Submitted to *Operations Research*
manuscript OPRE-2016-01-006

Tractable Sampling Strategies for Ordinal Optimization

Dongwook Shin
Hong Kong University of Science and Technology, School of Business and Management, Clear Water Bay, Kowloon, Hong Kong
{dwshin@ust.hk}

Mark Broadie, Assaf Zeevi
Graduate School of Business, Columbia University, New York, NY 10025
{mnb2@columbia.edu, assaf@gsb.columbia.edu}

We consider a problem of ordinal optimization where the objective is to select the best of several competing alternatives ("systems"), when the probability distributions governing each system's performance are not known, but can be learned via sampling. The objective is to dynamically allocate samples within a finite sampling budget to minimize the probability of selecting a system that is not the best. This objective does not possess an analytically tractable solution. We introduce a family of practically implementable sampling policies and show that the performance exhibits (asymptotically) near-optimal performance. Further, we show via numerical testing that the proposed policies perform well compared to other benchmark policies.

Key words: ranking and selection; optimal learning; dynamic sampling

1. Introduction

Given a finite number of populations, henceforth referred to as *systems*, we are concerned with the problem of dynamically learning the statistical characteristics of the systems to ultimately select the one with the highest mean. This is an instance of ordinal optimization where the systems cannot be evaluated analytically but it is possible to sequentially sample from each system subject to a given sampling budget. A commonly used performance measure for sampling policies is the probability of false selection, i.e., the likelihood of a policy mistakenly selecting a suboptimal

system. Unfortunately, as is well documented in the literature, this objective is not analytically tractable.

Glynn and Juneja (2004) focus on the large deviations rate function of the probability of false selection, hereafter referred to simply as the rate function. An oracle that knows the underlying probability distributions can determine the allocation that maximizes this rate function, and hence, asymptotically minimize the probability of false selection. Glynn and Juneja (2004) were primarily concerned with characterizing the static oracle allocation rule. Of course to implement this, since the probability distributions are not known, one would need to estimate the rate function, which requires estimation of the cumulant generating function associated with each system. This introduces significant implementation challenges as noted by Glynn and Juneja (2015).

This situation is drastically simplified when one *assumes* that the underlying distributions follow a particular parametric form and designs sampling policies based on that premise. Notably, Chen et al. (2000) suggest a sampling policy, known as the Optimal Computing Budget Allocation (OCBA), based on the premise of underlying normal distributions. Although this approach provides for an attractive and practically implementable allocation policy, any notion of optimality associated with such a policy is with respect to the particular distributional assumption that may not hold for the true underlying distributions.

The main contribution of this paper is to address some of these deficiencies by focusing on sampling procedures that are practically implementable, that are not restricted by parametric assumptions, and that simultaneously exhibit (asymptotic) performance guarantees. In more detail, our contributions are summarized as follows:

(i) We show that, if the difference in means between the best and the second-best systems, hereafter denoted by δ, is sufficiently small, the probability of false selection can be approximated by the rate function corresponding to a Gaussian distribution, which is structured around the first two moments of the underlying probability distributions (Theorem 2);

(ii) Based on the two-moment approximation, we propose a dynamic sampling policy, referred to as the Welch Divergence (WD) policy, and analyze its asymptotic performance as the sampling budget grows large (Theorem 1);

(iii) Building on the structural properties of the WD policy, we provide an adaptive variant of WD that performs more efficiently and exhibits attractive numerical performance when the number of systems is large.

The first contribution can be viewed from two perspectives. From a theoretical perspective, we characterize the class of problem instances where the misspecification due to the Gaussian assumption is not a primary concern. From a practical perspective, we address the implementation challenges regarding the rate function; it is approximated by estimating the first two moments of the underlying probability distributions, alleviating the need to estimate cumulant generating functions.

Figure 1 illustrates key qualitative findings of this paper. For a given sampling budget (T), sampling policies based on the first two moments (e.g., WD and OCBA) exhibit good performance if the difference in means between the best and the second-best systems (δ) is sufficiently small relative to $T^{-1/2}$. Thus, the regime favorable to the two-moment approximation can be roughly characterized with the aid of the curve along which $T \propto 1/\delta^2$; see §5.2. (This curve is not a sharp boundary between the two regions in the figure, rather, it provides a rough guideline for selecting algorithms in practice.)

Sampling policies that are structured around the first two moments are prevalent; see, e.g., Chen et al. (2000), Frazier et al. (2008), and the literature review in §2. It is one of the core distinctions of this paper that our proposed policies are closely related to the Welch's t-test statistic (Welch 1947), widely used for testing a hypothesis that two populations have equal means. Both of our analytical and numerical results show that this metric is indeed an appropriate measure of divergence between two probability distributions in an ordinal optimization setting.

Similar to majority of related literature, our analysis takes place in a setting where samples are taken independently within a system and across systems, and methodologies like common random numbers are outside the scope of the analysis. However, we allow sampling policies to take multiple samples in each stage, so our algorithms are applicable in parallel computing environments.

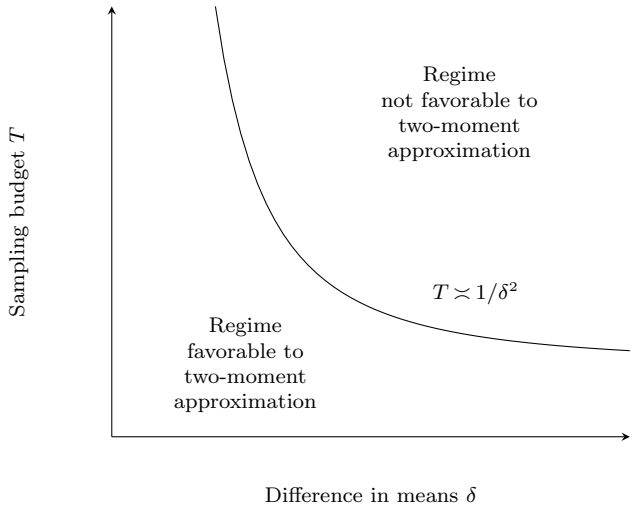

Figure 1 **Parameter region supporting the two-moment approximation.** For a fixed value of the difference in means (δ) between the best and the second-best systems, the efficacy of the two-moment approximation for the probability of false selection is determined by the relation of the gap and the sampling budget (T). As δ gets smaller, policies based on the two moments will perform well for a wider range of sampling budgets, but as this gap increases, the performance deteriorates and the approximation is less accurate.

