posterior-weighted Kullback-Leiber (PWKL) algorithm. The author also proposes a hybrid method that adds to $D_e(\hat{\alpha}_m, \alpha)$ further weights inversely proportional to the distance between α and $\hat{\alpha}_m$, so that alternative parameters closer to the current estimate receives even more weight.

2.2.3. The SHE algorithm

The Shannon entropy of the posterior distribution is defined as

$$H(\pi(\cdot|\mathbf{X}_m, \mathbf{e}_m)) \triangleq -\sum_{\alpha \in \mathcal{A}} \pi(\alpha|\mathbf{X}_m, \mathbf{e}_m) \log \pi(\alpha|\mathbf{X}_m, \mathbf{e}_m) = \log \kappa - D(\pi(\cdot|\mathbf{X}_m, \mathbf{e}_m) \parallel U_{\mathcal{A}}(\cdot)),$$

where $U_{\mathcal{A}}$ is the uniform distribution on the set $\mathcal{A}$ and $D(\cdot||\cdot)$ is the KL divergence defined in (4). Thus, the experiment that minimizes the Shannon entropy of the posterior distribution makes the posterior distribution as different from the uniform distribution as possible. In particular, let $f(x^{m+1}|e, \mathbf{X}_m, \mathbf{e}_m)$ be the posterior predictive distribution of the $(m+1)$-th outcome if the $(m+1)$-th experiment is chosen to be e. The sequential item selection algorithm chooses e^{m+1} to minimize the expected Shannon entropy $SHE(e)$, where

$$SHE(e) = \int H(\pi(\cdot|\mathbf{X}_m, X^{m+1} = x^{m+1}, \mathbf{e}_m, e^{m+1} = e)) f(x^{m+1}|e, \mathbf{X}_m, \mathbf{e}_m) dx^{m+1}.$$

The idea of minimizing the Shannon entropy is very similar to that of the minimum expected posterior variance method developed for IRT.

3. The misclassification probability, optimal design, and CAT

In this section, we present the main method of this paper. For a discrete parameter space, estimating the true parameter value is equivalent to classifying a subject into one of J groups. Given that the main objective is the estimation of the attribute parameter α, a natural goal of optimal test design would be the minimization of the misclassification probability. In the decision theory framework, this probability corresponds to the Frequentist risk associated with the zero-one loss function; see Chapter 11 of Cox and Hinkley (2000). Let α_0 denote the true parameter. The misclassification probability of some estimator $\hat{\alpha}(\mathbf{X}_m)$ based on m experiments is then

$$p(\mathbf{e}_m, \alpha_0) = P_{\mathbf{e}_m, \alpha_0}(\hat{\alpha}(\mathbf{X}_m) \neq \alpha_0). \tag{6}$$

We write $\mathbf{e}_m$ and α_0 in the subscript to indicate that the outcomes $X_1, ..., X_i$ are independent outcomes from $f(X^i|e^i, \alpha_0)$ respectively. Similarly, we will use $E_{\mathbf{e}_m,\alpha_0}$ to denote the corresponding expectation. Throughout this paper, we consider $\hat{\alpha}$ to be the posterior mode in (3). If one uses a uniform prior over the parameter space $\mathcal{A}$, i.e., $\pi(\alpha) = \frac{1}{J}$, then the posterior mode is identical to the maximum likelihood estimate. Thus, the current framework includes the situation in which the MLE is used. Under mild conditions, one can show that $p(\mathbf{e}_m, \alpha_0) \to 0$ as $m \to \infty$. A good choice of items should admit small $p(\mathbf{e}_m, \alpha_0)$. However, direct use of $p(\mathbf{e}_m, \alpha_0)$ as an efficiency measure is difficult, mostly due to the following computational limitations. The probability (6) is usually not in an analytic form. Regular numerical routines (such as Monte Carlo methods) fail to produce accurate estimate of $p(\mathbf{e}_m, \alpha_0)$. For instance, when $m = 50$, this probability could be as small as a few percentage points. Evaluating such a probability for a given relative accuracy is difficult, especially this probability has to be evaluated many times – essentially once for each possible combination of items. Therefore, (6) is not a feasible criterion from a computational viewpoint. Due to these concerns, we propose the use of an approximation of (6) based on large deviations theory. In particular, as we will show, under very mild conditions, the following limit can be established

$$-\frac{\log p(\mathbf{e}_m, \alpha_0)}{m} \to I \quad \text{as } m \to \infty. \tag{7}$$

That is, the misclassification probability decays to zero exponentially fast and it can be approximated by $p(\mathbf{e}_m, \alpha_0) \approx e^{-mI}$. We call the limit I the *rate function* that depends on both the experiment selection and the true parameter α_0. The selection of experiments

that maximizes the rate function is said to be *asymptotically optimal* in the sense that the misclassification probability based on the asymptotically optimal design achieves the same exponential decay rate as the optimal design that minimizes the probability in (6). In addition, the rate function has favorable properties from a computation point of view. It only depends on the proportion of each type of experiments. Therefore, the asymptotically optimal proportion does not depend on the total number of experiments m, which simplifies the computation. In addition, the rate I is in closed form for most standard distributions. Minimizing the misclassification probability is equivalent to adopting a zero-one loss function. In practice, there may be other sensible loss functions under specific scenarios. In this paper, we focus on the zero-one loss and the misclassification probability.

We emphasize that the asymptotic optimality discussed in the previous paragraph is different from the one in C. Tatsuoka and Ferguson (2003). In fact, these two criteria often yield different "optimal designs." To make a difference we refer to the latter as "Tatsuoka and Ferguson's asymptotic optimality" and reserve the term asymptotic optimality, as specified momentarily in Definition 1, for designs that maximize the rate function I.

3.1. The missclassification rate function and the asymptotically optimal design

We now consider the misclassification rate function. To facilitate the discussion, we assume that $m \times h_e$ independent outcomes are collected from experiment $e \in \mathcal{E}$. Furthermore, we assume that the proportion h_e does not change with m, except for some slight variation due to rounding. We say that such a selection of h_e is *stable*. Under the asymptotic regime where $m \to \infty$, the parameter $\mathbf{h} = (h_1, ..., h_\kappa)$ forms the exogenous experiment design

parameter to be tuned. Under this setting, the rate function (when exists) is

$$-\frac{1}{m}\log p(\mathbf{e}_m, \alpha_0) \to I_{\mathbf{h},\alpha_0}. \tag{8}$$

The limit depends on the proportion of each experiment contained in $\mathbf{h}$ and the true parameter α_0. We establish the above approximation and provide the specific form of $I_{\mathbf{h},\alpha_0}$ in Theorem 2.

For two sets of experiments corresponding to two vectors $\mathbf{h}_1$ and $\mathbf{h}_2$, if $I_{\mathbf{h}_1,\alpha_0} > I_{\mathbf{h}_2,\alpha_0}$, it suggests that the misclassification probability of $\mathbf{h}_1$ decays to zero at a faster speed than $\mathbf{h}_2$ and therefore $\mathbf{h}_1$ is a better design. We propose the use of $I_{\mathbf{h},\alpha_0}$ as a measure of efficiency.

Definition 1. We say that an experiment design corresponding to a set of proportions $\mathbf{h} = (h_1, ..., h_\kappa)$ is asymptotically optimal if $\mathbf{h}$ maximizes the rate function $I_{\mathbf{h},\alpha_0}$ as in (8).

One computationally appealing feature of asymptotically optimal design is that the asymptotically optimal proportion $\mathbf{h}$ generally does not depend on the particular form of the prior distribution. Since asymptotic optimality describes the amount of information in the data, item selection is relatively independent from the a priori information of the attributes.

3.2. The analytic form of the rate function

In this subsection, we present specific forms of the rate function. To facilitate discussion, we use a different set of superscripts. Among the m responses, $m \times h_e$ of them are from experiment e. Let $X^{e,l}$ be the l-th (independent) outcomes of type e experiments for $l = 1, ..., m \times h_e$. Note that the notation e includes all the information about an experiment. For instance, in the setting of the DINA model, e includes the attribute requirements (the

Q-matrix entries) as well as the slipping and the guessing parameters.

We start the discussion with a specific alternative parameter α_1. The posterior distribution prefers an alternative parameter $\alpha_1 \neq \alpha_0$ to α_0 if

$$\pi(\alpha_1|\mathbf{X}_m, \mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m, \mathbf{e}_m).$$

We insert the specific form of the posterior in (1) and the above inequality implies that

$$\frac{1}{m}\Big[\log(\pi(\alpha_0)) - \log(\pi(\alpha_1))\Big] < \sum_{e\in\mathcal{E}} \frac{h_e}{m \times h_e} \sum_{l=1}^{m\times h_e} \Big[\log f(X^{e,l}|e,\alpha_1) - \log f(X^{e,l}|e,\alpha_0)\Big]. \quad (9)$$

For each e, define

$$s_{\alpha_1}^{e,l} \triangleq \log f(X^{e,l}|e,\alpha_1) - \log f(X^{e,l}|e,\alpha_0), \quad l = 1, ..., m \times h_e \quad (10)$$

that is the log-likelihood ratio between the two parameter values α_0 and α_1. For a given e and different l, $s_{\alpha_1}^{e,l}$'s are i.i.d. random variables. Therefore, the right-hand side of (9) is the weighted sum of κ sample averages of i.i.d. random variables. Due to the entropy inequality, we have that $E_{e,\alpha_0}(s_{\alpha_1}^{e,l}) = -D(\alpha_0, \alpha_1) \leq 0$. If equality occurs, experiment e does not have power in differentiating α_1 from α_0. This occurs often in diagnostic classification models such as the DINA model (Section 2.1, Example 2).