The remainder of the paper is organized as follows. In §2 we survey related literature. In §3 we introduce a tractable objective function based on the theory of large deviations and formulate a dynamic optimization problem related to that objective. In §4 we propose the WD policy. In §5 we provide theoretical analyses of the proposed policy. In §6 we propose an adaptive version of the WD policy. In §7 we test the proposed policies numerically and compare with several benchmark policies. Appendices A and B contain the proofs for main theoretical results and auxiliary results, respectively. Appendix C discusses the effect of initial samples on the performance of the proposed policies.

2. Literature Review

The existing literature on (stochastic) ordinal optimization can be roughly categorized into *fixed budget* and *fixed confidence* settings; the goal in the former setting is to minimize the probability of false selection given a sampling budget, while in the latter setting the goal is to devise a sampling

procedure that satisfies a desired guarantee on the probability of false selection by taking as few samples as possible.

2.1. Fixed Budget Setting

In the case where the underlying probability distributions are assumed to be Gaussian with known variance, a series of papers beginning with Chen (1995) and continuing with Chen et al. (1996, 1997, 2000, 2003) suggest a family of policies known as Optimal Computing Budget Allocation (OCBA). They characterize the allocation that maximizes a lower bound on the probability of correct selection and suggest a dynamic policy that sequentially estimates the lower bound from past sample observations, and then makes sampling decisions accordingly.

The OCBA policy shares the same structure with our procedure in the sense that both myopically maximize certain objective functions. On the other hand, our work differs significantly from the stream of OCBA literature surveyed above: while the application of OCBA in non-Gaussian environments can be viewed as a heuristic, our work provides rigorous justification for the efficacy of our proposed policy in the non-Gaussian case.

Further, Chen et al. (2000) provides a simple budget allocation rule for OCBA that circumvents the computations involved in extracting the exact solution to their objective function. Pasupathy et al. (2015) show that such a workaround is valid asymptotically as the number of systems tends to infinity. The WD policy proposed in this paper requires the exact analysis of a convex optimization problem, and hence computational demands increase when the number of systems is large. To address that, we also provide a variant of WD whose computational burden is comparable to that of OCBA.

In the case with Gaussian distributions with unknown variances, Chick and Inoue (2001) derive a bound for the expected loss associated with potential incorrect selections and provide sampling policies to minimize that bound asymptotically.

Another example of a Gaussian-based procedure is the Knowledge Gradient (KG) policy proposed by Frazier et al. (2008). Like the OCBA policy, it assumes that each distribution is Gaussian

with known variance. However, while the probability of correct selection in the OCBA policy is classified as 0-1 loss, the KG policy aims to minimize expected linear loss: the difference in means between the best and selected system. The 0-1 loss is more appropriate in situations where identification of the best system is critical, while the linear loss is more appropriate when the ultimate value corresponds to the selected system's mean. An unknown-variance version of the KG policy is developed in Chick et al. (2007) under the name LL_1.

In cases with general distributions, the main difficulty is loss of practicality. Glynn and Juneja (2004) use large deviations theory with a certain family of light-tailed distributions to identify the rate function associated with the probability of false selection. The (asymptotically) optimal full information allocation is obtained by maximizing the rate function. Broadie et al. (2007) and Blanchet et al. (2008) study a heavy-tailed analog of this problem. These studies focus almost exclusively on structural insights for the rate function and static full information policies. This paper is significantly different in that we deal with dynamic sampling policies that need to learn and maximize the rate function simultaneously, which introduces challenges of dynamic sampling along with important implementation issues, some of which are discussed in Glynn and Juneja (2015).

Recently, Russo (2016) considers the problem where distributions are restricted to be members of an exponential family with a single unknown parameter. Contrary to the frequentiest setting in our paper, he formulates a posterior distribution for the unknown parameters according to Bayes rule and characterizes the rate of convergence for the posterior probability of false selection. He suggests sampling policies that can achieve a near-optimal rate asymptotically. However, from an implementation viewpoint, these policies require multi-dimensional integrals repeatedly, which are difficult to compute in general.

Our research is closely related to pure exploration in the multi-armed bandit (MAB) problem, often referred to as best-arm identification; see Bubeck and Cesa-Bianchi (2012) for a comprehensive overview. (Additional best-arm identification papers in the fixed confidence setting are

reviewed in §2.2.) The best-arm identification procedures seek the same goal as ordinal optimization procedures—i.e., selecting the unique best arm, or the system with highest mean in the language of this paper. Bubeck et al. (2009) derive bounds on the probability of false selection for two algorithms: a uniform allocation rule with selecting the arm with the highest empirical mean and the upper confidence bound (UCB) allocation rule in Auer et al. (2002) with selecting the most played arm. The rate of decrease of the probability of false selection is exponential for the former algorithm, but is only polynomial for the latter one, implying regret-minimizing allocation rules, such as UCB, are not well-suited for the pure exploration problem.

Further, Audibert and Bubeck (2010) suggest two algorithms: a variant of the UCB algorithm, in which the optimal value of its parameter depends on some measure of the complexity of the problem, and Successive Rejects algorithm, a parameter-free method based on progressively rejecting seemingly "bad" arms. Bubeck et al. (2013) generalize the preceding work to the problem of identifying multiple arms. They derive upper and lower bounds on the probability of false selection for a finite sampling budget. However, the asymptotic performance of these algorithms may not be optimized in terms of the rate function. In contrast, we propose and analyze a sampling policy that maximizes the rate function asymptotically.

2.2. Fixed Confidence Setting

In this setting the goal is to select the best system with a predetermined probability, by taking as few samples as possible. One of the early contributions in this setting traces back to the work of Bechhofer (1954) who established the indifference-zone (IZ) formulation. A large body of research on the IZ procedure followed (see, e.g., Paulson 1964, Rinott 1978, Nelson et al. 2001, Kim and Nelson 2001, Goldsman et al. 2002, Hong 2006). While most of procedures in the preceding papers are based on the premise of underlying Gaussian distributions, a recent procedure developed by Fan et al. (2016) allows for general, non-Gaussian distributions.

The stream of IZ literature surveyed above is closely related to best-arm identification problem in the fixed confidence setting. They differ along two dimensions: the best-arm procedures employ

different elimination mechanisms based on a confidence bound for the mean of each arm, while IZ procedures eliminate "arms" based on a confidence bound for the difference in means between two arms; and in the best-arm problem a standard assumption is that the underlying distribution for each arm (system) is either Bernoulli (Even-Dar et al. 2002, Mannor and Tsitsiklis 2004, Even-Dar et al. 2006, Kalyanakrishnan et al. 2012, Jamieson et al. 2014) or unbounded but explicit bounds on moments are known (Glynn and Juneja 2015).

When the underlying distributions have unbounded support, Glynn and Juneja (2015) show that one cannot use an empirical rate function to determine the number of samples that guarantees a desired probability of false selection. However, this issue is not critical in the fixed-budget setting of this paper, in which the number of samples is exogenously given; see §4.2-§4.3 for a further discussion.

3. Problem Formulation

3.1. Policy Preliminaries

Consider k stochastic systems, whose performance is governed by a distribution $F_j(\cdot)$, $j = 1, \ldots, k$. These distributions are unknown to the decision maker. We assume that the second moment of each distribution is finite, and let $\mu_j = \int x dF_j(x)$ and $\sigma_j = (\int x^2 dF_j(x) - \mu_j^2)^{1/2}$ be the mean and standard deviation for performance of the jth system. Denote $\boldsymbol{\mu}$ (respectively, $\boldsymbol{\sigma}$) the k-dimensional vector of means (respectively, standard deviations). We assume that $\mu_1 > \max_{j \neq 1} \mu_j$ and that each σ_j is strictly positive to avoid trivial cases.

A decision maker is given a *fixed* sampling budget T, which means T independent samples can be drawn from the k systems. A sampling policy $\boldsymbol{\pi}$ is defined as a sequence of random variables, $\pi_1, \pi_2, \ldots$, taking values in the set $\{1, 2, \ldots, k\}$; the event $\{\pi_t = j\}$ means a sample from system j is taken at stage t. Define X_{jt}, $t = 1, \ldots, T$, as a random sample from system j in stage t and let $\mathscr{F}_t$ be the σ-field generated by the samples and sampling decisions taken up to stage t, i.e., $\{(\pi_\tau, X_{\pi_\tau,\tau})\}_{\tau=1}^t$, with the convention that $\mathscr{F}_0$ is the nominal sigma algebra associated with underlying probability space. The set of non-anticipating policies is denoted as Π, in which the

sampling decision in stage t is determined based on all the sampling decisions and samples observed in previous stages, i.e., $\{\pi_t = j\} \in \mathscr{F}_{t-1}$ for $j = 1, \ldots, k$ and $t = 1, \ldots, T$.

For each system j, we denote by $N_{jt}(\boldsymbol{\pi})$ the cumulative number of samples up to stage t and let $\alpha_{jt}(\boldsymbol{\pi})$ be the sampling rate at stage t. Formally,

$$N_{jt}(\boldsymbol{\pi}) = \sum_{\tau=1}^{t} \mathbf{I}\{\pi_\tau = j\} \tag{1}$$

$$\alpha_{jt}(\boldsymbol{\pi}) = \frac{N_{jt}(\boldsymbol{\pi})}{t}, \tag{2}$$

where $\mathbf{I}\{A\} = 1$ if A is true, and 0 otherwise. Also, we denote by $\bar{X}_{jt}(\boldsymbol{\pi})$ and $S^2_{jt}(\boldsymbol{\pi})$ the sample mean and variance of system j in stage t:

$$\bar{X}_{jt}(\boldsymbol{\pi}) = \frac{\sum_{\tau=1}^{t} X_{j\tau} \mathbf{I}\{\pi_\tau = j\}}{N_{jt}(\boldsymbol{\pi})} \tag{3}$$

$$S^2_{jt}(\boldsymbol{\pi}) = \frac{\sum_{\tau=1}^{t} (X_{j\tau} - \bar{X}_{jt}(\boldsymbol{\pi}))^2 \mathbf{I}\{\pi_\tau = j\}}{N_{jt}(\boldsymbol{\pi})}. \tag{4}$$

Note that $X_{j\tau}$ is observed only when $\{\pi_\tau = j\}$. We use bold type for vectors: $\boldsymbol{\alpha}_t(\boldsymbol{\pi}) = (\alpha_{1t}(\boldsymbol{\pi}), \ldots, \alpha_{kt}(\boldsymbol{\pi}))$ and $\boldsymbol{N}_t(\boldsymbol{\pi}) = (N_{1t}(\boldsymbol{\pi}), \ldots, N_{kt}(\boldsymbol{\pi}))$ denote vectors of sampling rates and cumulative numbers of samples in stage t, respectively. Likewise, we let $\bar{\boldsymbol{X}}_t(\boldsymbol{\pi})$ and $\boldsymbol{S}^2_t(\boldsymbol{\pi})$ be the vectors of sample means and variances in stage t, respectively. For brevity, the argument $\boldsymbol{\pi}$ may be dropped when it is clear from the context. To ensure that $\bar{\boldsymbol{X}}_t$ and $\boldsymbol{S}^2_{jt}$ are well defined, each system is sampled once initially.