In what follows, we provide the specific form of the rate function for the probability

$$P\Big(\pi(\alpha_1|\mathbf{X}_m, \mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m, \mathbf{e}_m)\Big).$$

Let $g_e(s|\alpha_1)$ be the distribution of $s^{e,l}_{\alpha_1}$ under e and α_0, and let

$$g_e(s|\theta,\alpha_1) = g_e(s|\alpha_1)e^{\theta s - \varphi_{e,\alpha_1}(\theta)} \tag{11}$$

be its associated natural exponential family where $\varphi_{e,\alpha_1}(\theta) = \log\left[\int e^{\theta s} g_e(s|\alpha_1)ds\right]$is the log-moment-generating function. The exponential family is introduced for the purpose of defining the rate function. It is not meant for the data (response) generating process. In addition, it implicitly assumes that φ_{e,α_1} is finite in a neighborhood of the origin. Let

$$L_e(\theta_e|\alpha_1) = \theta_e\varphi'_{e,\alpha_1}(\theta_e) - \varphi_{e,\alpha_1}(\theta_e), \tag{12}$$

where φ'_{e,α_1} is the derivative. We define

$$I(\alpha_1,\mathbf{h}) = \inf_{\theta_1,...,\theta_\kappa} \sum_{e\in\mathcal{E}} h_e L_e(\theta_e|\alpha_1), \qquad \text{for } \mathbf{h} = (h_1,...,h_\kappa) \tag{13}$$

where the infimum is subject to the constraint that $\sum_{e\in\mathcal{E}} h_e\varphi'_{e,\alpha_1}(\theta_e) \geq 0$. Furthermore, we define a notation

$$I_e(\alpha_1) = I(\alpha_1,\mathbf{h}) \tag{14}$$

if all the elements of $\mathbf{h}$ are zero except for the one corresponding to the experiment e, that is, $I_e(\alpha_1)$ is the rate if all outcomes are generated from experiment e. Further discussion of evaluation of $I(\alpha_1,\mathbf{h})$ is provided momentarily in Remark 1.

The following two theorems establish the asymptotic decay rate of the misclassification probabilities. Their proofs are provided in the supplemental material (Appendix A). Recall that the vector $\mathbf{h} = (h_e : e \in \mathcal{E})$ represents the asymptotic proportions, i.e., $\frac{1}{m}\sum_{j=1}^{m} I(e^j =$

$e) \to h_e$ as $m \to \infty$.

Theorem 1. Suppose that for each $\alpha_1 \neq \alpha_0$ and each $e \in \mathcal{E}$, equation $\varphi'_{e,\alpha_1}(\theta_e) = 0$ has a solution. Then for every $\alpha_1 \neq \alpha_0$, $e \in \mathcal{E}$, and $h \in [0,1]^\kappa$ s.t. $\sum_{e\in\mathcal{E}} h_e = 1$, we have that

$$\lim_{m\to\infty} -\frac{1}{m} \log P_{\mathbf{e}_m,\alpha_0}\Big(\pi(\alpha_1|\mathbf{X}_m,\mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m,\mathbf{e}_m)\Big) = I(\alpha_1,\mathbf{h}). \tag{15}$$

Note that both $I(\alpha_1,\mathbf{h})$ and $I_e(\alpha_1)$ depend on the true parameter α_0. To simplify the notation, we omit the index of α_0 in the writing $I(\alpha_1,\mathbf{h})$ and $I_e(\alpha_1)$ because all the discussions are for the same true parameter α_0. Nonetheless, it is necessary and important to keep the dependence in mind. The rate function (15) has its root in statistical hypothesis testing. Consider testing the null hypothesis $H_0 : \ \alpha = \alpha_0$ against an alternative $H_A : \ \alpha = \alpha_1$. We reject the null hypothesis if $\pi(\alpha_1|\mathbf{X}_m,\mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m,\mathbf{e}_m)$. Thus, the misclassification probability is the same as the Type I error probability. Its asymptotic decay rate $I(\alpha_1,\mathbf{h})$ is known as the *Chernoff index* (Serfling, 1980, Chapter 10).

Remark 1. Without much effort, one can show that the constraint for the minimization in (13) can be reduced to $\sum_{e\in\mathcal{E}} h_e\varphi'_{e,\alpha_1}(\theta_e) = 0$, that is, the infimum is achieved on the boundary. Let $(\theta^*_e : e \in \mathcal{E})$ be the solution to the optimization problem. Using Lagrange multipliers, one can further simplify the optimization problem in (13) to the case in which the solution satisfies $\theta^*_1 = \cdots = \theta^*_\kappa$. Thus, the rate function can be equivalently defined as $I(\alpha_1,\mathbf{h}) = \sum_{e\in\mathcal{E}} h_e L_e(\theta|\alpha_1)$, where θ satisfies $\sum_{e\in\mathcal{E}} h_e\varphi'_{e,\alpha_1}(\theta) = 0$. Using the specific form

of L_e in (12) and the fact that φ_{e,α_1}'s are convex, we obtain that

$$I(\alpha_1, \mathbf{h}) = -\inf_{\theta} \sum_{e \in \mathcal{E}} h_e \varphi_{e,\alpha_1}(\theta). \tag{16}$$

Thus, the numerical evaluation of $I(\alpha, \mathbf{h})$ is a one dimensional convex optimization problem that can be stated in a closed form for most standard distributions.

While Theorem 1 provides the asymptotic decay rate of the probability that the posterior mode happens to be at one specific alternative parameter α_1, the next theorem gives the overall misclassification rate.

Theorem 2. Let $p(\mathbf{e}_m, \alpha_0)$ be the misspecification probability given by (6) and define the overall rate function

$$I_{\mathbf{h},\alpha_0} \triangleq \lim_{m \to \infty} -\frac{1}{m} \log p(\mathbf{e}_m, \alpha_0). \tag{17}$$

Then under the same conditions as in Theorem 1, $I_{\mathbf{h},\alpha_0} = \inf_{\alpha_1 \neq \alpha_0} I(\alpha_1, \mathbf{h})$.

Here, we include the index α_0 in the rate function $I_{\mathbf{h},\alpha_0}$ to emphasize that the misclassification probability depends on the true parameter α_0. Thus, the asymptotically optimal selection of experiments $\mathbf{h}^*$ is defined as

$$\mathbf{h}^* = \arg\sup_{\mathbf{h}} I_{\mathbf{h},\alpha_0} \tag{18}$$

An algorithm to compute $\mathbf{h}^*$ numerically is included in the supplemental material (Appendix C).

3.3. Intuitions and examples

According to Theorem 1, for a specific parameter α_1, the probability that the posterior favors α_1 over the true parameter α_0 for a given asymptotic design $\mathbf{h}$ admits the approximation $P(\hat{\alpha}(\mathbf{X}_m, \mathbf{e}_m) = \alpha_1 | e_1, ..., e_m, \alpha_0) \approx e^{-m \times I(\alpha_1, \mathbf{h})}$. Thus, the total misclassification probability is approximated by

$$p(\mathbf{e}_m, \alpha_0) = P(\hat{\alpha}(\mathbf{X}_m, \mathbf{e}_m) \neq \alpha_0 | e_1, ..., e_m, \alpha_0) \approx \sum_{\alpha_1 \neq \alpha_0} e^{-m \times I(\alpha_1, \mathbf{h})}. \tag{19}$$

Among the above summands, there is one (or possibly multiple) α' that admits the smallest $I(\alpha', \mathbf{h})$. Then, the term $e^{-mI(\alpha', \mathbf{h})}$ is the dominating term of the above sum. Note that the smaller $I(\alpha_1, \mathbf{h})$ is, the more difficult it is to differentiate between α_0 and α_1. According to the representation in (17), upon considering the overall misclassification probability, it is sufficient to consider the alternative parameter that is the most difficult to differentiate from α_0. This agrees with the intuition that if a set of experiments differentiates well between parameters that are similar to each other, it also differentiates well between less similar parameters. Thus, the misclassification probability only considers the most similar parameters to α_0. Similar observations have been made for the derivation of the Chernoff index, i.e., one only considers the alternative models most similar to the null. In practice it is usually easy to identify these most indistinguishable parameters so as to simplify the computation of (17). For instance, in the case of the DINA model, the most indistinguishable attribute parameter must be among those that have only one dimension misspecified.

For DCM's, the α' is generally not unique if $\mathbf{h}^*$ is chosen to be asymptotically optimal.

Consider the DINA model and a true parameter α_0. Let N_0 be the set of attributes closest to α_0. Each $\alpha_1 \in N_0$ is different from α_0 by only one attribute. Thus, N_0 is the set of parameters most difficult to distinguish from α_0. The asymptotically optimal design $\mathbf{h}^*$ must be chosen in such a way that $I(\alpha_1, \mathbf{h}^*)$ are identical for all $\alpha_1 \in N_0$. Thus, all $\alpha_1 \in N_0$ are equally difficult to distinguish from α_0 based on the item allocation $\mathbf{h}^*$. Otherwise, one can always reduce the proportion of certain items that are overrepresented and replace them with underrepresented items. Thus, the rate can be further improved. Note that the definition of N_0 may change under the reduced parameter space. See C. Tatsuoka and Ferguson (2003) for such a discussion when the parameter space is a partially ordered set.