In the optimization problem we consider in §3.3, we further restrict attention to *consistent* policies defined as follows.

Definition 1 (Consistency). A policy $\boldsymbol{\pi} \in \Pi$ is consistent if $N_{jt}(\boldsymbol{\pi}) \to \infty$ almost surely for each j as $t \to \infty$.

Under a consistent policy, no system has its sampling stopped prematurely. We denote by $\bar{\Pi} \subset \Pi$ the set of all non-anticipating, consistent policies. The naming convention stems from the fact that under such policies sample means and variances induced by the policy are consistent estimators of the population counterparts, as formalized in the following proposition.

PROPOSITION 1 **(Consistency of estimators)**. *For any consistent policy* $\boldsymbol{\pi} \in \bar{\Pi}$, $\bar{X}_{jt}(\boldsymbol{\pi}) \to \mu_j$ *and* $S^2_{jt}(\boldsymbol{\pi}) \to \sigma^2_j$ *almost surely as* $t \to \infty$.

The proof of the preceding proposition follows from straightforward application of the strong law of large numbers, and will be omitted. The following example illustrates some care is needed to ensure the consistency property holds.

EXAMPLE 1 (DYNAMIC SAMPLING WHICH IS NOT CONSISTENT). Suppose $k = 2$ and the populations are normal with $(\mu_1, \mu_2) = (2, 1)$ and unit variances. Let $\pi_1 = 1$, $\pi_2 = 2$, and $\pi_t = \arg\max_j\{\bar{X}_{jt}\}$ for each stage $t \geq 3$; that is, a sample is taken from the system with the greatest sample mean. Assume all systems are sampled once initially. After taking a sample from each system, it can be easily seen that the event, $A = \{\bar{X}_{12} \in (-\infty, 0) \text{ and } \bar{X}_{22} \in (1, \infty)\}$, occurs with positive probability. Conditional on this event, system 1 would not be sampled at all if $\bar{X}_{2t} \geq 0$ for all $t \geq 3$. The latter event occurs with positive probability since $\{\sum_{s=2}^{t} X_{2s} : t \geq 3\}$ is a random walk with positive drift and the probability that it falls below zero is strictly less than one. Combined with the fact that $\mathsf{P}(A) > 0$, this policy is not consistent.

3.2. Large Deviations Preliminaries

The probability of false selection, denoted $\mathsf{P}(\mathrm{FS}_T(\boldsymbol{\pi}))$ with $\mathrm{FS}_T(\boldsymbol{\pi}) \coloneqq \{\bar{X}_{1T}(\boldsymbol{\pi}) < \max_{j\neq 1} \bar{X}_{jT}(\boldsymbol{\pi}))\}$, is a widely used criterion for the efficiency of a sampling policy, but the exact evaluation of $\mathsf{P}(\mathrm{FS}_T(\boldsymbol{\pi}))$ is not analytically tractable (see, e.g., the survey paper by Kim and Nelson 2006). However, in an asymptotic regime where the sampling budget goes to infinity, $\mathsf{P}(\mathrm{FS}_T(\boldsymbol{\pi}))$ can be expressed in a closed form.

In particular, for a fixed vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k) \in \Delta^{k-1}$, where Δ^{k-1} a $(k-1)$-simplex defined as

$$\Delta^{k-1} = \left\{ (\alpha_1, \ldots, \alpha_k) \;\middle|\; \sum_{j=1}^{k} \alpha_j = 1 \text{ and } \alpha_j \geq 0 \right\}, \tag{5}$$

we let $\boldsymbol{\pi}^{\boldsymbol{\alpha}} \in \Pi$ be a static allocation rule that targets the (asymptotic) fractional allocation vector $\boldsymbol{\alpha} \in \Delta^{k-1}$. This rule can be easily implemented in several ways. For simplicity consider drawing one

sample from each system and thereafter setting $\pi^{\boldsymbol{\alpha}}_{t+1} = \arg\max_j\{\alpha_j - \alpha_{jt}(\boldsymbol{\pi}^{\boldsymbol{\alpha}})\}$ for $t = k, k+1\ldots$, with ties broken arbitrarily. This policy is *static* in the sense that it is independent of sample observations. Define $M_j(\theta) \coloneqq \mathsf{E}[e^{\theta X_j}]$ and let $\Lambda_j(\theta) \coloneqq \log M_j(\theta)$ be the cumulant generating function. Let $I_j(\cdot)$ denote the Fenchel-Legendre transform of $\Lambda_j(\cdot)$, i.e.,

$$I_j(x) = \sup_\theta\{\theta x - \Lambda_j(\theta)\}, \quad j = 1, \ldots, k. \tag{6}$$

Let $\mathscr{D}_j = \{\theta \in \mathscr{R} : \Lambda_j(\theta) < \infty\}$ and $\mathscr{H}_j = \{\Lambda_j'(\theta) : \theta \in \mathscr{D}_j^0\}$, where $\mathscr{D}^0$ denotes the interior of a set $\mathscr{D}$ and $\Lambda_j'(\theta)$ denotes the derivative of $\Lambda_j(\cdot)$ at θ. In the current section we make the following assumption, which will also be needed for our theoretical results in §5.

Assumption 1. *The interval* $[\min_{j\neq 1}\mu_j, \mu_1] \subset \cap_{j=1}^k \mathscr{H}_j^0$.