Example 1: a simple example to illustrate α'. Consider the DINA model with three attributes. The true attribute is $\alpha_0 = (1, 1, 0)$. There are three types of experiments in the item bank, $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, and $e_3 = (0, 0, 1)$. Each type of item is used to measure one attribute. The corresponding slipping and guessing parameters are $(s_1, g_1) = (0.1, 0.1)$, $(s_2, g_2) = (0.2, 0.2)$, and $(s_3, g_3) = (0.1, 0.1)$. Consider a design $\mathbf{h} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, that is, equal numbers of items are selected for each of the three types. Then, the second attribute is the most difficult to identify because its slipping and guessing parameters (s_2, g_2) are larger. In this case, the attribute parameter that minimizes $I(\alpha, \mathbf{h})$ over $\alpha \neq \alpha_0$ and is most indistinguishable from α_0 is $\alpha' = (1, 0, 0)$, which differs from α_0 in its second attribute . The asymptotically optimal design should spend more items to measure the second attribute than the other two. We will later revisit this example and compute $\mathbf{h}^*$.

Example 2: calculations for the Bernoulli distribution. We illustrate the calculation of the rate function for models such as the DINA with Bernoulli responses. It is sufficient to compute $I(\alpha_1, \mathbf{h})$ for each true parameter α_0 and alternative $\alpha_1 \neq \alpha_0$. A separating experiment e will produce two different Bernoulli outcome distributions $f(x|\alpha_0, e) = p_0^x(1-p_0)^{1-x}$ and $f(x|\alpha_1, e) = p_1^x(1-p_1)^{1-x}$ where $p_1 \neq p_0$. Then, the log-likelihood ratio is

$$s^e_{\alpha_1} \triangleq \log f(X|e, \alpha_1) - \log f(X|e, \alpha_0) = X\Big(\log\frac{p_1}{1-p_1} - \log\frac{p_0}{1-p_0}\Big) + \log\frac{1-p_1}{1-p_0}.$$

The log-moment-generating function of $s^e_{\alpha_1}$ under $f(x|\alpha_0, e)$ is

$$\varphi_{e,\alpha_1}(\theta) = \log\left[(1-p_1)^\theta(1-p_0)^{1-\theta} + p_1^\theta p_0^{1-\theta}\right]. \tag{20}$$

For the purpose of illustration, we compute $I_e(\alpha_1)$. The solution to $\varphi'_{e,\alpha_1}(\theta^*) = 0$ is

$$\theta^* = \left[\log\frac{p^*}{1-p^*} - \log\frac{p_0}{1-p_0}\right] \Big/ \left[\log\frac{p_1}{1-p_1} - \log\frac{p_0}{1-p_0}\right],$$

where $p^* = \frac{\log(1-p_1)-\log(1-p_0)}{\log(1-p_1)-\log(1-p_0)-\log p_1+\log p_0}$. Then, the misclassification probability is approximated by

$$-\lim\frac{\log P(\hat{\alpha}_m = \alpha_1)}{m} = -\varphi_{e,\alpha_1}(\theta^*) = I_e(\alpha_1). \tag{21}$$

The parameter θ^* depends explicitly on p_1, p_0, and φ_{e,α_1}. Supposing that $p_0 < p_1$, the parameter p^* is the cutoff point. If the averaged positive responses is below p^*, then the likelihood is in favor of α_0 and vice versa.

When there are more types of items, the rate function $I(\alpha_1, \mathbf{h})$ is the sum of all the $\varphi_{e,\alpha_1}(\theta)$ in (20) weighted by their own proportions, i.e., $I(\alpha_1, \mathbf{h}) = -\inf_\theta \sum_{e\in\mathcal{E}} h_e\varphi_{e,\alpha_1}(\theta)$. Then, take the infimum over $\alpha_1 \neq \alpha_0$ for the overall rate function $I_{\mathbf{h},\alpha_0}$.

Example 1 revisited. We apply the calculations in Example 2 to the specific setting in Example 1, where $\alpha_0 = (1, 1, 0)$. We consider the alternative parameters closest to α_0: $\alpha_1 = (0, 1, 0)$, $\alpha_2 = (1, 0, 0)$, and $\alpha_3 = (1, 1, 1)$. Using the notation in (14) for each individual item and the formula in (21), we have that $I_{e_1}(\alpha_1) = 0.51$, $I_{e_2}(\alpha_2) = 0.22$, and $I_{e_3}(\alpha_3) = 0.51$.

For each $\mathbf{h} = (h_{e_1}, h_{e_2}, h_{e_3})$ and $i = 1, 2, 3$, we have that $I(\alpha_i, \mathbf{h}) = h_{e_i} I_{e_i}(\alpha_i)$. Thus, the asymptotically optimal allocation of experiments is inversely proportional to the rates $I_{e_i}(\alpha_i)$, that is, $h^*_{e_1} = h^*_{e_3} = 0.23$ and $h^*_{e_2} = 0.54$.

Comparison with the KL information. Typically, an experiment with a large KL information $D_e(\alpha_0, \alpha_1)$ also has a large rate $I_e(\alpha_1)$. This positive association is expected in that both indexes describe the information contained in the outcome of experiment e. However, these two indexes sometimes do yield different choices of items. Consider the setting in Example 2. Let experiment e_1 have parameters $s_1 = 0.1$ and $g_1 = 0.5$ and let e_2 have parameters $s_2 = 0.6$ and $g_2 = 0.01$. Both e_1 and e_2 differentiate α_0 and α_1 in such a way that the ideal response of α_0 is negative and that of α_1 is positive. We compute the rate functions and the KL informations

$$D_{e_1}(\alpha_0, \alpha_1) = 0.51, \quad D_{e_2}(\alpha_0, \alpha_1) = 0.46, \quad I_{e_1}(\alpha_1) = 0.11, \quad I_{e_2}(\alpha_1) = 0.19.$$

Thus, according to the rate function, e_2 is a better experiment, while the KL information gives the opposite answer. Thus, the KL information method does not always coincide with the rate function. This is also reflected in the simulation study. More detailed connections and comparisons are provided in Section 4. A similar difference exists between Tatsuoka and

Ferguson's criterion and the rate function. Therefore, the criterion proposed in this paper is fundamentally different from the existing ideas.

Remark 2. In practice, it is generally impossible to have two responses collected from exactly the same item. The approximation of the misclassification probability can be easily adapted to the situation when all the m items are distinct. Consider that a sequence of outcomes $X^1, ..., X^m$, has been collected from experiments $e^1, ..., e^m$. For $\alpha_1 \neq \alpha_0$, we slightly abuse the notation and let $s^i_{\alpha_1} = \log f(X^i|e^i, \alpha_1) - \log f(X^i|e^i, \alpha_0)$ be the log-likelihood ratio of the i-th outcome (generated from experiment e^i) and let $\varphi_{e^i,\alpha_1}(\theta) = \log[E_{e_i,\alpha_0}(e^{\theta s^i_{\alpha_1}})]$ be the log-MGF. Then, an analogue of Theorem 1 would be

$$P_{\mathbf{e}_m,\alpha_0}\Big(\pi(\alpha_1|\mathbf{X}_m, \mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m, \mathbf{e}_m)\Big) \approx e^{-I_{\mathbf{e}_m}(\alpha_1)},$$

where $I_{\mathbf{e}_m}(\alpha_1) = -\inf_\theta \sum_{i=1}^m \varphi_{e^i,\alpha_1}(\theta)$. Furthermore, the misclassification probability can be approximated by $-\log p(\mathbf{e}_m, \alpha_0) \approx \inf_{\alpha_1 \neq \alpha_0} I_{\mathbf{e}_m}(\alpha_1)$.

3.4. An adaptive algorithm based on the rate function

We now describe a CAT procedure corresponding to the asymptotic theorems. At each step, one first obtains an estimate of α based on the existing outcomes, $X^1, ..., X^m$. Then each successive item is selected as if the true parameter is identical to the current estimate.

Algorithm 1. Start with a prefixed set of experiments $(e^1, ..., e^{m_0})$. For every m, Outcomes $\mathbf{X}_m = (X^1, ..., X^m)$ of experiments $\mathbf{e}_m = (e^1, ..., e^m)$ have been collected. Choose e^{m+1} as follows.

1. Compute the posterior mode $\hat{\alpha}_m \triangleq \hat{\alpha}(\mathbf{X}_m, \mathbf{e}_m)$. Let α' be the attribute that has the second highest posterior probability, that is, $\alpha' = \arg\max_{\alpha \neq \hat{\alpha}_m} \pi(\alpha|\mathbf{X}_m, \mathbf{e}_m)$.
2. Let $\alpha_0 = \hat{\alpha}_m$. The next item e^{m+1} is chosen to be the one that admit the largest rate function with respect to α', that is, $e^{m+1} = \arg\sup_e I_e(\alpha')$ where $I_e(\alpha')$ is defined as in (14) (recall that $I_e(\alpha')$ depends on the true parameter value α_0 that is set to be $\hat{\alpha}_m$).

The attribute is estimated by the posterior mode based on previous responses. Then, α' is selected to have the second highest posterior probability. Thus, α' is the attribute profile that is most difficult to differentiate from $\alpha_0 = \hat{\alpha}_m$ given the currently observed responses $\mathbf{X}_m$. Thus, the experiment e^{m+1} maximizing $I_e(\alpha_1)$ is the experiment that best differentiates between $\hat{\alpha}_m$ and α'. Thus, the rationale behind Algorithm 1 is to first find the attribute profile α' most "similar" to $\hat{\alpha}_m$ and then to select the experiment that best differentiates between the two. We implement this algorithm and compare it with other existing methods in Section 6.

4. Relations to existing methods

4.1. Connection to the continuous parameter space CAT for IRT models

If the parameter space is continuous (as is the case in IRT), the efficiency measure (6) and its approximation are not applicable, in that $P(\hat{\theta}(\mathbf{X}) \neq \theta_0) = 1$, where θ is the continuous ability level. To make an analogue, we need to consider a slightly different probability

$$P(|\hat{\theta}(\mathbf{X}) - \theta_0| > \varepsilon | e_1, ..., e_m, \theta_0) \qquad \text{for some } \varepsilon > 0 \text{ small.} \tag{22}$$

This is known as an *indifference zone* in the literature of sequential hypothesis testing. One can establish similar large deviations approximations so that the above probability is approximately $e^{-mI(\varepsilon)}$. Thus, the proposed measure is closely related to the maximum Fisher information criterion and expected minimum posterior variance criterion. For IRT models, $\hat{\theta}(\mathbf{X})$ asymptotically follows a Gaussian distribution with mean centered around θ_0. Then, minimizing its variance is the same as minimizing the probability

$$P(|\hat{\theta}(\mathbf{X}) - \theta_0| > \delta m^{-1/2} | e_1, ..., e_m, \theta_0)$$

for all $\delta > 0$. One may consider that, by choosing ε very small in (22), in particular $\varepsilon \approx \delta m^{-1/2}$, maximizing the rate function is approximately equivalent to minimizing the the asymptotic variance. This connection can be made rigorous by the smooth transition from the large deviations to the moderate deviations approximations.

4.2. Connection to the KL information methods and global information

The proposed efficiency measure is closely related to the KL information. Consider a specific alternative α_1. We provide another representation of the rate function for $P(\pi(\alpha_1|\mathbf{X}_m, \mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m, \mathbf{e}_m))$. To simplify our discussion and without loss of too much generality, suppose that only one type of experiment, e, is used and that $X^1, ..., X^m$ are thus i.i.d. The calculations for multiple types of experiments are completely analogous, but more tedious. The alternative parameter α_1 admits a higher posterior probability if $\sum_{i=1}^{m} s^i > \log \pi(\alpha_0) - \log \pi(\alpha_1)$ where $s^i = \log f(X^i|\alpha_1, e) - \log f(X^i|\alpha_0, e)$ are i.i.d. random variables following distribution $g(s)$. Then, the rate function takes the form

$I_e(\alpha_1) = -\inf_\theta[\varphi(\theta)]$, where $\varphi(\theta) = \log \int g(s)e^{\theta s}$ is the log-MGF of the log likelihood ratio. Let $\theta^* = \arg\inf_\theta \varphi(\theta)$ and $g(s|\theta) = g(s)e^{\theta s - \varphi(s)}$. With some simple calculation, we obtain that

$$I_e(\alpha_1) = \int \log \frac{g(s|\theta^*)}{g(s)} g(s|\theta^*) ds,$$

which is the Kullback-Leibler information between $g(s|\theta^*)$ and $g(s)$.

Then, the rate function is the minimum KL information between the log-likelihood ratio distribution and the zero-mean distribution within its exponential family. An intuitive connection between the proposed method and the existing method based on KL information is as follows. The KL, or the posterior-weighted KL, method maximizes the KL information between the response distributions under the true and alternative model. Our method maximizes the KL information of the log-likelihood ratio instead of directly that of the outcome variable. This is because the maximum likelihood estimator (or the posterior mode estimator) maximizes the sum of the log-likelihoods.

The rate function in (17) is the minimum of $I(\alpha, \mathbf{h})$ over all $\alpha \neq \alpha_0$. This is different from the approach taken by most existing methods, which typically maximize a (possibly weighted) average of the KL information or Shannon entropy over the parameter space. Instead, this approach recalls Tatsuoka and Ferguson's asymptotically optimal experiment selection, which involves maximizing the smallest KL distance.

4.3. Connection to the Tatsuoka and Ferguson's criterion

Using the notation in Section 2.2.1, the posterior mode converges to unity exponentially fast $\frac{1}{m}\log[1-\pi(\alpha_0|\mathbf{X}_m,\mathbf{e}_m)] \to -H$ almost surely. This convergence implies that

$$P(\pi(\alpha_0|\mathbf{X}_m,\mathbf{e}_m) < \pi(\alpha'|\mathbf{X}_m,\mathbf{e}_m)) \to 0$$

as $m \to \infty$ where α' is the largest posterior mode other than α_0. The above probability is approximately the missclassification probability $p(\mathbf{e}_m, \alpha_0)$. TF's criterion and ours are very closely related in that a large H typically implies a small $p(\mathbf{e}_m, \alpha_0)$. This is because it is unlikely for $\pi(\alpha_0|\mathbf{X}_m,\mathbf{e}_m)$ to fall below some level if it converges faster to unity. However, these two criteria are technically distinct. As shown later in Section 6.1, they yield different optimal designs and the corresponding misclassification probabilities could be different.

5. Discussion

Finite sample performance. The asymptotically optimal allocation of experiments $\mathbf{h}^*$ maximizes the rate function. It can be shown that $\mathbf{h}^*$ converges to the optimal allocation that minimizes the misclassification probability as m tends to infinity. However, for finite samples, $\mathbf{h}^*$ may be different from the optimal design. In addition, $\mathbf{h}^*$ is derived under a setting in which outcomes are independently generated from experiments whose selection has been fixed (independent of the outcome). Therefore, the theorems here answer the question of what makes a good experiment and they serve as theoretical guidelines for the design of CAT procedures. Also, as the simulation study shows, the algorithms perform well.

Results associated with other estimators. The particular form of the rate function relies very much on the distribution of the log-likelihood ratios. This is mostly because we primarily focus on the misclassification probability of the maximum likelihood estimator or posterior mode. If one is interested in (asymptotically) minimizing the misclassification probability of other estimators, such as method of moment based estimators, similar exponential decay rates can be derived and they will take different forms.

Infinite discrete parameter space. We assume that the parameter space is finite. In fact, the analytic forms of $I_{\mathbf{h},\alpha_0}$ and $I(\alpha, \mathbf{h})$ can be extended to the situation when the parameter space is infinite but still discrete. The approximation results in Theorem 2 can be established with additional assumptions on the likelihood function $f(x|e, \alpha)$. With such approximation results established, Algorithm 1 can be straightforwardly applied.

Summary. To conclude the theoretical discussion, we would like to emphasize that using the misclassification probability as an efficiency criterion is a very natural approach. However, due to computational limitations, we use its approximation via large deviations theory. The resulting rate function has several appealing features. First, it only depends on the proportion of outcomes collected from each type of experiment and is free of the total number of experiment m. In addition, the rate function is usually in a closed form for stylized parametric families. For more complicated distributions, its evaluation only consists of a one dimensional minimization and is computationally efficient. In addition, as the simulation study shows, the asymptotically optimal design shows nice finite sample properties.

6. Simulation

6.1. A simple demonstration of the asymptotically optimal design

In this subsection, we consider (nonadaptive) prefixed designs selected by different criteria. We study the performance of the asymptotically optimal design $\mathbf{h}^*$ for a given α_0. We consider the DINA model with true attribute profile $\alpha_0 = (1, 1, 0)$. We also consider four types of experiments with the following attribute requirements:

$$e_1 = (0, 0, 1), \quad e_2 = (1, 0, 0), \quad e_3 = (0, 1, 0), \quad e_4 = (1, 1, 0).$$

We compare the asymptotically optimal design proposed by the current paper, denoted by LYZ, and the optimal design by Tatsuoka and Ferguson (2003), denoted by TF. We consider two sets of different slipping and guessing parameters that represent two typical situations.

Setting 1. $s_1 = s_2 = s_3 = s_4 = 0.05$ and $g_1 = g_2 = g_3 = g_4 = 0.5$. Under this setting, the asymptotically optimal proportions by LYZ are $h_1^{LYZ} = h_4^{LYZ} = 0.5$, and $h_2^{LYZ} = h_3^{LYZ} = 0$; the optimal proportions by TF are $h_1^{TF} = 0.3733$, $h_4^{TF} = 0.6267$, and $h_2^{TF} = h_3^{TF} = 0$.

Setting 2. $c_1 = c_2 = c_3 = c_4 = 0.05$, $g_1 = g_2 = g_3 = 0.5$, and $g_4 = 0.8$. Under this selection, the asymptotically optimal proportions by LY are $h_1^{LYZ} = h_2^{LYZ} = h_3^{LYZ} = \frac{1}{3}$, and $h_4^{LYZ} = 0$; the optimal proportions by TF are $h_1^{TF} = 0.2295$, $h_2^{TF} = h_3^{TF} = 0.3853$, and $h_4^{TF} = 0$.

We simulate outcomes from the above prefixed designs with different test lengths $m = 20, 50, 100$. Tables 1 and 2 show the misclassification probabilities computed via Monte Carlo. LYZ admits smaller misclassification probabilities (MCP) for all the samples sizes. The advantage of LYZ manifests even with small sample sizes. For instance, when $m = 20$ in Table 2, the misclassification probability of LYZ is 13% and that of TF is 27%.