To rephrase the preceding assumption, the maximizer of the sup in (6) can be denoted by $\theta_j(x)$ such that $\Lambda_j'(\theta_j(x)) = x$, which is well defined for any $x \in [\min_{j\neq 1}\mu_j, \mu_1]$. Under this assumption, for fixed $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ we have that

$$\frac{1}{T}\log \mathsf{P}(\mathrm{FS}_T(\boldsymbol{\pi}^{\boldsymbol{\alpha}})) \to -\rho(\boldsymbol{\alpha}) \coloneqq -\min_{j\neq 1} G_j(\boldsymbol{\alpha}) \text{ as } T \to \infty, \tag{7}$$

where

$$G_j(\boldsymbol{\alpha}) = \inf_x\{\alpha_1 I_1(x) + \alpha_j I_j(x)\}. \tag{8}$$

The proof for the preceding equations, provided in Appendix A, essentially follows that of Glynn and Juneja (2004). From (7) it follows that $\mathsf{P}(\mathrm{FS}_T(\boldsymbol{\pi}^{\boldsymbol{\alpha}}))$ behaves roughly like $\exp(-\rho(\boldsymbol{\alpha})T)$ for large values of T. Hence, under Assumption 1, $\rho(\cdot)$ is an appropriate measure of asymptotic efficiency closely associated with the probability of false selection, at least for static allocations.

3.3. Problem Formulation

For a fixed budget T, we define the *relative efficiency* $\mathscr{R}_T(\boldsymbol{\pi})$ for a policy $\boldsymbol{\pi} \in \Pi$ to be

$$\mathscr{R}_T(\boldsymbol{\pi}) = \frac{\rho(\boldsymbol{\alpha}_T(\boldsymbol{\pi}))}{\rho^*}, \tag{9}$$

where $\rho^* = \max_{\boldsymbol{\alpha}\in\Delta^{k-1}}\{\rho(\boldsymbol{\alpha})\}$. By definition, the value of $\mathscr{R}_T(\boldsymbol{\pi})$ lies in the interval $[0,1]$; an allocation is considered efficient when $\mathscr{R}_T(\boldsymbol{\pi})$ is close to 1. We are interested in the policy that maximizes the expected relative efficiency with the fixed sampling budget T:

$$\sup_{\boldsymbol{\pi}\in\bar{\Pi}} \mathsf{E}[\mathscr{R}_T(\boldsymbol{\pi})]. \tag{10}$$

The following definition characterizes consistent policies that have "good" asymptotic properties.

Definition 2 (Asymptotic optimality). A policy $\boldsymbol{\pi} \in \bar{\Pi}$ is asymptotically optimal if $\mathsf{E}[\mathscr{R}_T(\boldsymbol{\pi})] \to 1$ as $T \to \infty$.

The asymptotic optimality implies the sampling budget is allocated in a way that the probability of false selection converges to zero at an exponential rate and that the exponent governing the rate of convergence is asymptotically the best possible. It is important to note that a sufficient condition for asymptotic optimality is that $\boldsymbol{\alpha}_T(\boldsymbol{\pi}) \to \boldsymbol{\alpha}^* \in \arg\max_{\boldsymbol{\alpha}\in\Delta^{k-1}}\{\rho(\boldsymbol{\alpha})\}$ in probability as $T \to \infty$. Also note each element of $\boldsymbol{\alpha}^*$ is strictly positive; otherwise $\rho(\boldsymbol{\alpha}^*) = 0$ but we know that $\rho(\boldsymbol{\alpha}) > 0$ for $\boldsymbol{\alpha} = (1/k, \ldots, 1/k)$. Therefore, asymptotic optimality implies consistency.

4. A Two-Moment Approximation and the Proposed Policy

4.1. Motivation

In a general setting with unknown underlying distributions, the functional form of the (asymptotic) probability of false selection $\rho(\cdot)$ is not known, and therefore, it needs to be estimated from sample observations non-parametrically. One such approach would proceed as follows. Let $\hat{\Lambda}_{jt}(\theta;\boldsymbol{\pi})$ be the empirical estimate of the cumulant generating function for system j, i.e.,

$$\hat{\Lambda}_{jt}(\theta;\boldsymbol{\pi}) = \log \frac{\sum_{\tau=1}^{t} \exp(\theta X_{j\tau})\mathbf{I}(\pi_\tau = j)}{N_{jt}}. \tag{11}$$

Define $\hat{I}_{jt}(x;\boldsymbol{\pi})$ as the empirical counterpart of $I_j(x)$ in (6), where $\Lambda_j(\theta)$ is replaced with $\hat{\Lambda}_{jt}(\theta;\boldsymbol{\pi})$. Then, the rate function can be estimated by replacing $I_j(x)$ with $\hat{I}_{jt}(x;\boldsymbol{\pi})$ in (8). However, as noted by Glynn and Juneja (2015), this approach runs into problems at the very first step. Specifically, the estimator of the cumulant generating function, $\hat{\Lambda}_{jt}(\theta;\boldsymbol{\pi})$, tends to be heavy-tailed in most settings,

thereby increasing the possibility of large errors for the corresponding rate function estimate. To see this, observe that if $X_{j\tau}$ has an exponential right tail, then $\exp(\theta X_{j\tau})$ is heavy-tailed for $\theta > 0$ and therefore so is $\hat{\Lambda}_{jt}(\theta; \boldsymbol{\pi})$. This in turn could mean that a significant portion of the sampling budget is "wasted" on estimating the rate function, leaving only a small budget to maximize it.