==========

Insert Table 1 about here

==========

==========

Insert Table 2 about here

==========

6.2. The CAT algorithms

We compare Algorithm 1 with other adaptive algorithms in the literature, such as the SHE and PWKL as given in Section 2.2. We compare the behavior of these three algorithms, along with the random selection method (i.e., at each step an item is randomly selected from the item bank), in several settings.

General simulation structure. Let K be the length of the attribute profile. The true attribute α_0 is uniformly sampled from the space $\{0, 1\}^K$, i.e., each attribute has a 50% chance of being positive. Each test begins with a fixed choice of $m_0 = 2K$ items with slipping and guessing probabilities, $s = g = 0.05$. In particular, each attribute is tested

by 2 items testing solely that attribute, i.e., items with attribute requirements of the form $(0, \ldots, 0, 1, 0, \ldots, 0)$. After the prefixed choice of items, subsequent items are chosen from a bank containing items with all possible attribute requirement combinations and pre-specified slipping and guessing parameters. Items are chosen sequentially based on either Algorithm 1, SHE, PWKL, or random (uniform) selection over all possible items. The misclassification probabilities are computed based on 500,000 independent simulations that provide enough accuracy for the misclassification probabilities.

For illustration purposes, we choose the random selection method as the benchmark. For each adaptive method, we compute the ratio of the misclassification probability of that method and the misclassification probability of the random (uniform) selection method. The log of this ratio as test length increases is plotted under each setting in Figures 1, 2, 3, and 4.

A summary of the simulation results is as follows. The PWKL immediately underperforms the other two methods in all settings. The SHE and the LYZ methods perform similarly early on, but eventually the LYZ method begins to achieve significantly lower misclassification probabilities. From the plots, we can see that this pattern of behavior does not change as we vary K. However, as K grows larger, more items are needed for the asymptotic benefits of the LYZ method to become apparent. In addition, the CPU time varies for different dimension and different methods. To run 100 independent simulations, the LYZ and the KL methods take less than 10 seconds for all K. The SHE method is slightly more computationally costly and takes as much as a few minutes for 100 simulations when $K = 8$.

The specific simulation setting as given as follows.

Setting 3. The test bank contains two sets of items. Each set contains $2^K - 1$ types items containing all the possible attribute requirements. For one set, the slipping and the guessing parameters are $(s, g) = (0.10, 0.50)$; for the other set, the parameters are $(s, g) = (0.60, 0.01)$. Thus, there are $2(2^K - 1)$ types items in the bank that can be selected repeatedly. The simulation is run for $K = 3, 4, 5, 6$. The results are presented in Figure 1.

Setting 4. With a similar setup, we have two different sets of the slipping and the guessing parameters $(s, g) = (0.15, 0.15)$ and $(s, g) = (0.30, 0.01)$. The basic pattern remains. The results are presented in Figure 2.

Setting 5. We increase the variety of items available. The test bank contains items with any of four possible pairs of slipping and guessing parameters: $(s_1, g_1) = (0.01, 0.60)$, $(s_2, g_2) = (0.20, 0.01)$, $(s_3, g_3) = (0.40, 0.01)$, and $(s_4, g_4) = (0.01, 0.20)$; in addition, items corresponding to each of the $2^K - 1$ possible attribute requirements are available. Items corresponding to a particular set of attribute are limited to either (s_1, g_1) and (s_2, g_2) or (s_3, g_3) and (s_4, g_4). Thus, combining the different attribute requirements and item parameters, there are a total of $2(2^K - 1)$ types of items in the bank, each of which can be selected repeatedly. The simulation is run for $K = 3, 4, \ldots, 8$. The results are presented in Figure 3.

Setting 6. We add correlation by generating a continuous ability parameter $\theta \sim N(0, 1)$. The individual α_k are independently distributed given θ, such that

$$p(\alpha_k = 1|\theta) = \exp(\theta)/[1 + \exp(\theta)], \quad k = 1, 2, \ldots, K.$$

Setting 6 follows Setting 5 in all other respects. The results are presented in Figure 4.

==========

Insert Figure 1 about here

==========

==========

Insert Figure 2 about here

==========

==========

Insert Figure 3 about here

==========

==========

Insert Figure 4 about here

==========

7. Acknowledgement

We would like to thank the editors and the reviewers for providing valuable comments.

This research is supported in part by NSF and NIH.

References

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213-229.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*, 619-632.

Chiu, C., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633-665.

Cox, D., & Hinkley, D. (2000). *Theoretical statistics.* Chapman & Hall.

de la Torre, J., & Douglas, J. (2004). Higher order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333-353.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (p. 361-390). Hillsdale, NJ: Erlbaum Associates.

Edelsbrunner, H., & Grayson, D. R. (2000). Edgewise subdivision of a simplex. *Discrete & Computational Geometry*, *24*, 707-719.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Junker, B. (2007). Using on-line tutoring records to predict end-of-year exam scores: experience with the ASSISTments project and MCAS 8th grade mathematics. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: intellectual growth and standard settings.* Maple Grove, MN: JAM Press.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205-237.

Lord, F. M. (1971). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement*, *31*, 3-31.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale: Erlbaum.

Owen, R. J. (1975). Bayesian sequential procedure for quantal response in context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351-356.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York: Guilford Press.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics* (W. Shewhart & S. Wilks, Eds.). New York: Wiley-Interscience.

Tatsuoka, C. (1996). *Sequential classification on partially ordered sets. doctoral dissertation.* dissertation, Cornell University.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Applied Statistics (JRSS-C)*, *51*, 337-350.

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *65*, 143-157.

Tatsuoka, K. (1991). Boolean algebra applied to determination of the universal set of misconception states. *Princeton, NJ: Educational Testing Services*, *ONR-Technical Report No. RR-91-44*.

Tatsuoka, K. K. (1983). Rule space: an approch for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345-354.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, *12*, 55-73.

Tatsuoka, K. K. (2009). *Cognitive assessment: an introduction to the rule space method*. New York: Routledge.

Templin, J., He, X., Roussos, L. A., & Stout, W. F. (2003). The pseudo-item method: a simple technique for analysis of polytomous data with the fusion model. *External Diagnostic Research Group Technical Report*.

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287-305.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer et al. (Eds.), *Computerized adaptive testing: a primer* (2nd ed., p. 101-133). Mahwah, NJ: Lawrence Erlbaum Associates.

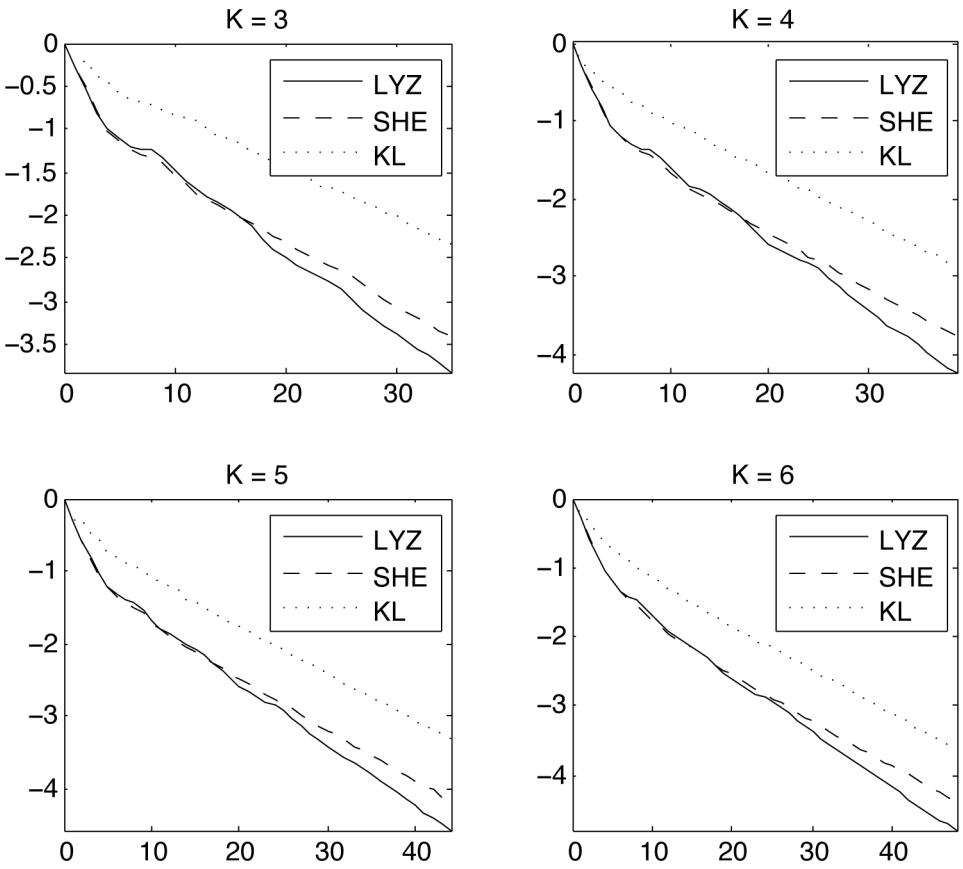
van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*, 201-216.

von Davier, M. (2005). *A general diagnosis model applied to language testing data* (Research report). Princeton, NJ: Educational Testing Service.

Xu, X., Chang, H.-H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

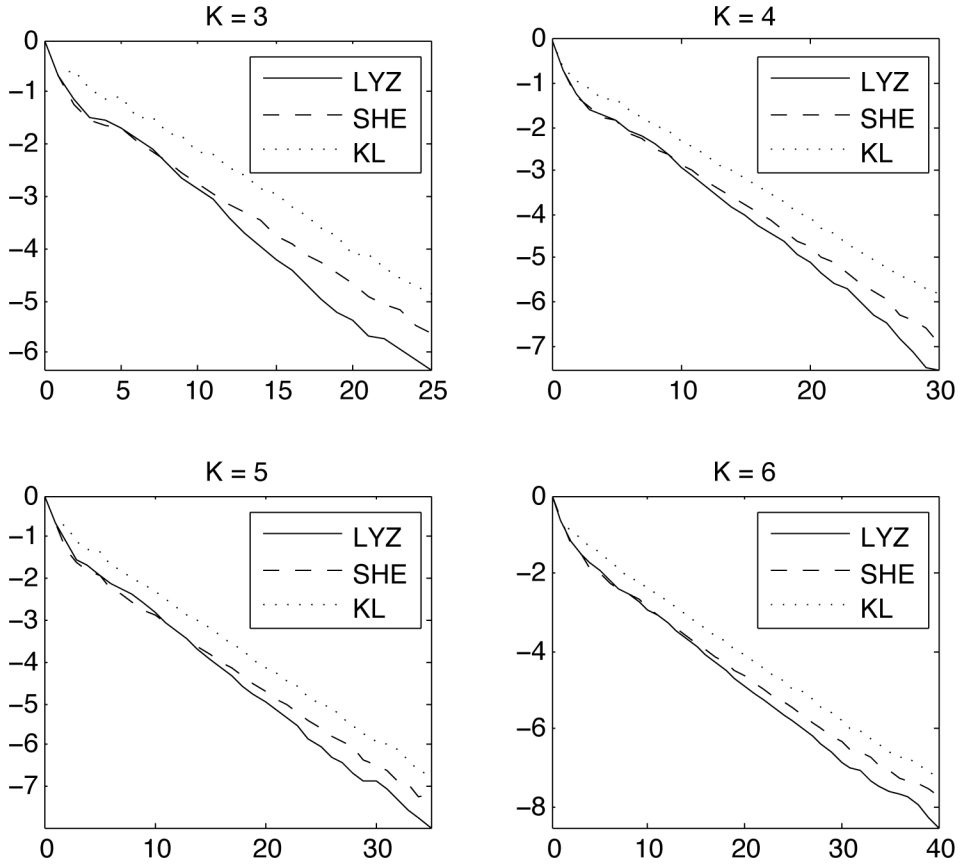
Figures

FIGURE 1.
Log-ratio of the misclassification probabilities for Setting 3



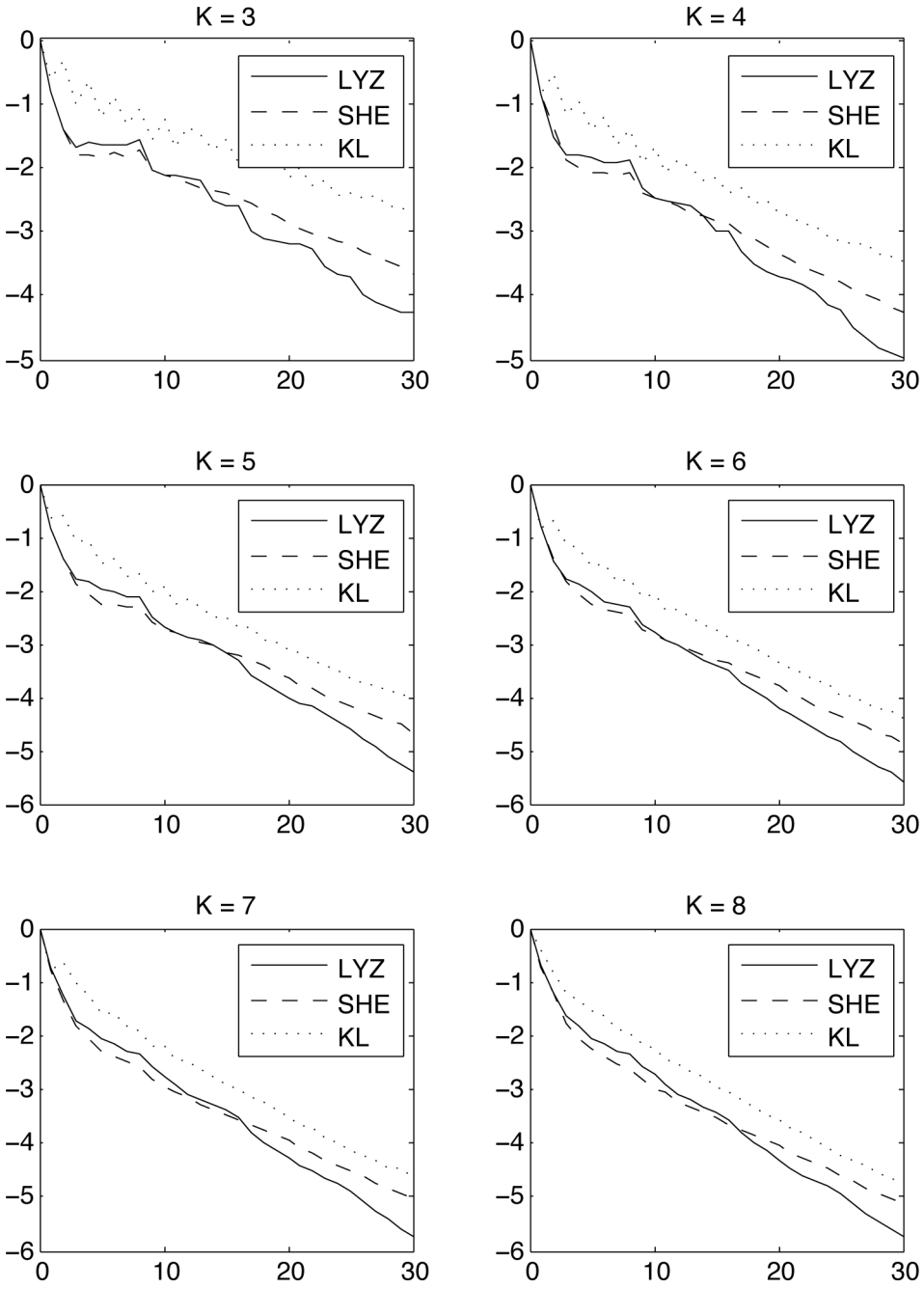
This plot shows the log-ratio of the misclassification probabilities of the given method and those of the random selection method. The x-coordinate is the test length, that is counted beginning with the first adaptive item (beyond the buffer).

FIGURE 2.
Log-ratio of the misclassification probabilities for Setting 4



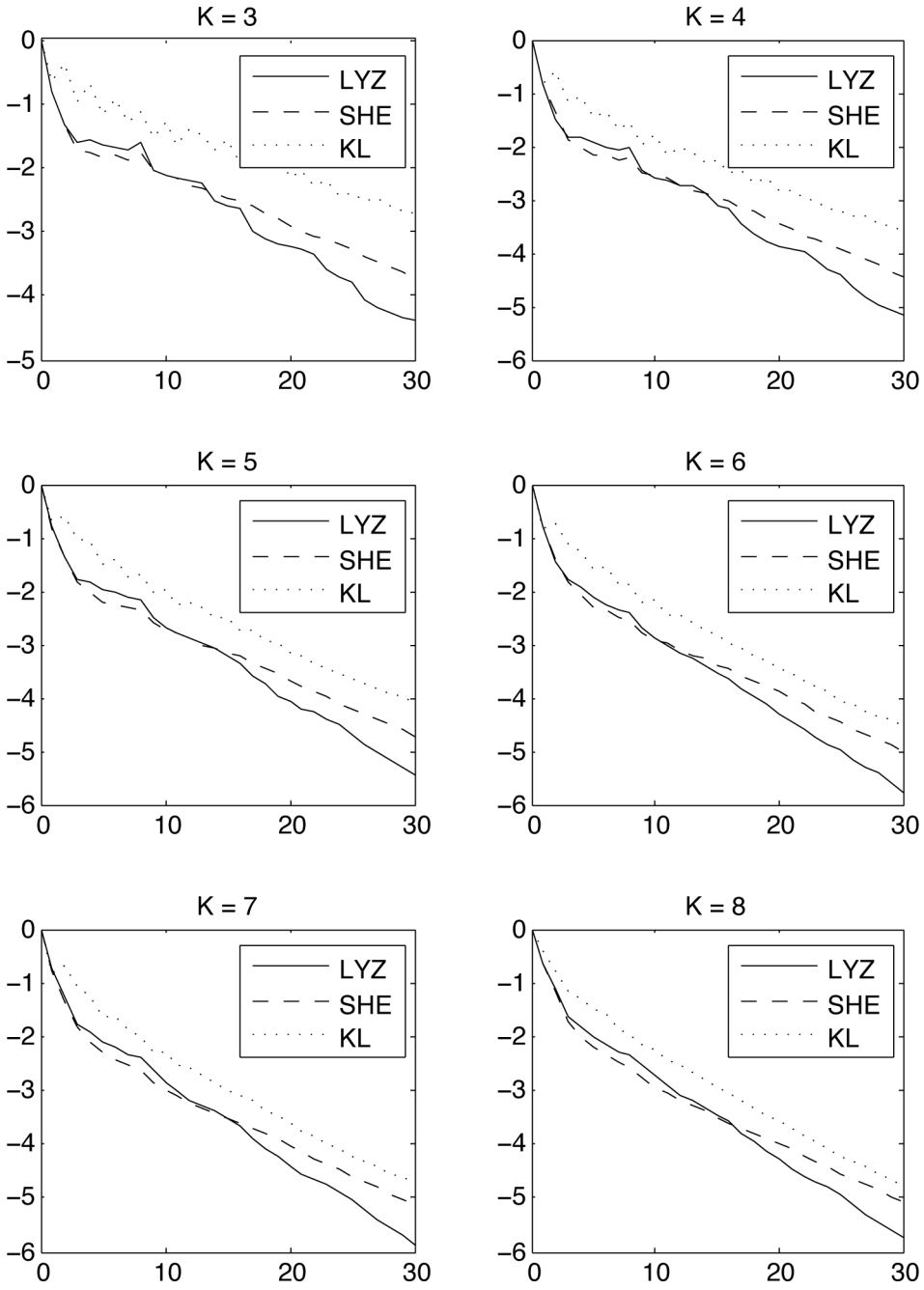
This plot shows the log-ratio of the misclassification probabilities of the given method and those of the random selection method. The x-coordinate is the test length, that is counted beginning with the first adaptive item (beyond the buffer).