Further, maximization of the rate function (or its empirical estimate) involves a multi-level optimization: the first inner layer to evaluate $I_j(x)$ in (6) for each j; the second inner layer to evaluate $\rho(\boldsymbol{\alpha})$ in (7); and the outer layer to maximize $\rho(\boldsymbol{\alpha})$ over $\boldsymbol{\alpha} \in \Delta^{k-1}$. This multi-level optimization problem becomes increasingly difficult to solve numerically as the number of systems gets large. In the rest of this section, we suggest a close approximation to the rate function, which allows one to circumvent the difficulties mentioned above.

4.2. A Two-Moment Approximation

We suggest an approximation to $\rho(\boldsymbol{\alpha})$, which is structured around the first two moments of the underlying probability distributions. For each $\boldsymbol{\alpha} \in \Delta^{k-1}$, define $\rho^G(\boldsymbol{\alpha})$ as

$$\rho^G(\boldsymbol{\alpha}) = \min_{j \neq 1} \frac{(\mu_1 - \mu_j)^2}{2(\sigma_1^2/\alpha_1 + \sigma_j^2/\alpha_j)}. \tag{12}$$

PROPOSITION 2 **(Characteristics of $\rho^G(\cdot)$).** *There exists a unique $\boldsymbol{\alpha}^G = \arg\max_{\boldsymbol{\alpha} \in \Delta^{k-1}} \{\rho^G(\boldsymbol{\alpha})\}$ that satisfies the following system of equations:*

$$\frac{(\mu_1 - \mu_i)^2}{\sigma_1^2/\alpha_1^G + \sigma_i^2/\alpha_i^G} = \frac{(\mu_1 - \mu_j)^2}{\sigma_1^2/\alpha_1^G + \sigma_j^2/\alpha_j^G}, \quad \textit{for all} \;\; i, j = 2, \ldots, k \tag{13}$$

$$\frac{\alpha_1^G}{\sigma_1} = \sqrt{\sum_{j=2}^{k} \frac{(\alpha_j^G)^2}{\sigma_j^2}}. \tag{14}$$

The proof for (13)-(14) in the preceding proposition follows from Theorem 1 in Glynn and Juneja (2004). The uniqueness of $\boldsymbol{\alpha}^G$ follows from the fact that $\rho^G(\boldsymbol{\alpha})$ is a strictly concave function of $\boldsymbol{\alpha}$. An immediate corollary of Proposition 2 is that $\boldsymbol{\alpha}^G$ lies in the interior of Δ^{k-1}; if $\alpha_j^G = 0$ for some j, then (13) and (14) would not hold.

REMARK 1 (RELATION TO WELCH'S t-TEST STATISTIC). With an appropriate scaling, each term in the min operator of (12) can be viewed as a population counterpart of Welch's t-test statistic (Welch 1947) given by

$$\frac{\bar{X}_{1t} - \bar{X}_{jt}}{S_{1t}^2/N_{1t} + S_{jt}^2/N_{jt}}, \tag{15}$$

which is used to test the hypothesis that systems 1 and j have equal means when the underlying distributions are Gaussian. Further, in the case with general, non-Gaussian distributions, Chernoff (1952) shows that there is no significant loss in efficiency by using Welch's t-test statistic as long as the probability distributions of systems 1 and j are sufficiently close to each other.

It can be easily seen that $\rho^G(\cdot)$ is identical to $\rho(\cdot)$ when the underlying distributions are Gaussian—hence the superscript G (see Example 1 of Glynn and Juneja 2004). This observation, together with Remark 1, suggests that $\rho^G(\cdot)$ may be closely aligned with $\rho(\cdot)$. If this is the case, then $\boldsymbol{\alpha}_T(\boldsymbol{\pi}) \approx \boldsymbol{\alpha}^G$ implies

$$\begin{aligned} \mathscr{R}_T(\boldsymbol{\pi}) &\approx \frac{\rho(\boldsymbol{\alpha}^G)}{\rho^*} \\ &\approx 1, \end{aligned} \tag{16}$$

where '$\approx$' signifies approximate equality (to be made precise in §5). The preceding observation will be formally stated and proved in Theorem 2.

4.3. Welch Divergence Policy

Motivated by the two-moment approximation, we now propose a dynamic sampling policy, called Welch Divergence (WD), which iteratively estimates $\boldsymbol{\alpha}^G = \arg\max_{\boldsymbol{\alpha} \in \Delta^{k-1}} \{\rho^G(\boldsymbol{\alpha})\}$ from the history of sample observations. The name of the policy stems from its connection to Welch's t-test statistic (Welch 1947) in Remark 1. Denote $\hat{\boldsymbol{\alpha}}_t^G$ the estimator of $\boldsymbol{\alpha}^G$ in stage t. Formally,

$$\hat{\boldsymbol{\alpha}}_t^G = \underset{\boldsymbol{\alpha} \in \Delta^{k-1}}{\arg\max} \left\{ \min_{j \neq b} \frac{(\bar{X}_{bt} - \bar{X}_{jt})^2}{2(S_{bt}^2/\alpha_b + S_{jt}^2/\alpha_j)} \right\}, \tag{17}$$

where $b = \arg\max_j \{\bar{X}_{jt}\}$. If $\bar{X}_{jt} = \bar{X}_{bt}$ for some $j \neq b$, then the objective function of (17) is zero for any $\boldsymbol{\alpha} \in \Delta^{k-1}$ so that $\hat{\boldsymbol{\alpha}}_t^G$ may not be well defined. This event can happen with positive probability

Algorithm 1 WD(n_0, m, ϵ)

(Initialization) For each j, take n_0 samples or until $S_{jt}^2 > 0$

While $t \leq T$ **do**

(Myopic optimization) Solve (17) to obtain $\hat{\boldsymbol{\alpha}}_t^G$; if $\bar{X}_{jt} = \bar{X}_{bt}$ for some $j \neq b$, replace $(\bar{X}_{bt} - \bar{X}_{jt})^2$ in (17) with ϵ/N_{jt}

(Sampling) Let $\pi_{t+s} = \arg\max_{j=1,\ldots,k}\{\hat{\alpha}_{jt}^G - \alpha_{jt}\}$ for $s = 1, \ldots, m$. Let $t = t + m$

End While

when the underlying distributions have jumps. To avoid technical difficulties due to ties, we replace $(\bar{X}_{bt} - \bar{X}_{jt})^2$ with ϵ/N_{jt} for each $j \neq b$ such that $\bar{X}_{jt} = \bar{X}_{bt}$, where $\epsilon > 0$ is a sufficiently small number.

The WD policy matches $\boldsymbol{\alpha}_t(\boldsymbol{\pi})$ with $\hat{\boldsymbol{\alpha}}_t^G$ in each stage, simultaneously making $\hat{\boldsymbol{\alpha}}_t^G$ approach $\boldsymbol{\alpha}^G$ as $t \to \infty$. The policy is summarized in Algorithm 1, with n_0, m, and ϵ being parameters of the policy; n_0 is the number of initial samples from each system, m is the batch size, and ϵ is the constant for the perturbation in case of ties. For ease of exposition, we assume T is a multiple of m.

Remark 2 (Generality of WD). Although we impose Assumption 1 to ensure that the rate function in (7) is well defined, it is important to note that the WD policy does not involve the rate function; it is structured around the first two moments of the underlying distributions. Hence, the WD policy can be implemented in general problem instances whenever two moments exist; for example, see §7.2.3 for the performance of WD with respect to the probability of false selection, the objective of original interest, when the underlying distributions are heavy-tailed and the rate function does not exist.

5. Main Theoretical Results and Qualitative Insights

5.1. Main Theoretical Results

This section contains our three main theoretical results. Theorem 1 provides the asymptotic performance of the WD policy as $T \to \infty$. Theorem 2 validates the two-moment approximation in an

asymptotic regime where $\delta = \mu_1 - \max_{j\neq 1}\{\mu_j\} \to 0$. Finally, Theorem 3 strengthens the preceding theorem in a stylized asymptotic setting where both $T \to \infty$ and $\delta \to 0$.

THEOREM 1 **(Asymptotic performance of WD)**. *The WD policy is consistent. Further, if Assumption 1 is satisfied, then* $\boldsymbol{\alpha}_T(\boldsymbol{\pi}) \to \boldsymbol{\alpha}^G$ *almost surely as* $T \to \infty$ *for* $\boldsymbol{\pi} =$ *WD, and therefore,*

$$\mathsf{E}[\mathscr{R}_T(\boldsymbol{\pi})] \to \frac{\rho(\boldsymbol{\alpha}^G)}{\rho^*} \text{ as } T \to \infty. \tag{18}$$

An important implication of Theorem 1 is that if $\boldsymbol{\alpha}^G \approx \boldsymbol{\alpha}^*$, then $\boldsymbol{\alpha}_T(\boldsymbol{\pi})$ for $\boldsymbol{\pi} = \text{WD}$ is close to $\boldsymbol{\alpha}^*$ for sufficiently large T, and therefore, the WD policy exhibits near-optimal performance asymptotically. The following theorem characterizes the class of problem instances where $\rho(\cdot)$ and $\rho^G(\cdot)$ are closely aligned so that $\boldsymbol{\alpha}^G \approx \boldsymbol{\alpha}^*$.

THEOREM 2 **(Validity of the two-moment approximation)**. *Consider a class of system configurations for which Assumption 1 is satisfied and let* $\delta = \mu_1 - \max_{j\neq 1}\{\mu_j\}$. *If* $\sigma_j \in [\sigma_{\min}, \sigma_{\max}]$ *for* $0 < \sigma_{\min} \leq \sigma_{\max} < \infty$ *for each* j, *then*

$$\frac{\rho(\boldsymbol{\alpha}^G)}{\rho^*} \to 1 \text{ as } \delta \to 0, \tag{19}$$

where $\boldsymbol{\alpha}^G = \arg\max_{\boldsymbol{\alpha}\in\Delta^{k-1}}\{\rho^G(\boldsymbol{\alpha})\}$.

Note that $\rho(\boldsymbol{\alpha})$ tends to zero for each $\boldsymbol{\alpha} \in \Delta^{k-1}$ as $\delta \to 0$, and therefore, so does $|\rho^* - \rho(\boldsymbol{\alpha}^G)|$. Theorem 2 strengthens the preceding argument; $\rho(\boldsymbol{\alpha}^G)$ and ρ^* converge to zero at the *same rate* as $\delta \to 0$, and therefore, maximizers of $\rho(\cdot)$ and $\rho^G(\cdot)$ should coincide in the limit. This suggests that the WD policy, which maximizes $\rho^G(\boldsymbol{\alpha})$, can achieve near-optimal performance with respect to ρ^* when the gap between the best and second-best means is sufficiently close to 0. More precisely, after taking $T \to \infty$, the probability of false selection on a logarithmic scale depends on the underlying distributions only through the first two moments as the gap in the means shrinks to zero. Next we present a simple example to verify the result in Theorem 2.