FIGURE 3.
Log-ratio of the misclassification probabilities for Setting 5



This plot shows the log-ratio of the misclassification probabilities of the given method and those of the random selection method. The x-coordinate is the test length, that is counted beginning with the first adaptive item (beyond the buffer).

FIGURE 4.
Log-ratio of the misclassification probabilities for Setting 6



This plot shows the log-ratio of the misclassification probabilities of the given method and those of the random selection method. The x-coordinate is the test length, that is counted beginning with the first adaptive item (beyond the buffer).

Tables

TABLE 1.
The misclassification probabilities (MCP) under Setting 1.

m	LYZ	TF
20	6.5E-02	8.5E-02
50	3.2E-03	1.0E-02
100	4.5E-05	3.7E-04

TABLE 2.
The misclassification probabilities (MCP) under Setting 2.

m	LYZ	TF
20	1.3E-01	2.7E-01
50	2.2E-02	3.7E-02
100	1.1E-03	5.5E-03

Supplemental Material

A. Technical Proofs

Proof of Theorem 1. The proof of Theorem 1 uses standard large deviations technique and exponential change of measure. Consider a specific alternative parameter $\alpha_1 \neq \alpha_0$. We use $e \in \mathcal{E}$ to indicate different types of experiments.

Suppose that m_e independent outcomes have been generated from experiment e. Note that $m_e/m \to h_e$. The log-likelihood ratios are as defined in (10), and follow joint distribution

$$\prod_{e\in\mathcal{E}} \prod_{l=1}^{m_e} g_e(s_{\alpha_1}^{e,l}|\alpha_1).$$

Let $A = \log \pi(\alpha_0) - \log \pi(\alpha_1)$. We choose θ_m such that $\sum_{e\in\mathcal{E}} m_e \varphi_e'(\theta_m) = A$. Let θ_e^* be chosen as in the statement of the theorem. According to Remark 1, we have that $\theta_1^* = ... = \theta_\kappa^*$ and further that $\theta_m - \theta_1^* \to 0$ as $m \to \infty$. We further consider the exponential change of measure, Q, under which the log-likelihood ratios follow joint density

$$\prod_{e\in\mathcal{E}} \prod_{l=1}^{m_e} g_e(s_{\alpha_1}^{e,l}|\theta_m, \alpha_1), \tag{A1}$$

where $g_e(s|\theta, \alpha_1)$ is the exponential family defined in (11).

Note that under Q or equivalently under the joint density (A1), the $s_{\alpha_1}^{e,l}$'s are jointly independent. For a given experiment e, the $s_{\alpha_1}^{e,l}$'s are i.i.d. Following the standard results for natural exponential families, for each e, $E^Q s_{\alpha_1}^{e,l} = \varphi_e'(\theta_m)$, where E^Q denotes the expectation with respect to the density (A1). The total sum has expectation

$$E^Q \sum_{e,l} s_{\alpha_1}^{e,l} = \sum_e m_e \varphi_e'(\theta_m) = A.$$

To simplify the notation, we use $\sum_{e,l}$ and $\prod_{e,l}$ to denote the sum and the product over all the outcomes. We write

$$\begin{aligned} P\Big(\pi(\alpha_1|\mathbf{X}_m, \mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m, \mathbf{e}_m)\Big) &= P\Big(\sum_{e,l} s_{\alpha_1}^{e,l} > A \Big) \\ &= E^Q\Big(\prod_{e,l} \frac{g_e(s_{\alpha_1}^{e,l}|\alpha_1)}{g_e(s_{\alpha_1}^{e,l}|\theta_m, \alpha_1)} \; ; \; \sum_{e,l} s_{\alpha_1}^{e,l} > A \Big). \end{aligned}$$

We plug in the forms of $g_e(s|\alpha_1)$ and $g_e(s|\theta,\alpha_1)$ and continue the calculation:

$$P\Big(\pi(\alpha_1|\mathbf{X}_m,\mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m,\mathbf{e}_m)\Big) = E^Q\Big(\ \prod_{e,l} e^{\varphi_e(\theta_m)-\theta_m s^{e,l}_{\alpha_1}}\ ;\ \sum_{e,l} s^{e,l}_{\alpha_1} > A\ \Big)$$
$$= e^{-\sum_e m_e L_e(\theta_m|\alpha_1)} E^Q\Big(\prod_{e,l} e^{-\theta_m(s^{e,l}_{\alpha_1}-\varphi'_e(\theta_m))}; \sum_{e,l} s^{e,l}_{\alpha_1} > A\Big),$$

where L_e is defined as in (12). Note that $\sum_e m_e\varphi'_e(\theta_m) = A$. We continue the above calculation and obatin that

$$P\Big(\pi(\alpha_1|\mathbf{X}_m,\mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m,\mathbf{e}_m)\Big) = e^{-\sum_e m_e L_e(\theta_m|\alpha_1)} E^Q\Big(e^{-\theta_m\sum_{e,l} s^{e,l}_{\alpha_1}+\theta_m A}; \sum_{e,l} s^{e,l}_{\alpha_1} > A\Big)$$
$$\le e^{-\sum_e m_e L_e(\theta_m|\alpha_1)}$$
$$= e^{-(1+o(1))m\sum_e h_e L_e(\theta_m|\alpha_1)}. \tag{A2}$$

Thus, we have shown an upper bound.

For the lower bound, by the central limit theorem, there exists $\varepsilon,\delta>0$ such that for m large enough, we may write the expectation term in the above display as

$$E^Q\Big(e^{-\theta_m\sum_{e,l} s^{e,l}_{\alpha_1}+\theta_m A}; \sum_{e,l} s^{e,l}_{\alpha_1} > A\Big) \ge E^Q\Big(e^{-\theta_m\sum_{e,l} s^{e,l}_{\alpha_1}+\theta_m A}; A+\sqrt{m}\delta > \sum_{e,l} s^{e,l}_{\alpha_1} > A\Big)$$
$$\ge E^Q\Big(e^{-\theta_m\delta\sqrt{m}}; A+\sqrt{m}\delta > \sum_{e,l} s^{e,l}_{\alpha_1} > A\Big)$$
$$\ge \varepsilon e^{-\theta_m\delta\sqrt{m}}.$$

Thus, we obtain a lower bound that

$$P\Big(\pi(\alpha_1|\mathbf{X}_m,\mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m,\mathbf{e}_m)\Big) \ge e^{-(1+o(1))m\sum_e h_e L_e(\theta_m|\alpha_1)}. \tag{A3}$$

Combining (A2), (A3), and the fact that $\theta_m\to\theta_1^*$, we conclude the proof of Theorem 1 using the definition of $I(\mathbf{h},\alpha_1)$ in (13).

Proof of Theorem 2. Based on the proof of Theorem 1, the proof of Theorem 2 is simply an application of the Bernoulli's inequality. Thus, we only lay out the key steps. First,

$$p(\mathbf{e}_m,\alpha_0) = P\Big[\ \cup_{\alpha_1\neq\alpha_0}\ \{\pi(\alpha_1|\mathbf{X}_m,\mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m,\mathbf{e}_m)\}\Big].$$

Let α' be an alternative parameter admitting the smallest rate, that is, $I(\alpha', \mathbf{h}) = I_{\mathbf{h},\alpha_0}$. Thus, we have that

$$\begin{aligned} P\Big[\pi(\alpha'|\mathbf{X}_m, \mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m, \mathbf{e}_m)\Big] &\leq p(\mathbf{e}_m, \alpha_0) \\ &\leq \sum_{\alpha_1 \neq \alpha 0} P\Big[\pi(\alpha_1|\mathbf{X}_m, \mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m, \mathbf{e}_m)\Big] \\ &\leq \kappa P\Big[\pi(\alpha'|\mathbf{X}_m, \mathbf{e}_m) > \pi(\alpha_0|\mathbf{X}_m, \mathbf{e}_m)\Big]. \end{aligned}$$

We take the log on both sides and obtain that

$$-(1+o(1))I(\alpha', \mathbf{h}) \quad \leq \quad \frac{\log p(\mathbf{e}_m, \alpha_0)}{m} \quad \leq \quad -(1+o(1))I(\alpha', \mathbf{h}) + \frac{\log \kappa}{m}.$$

B. Identifiability issues

Throughout this paper, we assume that all the parameters are separable from each other by the set of experiments. In the case that there are two or more parameters that are not separable, we need to reduce the parameter spaces as follows. We write $\alpha_1 \sim \alpha_2$ if $D_e(\alpha_1, \alpha_2) = 0$ for all $e \in \mathcal{E}$. It is not difficult to verify that the binary relationship "$\sim$" is an equivalence relation. Let $[\alpha] = \{\alpha_1 \in \mathcal{E} : \alpha_1 \sim \alpha\}$ be the set of parameters related to α by $\sim$. Then, the reduced parameter set is defined as the quotient set

$$\tilde{\mathcal{E}} = \mathcal{E}/\sim = \{[\alpha] : \alpha \in \mathcal{E}\}.$$

To further explain, if $\alpha_1 \sim \alpha_2$, then the response distributions are identical $f(x|e, \alpha_1) = f(x|e, \alpha_2)$ for all e. We are not able to distinguish α_1 and α_2. If $[\alpha_1] \neq [\alpha_2]$, then there exists at least one e such that $f(x|e, \alpha_1)$ and $f(x|e, \alpha_2)$ are distinct distributions. Therefore, all equivalence classes in the new parameter space $\tilde{\mathcal{E}}$ are identifiable.

C. Computation of the asymptotically optimal design

For some true parameter value $\alpha_0 \in \mathcal{A}$, we wish to optimize

$$\sup_{\mathbf{h}} I_{\mathbf{h},\alpha_0} = \sup_{\mathbf{h}} \inf_{\alpha \in \mathcal{A}} I(\alpha, \mathbf{h})$$

over all nonnegative $\mathbf{h}$ such that $\sum_j h_j = 1$. Combine Equations (16), (17), and (18) to rewrite the problem as that of finding

$$\mathbf{h}^* = \arg \sup_{\mathbf{h}: \sum_j h_j = 1} \inf_{\alpha \in \mathcal{A}} \sup_{\theta} \sum_j h_j(-\varphi_{j,\alpha}(\theta)) \tag{A4}$$

Consider the innermost quantity as a function of $\mathbf{h}$ and θ. For any particular α, $f_\alpha(\mathbf{h},\theta) = \sum_j h_j(-\varphi_{j,\alpha}(\theta))$ is linear in $\mathbf{h}$, and so $I(\alpha,\mathbf{h}) = \sup_\theta f_\alpha(\mathbf{h},\theta)$ is convex in $\mathbf{h}$. Additionally, the set $\{\mathbf{h} \in \mathbb{R}_+^d : \sum_{j=1}^d h_j = 1\}$ forms a $(d-1)$-simplex with its d vertices at the standard basis vectors; a $(d-1)$-simplex is simply a $(d-1)$-dimensional polytope formed from the convex hull of its d vertices. By convexity, for each α, $I(\alpha,\mathbf{h})$ must attain its maximal value at one of these vertices. Let $s_{\mathbf{V}}$ be a generic notation for a d-dimensional simplex with vertices at $\mathbf{v} = \{v_1,\ldots,v_d\}$. Based on the above discussion, we can find supper and lower bounds for $\sup_{\mathbf{h}\in s_{\mathbf{V}}} \inf_{\alpha\in A} I(\alpha,\mathbf{h})$. In particular, we have that

$$\sup_{\mathbf{h}\in s_{\mathbf{V}}} \inf_{\alpha\in\mathcal{A}} I(\alpha,\mathbf{h}) \le \sup_{\mathbf{h}\in s_{\mathbf{V}}} \inf_{\alpha\in\mathcal{A}} \sup_{\mathbf{h}\in\mathbf{V}} I(\alpha,\mathbf{h}) = \inf_{\alpha\in\mathcal{A}} \sup_{\mathbf{h}\in\mathbf{V}} I(\alpha,\mathbf{h}) \triangleq UB(s_{\mathbf{V}})$$

and that

$$\sup_{\mathbf{h}\in s_{\mathbf{V}}} \inf_{\alpha\in\mathcal{A}} I(\alpha,\mathbf{h}) \ge \sup_{\mathbf{h}\in\mathbf{V}} \inf_{\alpha\in\mathcal{A}} I(\alpha,\mathbf{h}) \triangleq LB(s_{\mathbf{V}}).$$

Furthermore, as $I(\alpha,\mathbf{h})$ is a continuous function of $\mathbf{h}$, the two bounds converge to each other as the size of the simplex s_v converges to zero. With these constructions, now consider the following algorithm for finding $\mathbf{h}^*$ and $I_{\mathbf{h}^*,\alpha_0}$. In the algorithm, we use L to denote a set each element of which is a simplex and "$\leftarrow$" to denote value assignment.

Algorithm 2. Set $\varepsilon > 0$ indicating the accuracy level of the algorithm. Let

$$\mathbf{v}_0 = \{(1,0,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,\ldots,0,1)\}$$

and $L = \{s_{\mathbf{V}_0}\}$, i.e., $s_{\mathbf{V}_0} = \{\mathbf{h} \in \mathbb{R}_d^d : \sum h_j = 1\}$. Set $LB \leftarrow LB(s_{\mathbf{V}_0})$ and $UB \leftarrow UB(s_{\mathbf{V}_0})$.

Perform the following steps.

1. Let $s_{\mathbf{V}^*} \in L$ be the simplex with the largest $UB(s_{\mathbf{V}^*})$, i.e.,

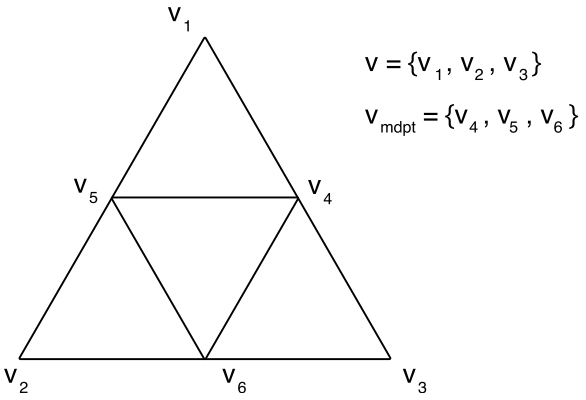
$$s_{\mathbf{V}^*} = \arg\sup_{s_{\mathbf{V}}\in L} UB(s_{\mathbf{V}}).$$

   Divided $s_{\mathbf{V}^*}$ into $2^{\kappa-1}$ smaller simplexes, with their vertices at either the original vertices $\mathbf{v}^*$ or their midpoints $\mathbf{v}^*_{mdpt}$ ((Edelsbrunner & Grayson, 2000)). Denote these $2^{\kappa-1}$ sub-simplexes by $s_{\mathbf{V}_1},...,s_{\mathbf{V}_{2^{\kappa-1}}}$. A simple example for the $\kappa = 3$ case is illustrated in Figure 5.
2. Remove $s_{\mathbf{V}^*}$ from L and add $s_{\mathbf{V}_1},...,s_{\mathbf{V}_{2^{\kappa-1}}}$ to L, i.e.,

$$L \leftarrow (L\backslash\{s_{\mathbf{V}^*}\}) \cup \{s_{\mathbf{V}_1},...,s_{\mathbf{V}_{2^{\kappa-1}}}\}.$$

3. Let $LB \leftarrow \max\left\{LB, \sup_{\mathbf{h}\in\mathbf{V}^*_{mdpt}} \inf_{\alpha\in\mathcal{A}} I(\alpha,\mathbf{h})\right\}$.

FIGURE 5.
A 2-simplex



This figure depicts the 2-simplex s_v with vertices v and their midpoints v_{mdpt}. This simplex has 4 subdivisions associated with the following sets of vertices: $\{v_1, v_4, v_5\}$, $\{v_2, v_5, v_6\}$, $\{v_3, v_4, v_6\}$, and $\{v_4, v_5, v_6\}$.

4. For each $s_{\mathbf{V}} \in L$, if $UB(s_{\mathbf{V}}) < LB$ then remove $s_{\mathbf{V}}$ from L, that is, $L \leftarrow L \backslash \{s_{\mathbf{V}}\}$.
5. Set $UB \leftarrow \sup_{s_{\mathbf{V}} \in L} UB(s_{\mathbf{V}})$.

Repeat the above steps until $UB - LB < \varepsilon$ and output

$$\mathbf{h}^* = \arg \sup_{\mathbf{h} \in \mathbf{V}, s_{\mathbf{V}} \in L} \inf_{\alpha \in \mathcal{A}} I(\alpha, \mathbf{h}).$$

This algorithm will efficiently solve the problem of finding the optimal $\mathbf{h}$, with easily controllable error in both the objective function and $\mathbf{h}$. This algorithm can in fact be used to find the maximum over the simplex of the minimum of any assortment of convex functions. In particular, this can be used to solve Tatsuoka and Ferguson's algorithm, since the KL distance is linear (and hence convex) in $\mathbf{h}$.