EXAMPLE 2 (BERNOULLI SYSTEMS). Consider two Bernoulli systems with parameters $\mu_1, \mu_2 \in (0,1)$ with $\mu_1 > \mu_2$ so that $\mathsf{P}(X_j = 1) = \mu_j = 1 - \mathsf{P}(X_j = 0)$ for $j = 1, 2$. The rate functions are

$$I_j(x) = x \log\left(\frac{x}{\mu_j}\right) + (1-x)\log\left(\frac{1-x}{1-\mu_j}\right), \quad j = 1, 2. \tag{20}$$

Using a second-order Taylor expansion of $I_j(x)$ at $x = \mu_j$ and the fact that $\sigma_j^2 = \mu_j(1-\mu_j)$, it can be seen that

$$I_j(x) = \frac{(x-\mu_j)^2}{2\sigma_j^2} + o\left((\mu_1 - \mu_2)^2\right) \tag{21}$$

as $\mu_2 \uparrow \mu_1$, where $f(x) = o(x)$ if $f(x)/x \to 0$ as $x \to 0$. Since $\rho(\boldsymbol{\alpha}) = \inf_x\{\alpha_1 I_1(x) + \alpha_2 I_2(x)\}$, it is not difficult to show that

$$\rho(\boldsymbol{\alpha}) = \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2/\alpha_1 + \sigma_2^2/\alpha_2)} + o\left((\mu_1 - \mu_2)^2\right), \tag{22}$$

where the right-hand side converges to (12) as $\mu_2 \uparrow \mu_1$. Further, it can be easily checked that

$$|\rho(\boldsymbol{\alpha}^*) - \rho(\boldsymbol{\alpha}^G)| \le \max\left\{|\rho(\boldsymbol{\alpha}^*) - \rho^G(\boldsymbol{\alpha}^*)|, |\rho(\boldsymbol{\alpha}^G) - \rho^G(\boldsymbol{\alpha}^G)|\right\}. \tag{23}$$

Combined with the fact that (22) holds for any $\boldsymbol{\alpha} \in \Delta^{k-1}$, the desired result (19) follows.

Figure 2 uses Example 2 to illustrate the proximity between $\rho(\cdot)$ and $\rho^G(\cdot)$ for the case with Bernoulli systems, as a function of $\mu_1 - \mu_2$. Observe that the maximizers of $\rho(\cdot)$ and $\rho^G(\cdot)$ approach each other as the gap in means $(\mu_1 - \mu_2)$ gets smaller, as predicted by the analysis above, in particular, (22).

As noted by Glynn and Juneja (2004), when the true underlying distributions are incorrectly specified as Gaussian, substantially sub-optimal allocations can result; further, the preceding argument is unaffected by taking batches of samples, one of the standard ways to justify the Gaussian assumption. In contrast, Theorem 2 implies that one may approximate $\rho(\cdot)$ by assuming that the underlying distributions are Gaussian, when the difference in means between the best and second-best systems is small enough. This provides a rigorous justification for a two-moment approximation in a specific class of problem instances, primarily affected by the distance between means. We

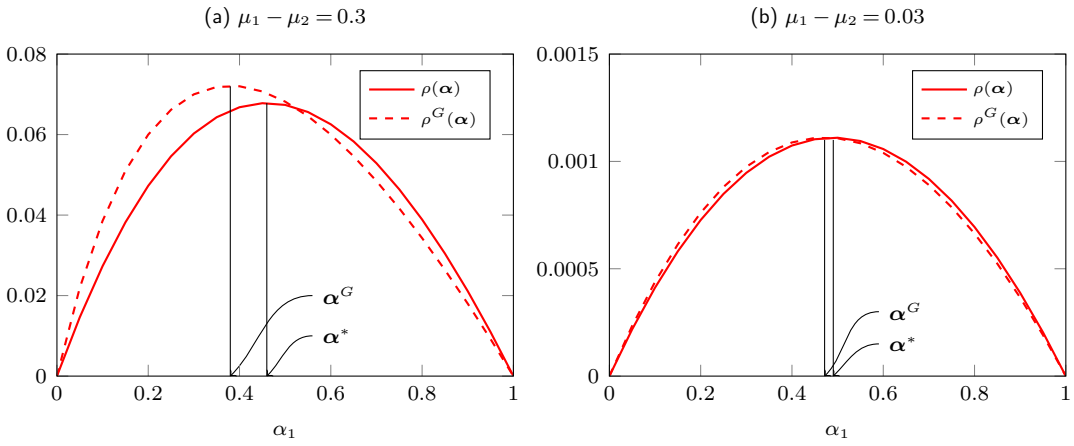

Figure 2 **Approximation error.** Proximity between $\rho(\boldsymbol{\alpha})$ (solid) and $\rho^G(\boldsymbol{\alpha})$ (dashed) for the case with two Bernoulli systems. The ratio $\rho(\boldsymbol{\alpha}^G)/\rho^*$ is 0.9728 in the left panel and 0.9986 in the right panel. The maximizers $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}^G$ tend to be close to each other as $\mu_1 - \mu_2 \to 0$.

should note that the case when the means are "close" implies the problem is "harder" insofar as differentiating the best system from the rest; if the means are "far apart," the problem is simpler. Theorem 2 addresses the former case.

To strengthen our observations in Theorem 2, we consider a stylized sequence S of system configurations defined as follows. Each system configuration is indexed by superscript t. The t-th configuration is denoted by $\mathbf{F}^t = (F_1^t, \ldots, F_k^t)$, where $F_j^t(\cdot)$ is the distribution function for system j; that is, in the t-th configuration, sample observations from system j are independently and identically distributed as $F_j^t(\cdot)$. Let μ_j^t be the mean of the distribution $F_j^t(\cdot)$. The initial configuration is $\mathbf{F}^1 = \mathbf{F}$. We let $\mu_j^1 = \mu_j$ for each j and define $\Delta_{1j} = \mu_1 - \mu_j$ for $j \neq 1$. Also, the configuration $\mathbf{F}^t$, $t \geq 2$, is defined by change of measure from the initial configuration: we fix $F_1^t(x) = F_1(x)$ and the distribution of system $j \neq 1$ is shifted so that $F_j^t(x) = F_j(x - \delta_t \Delta_{1j})$, where $\{\delta_t\}_{t=1}^{\infty}$ is a sequence of nonincreasing, positive numbers with $\delta_1 = 0$. This implies that $(\mu_1^t - \mu_j^t) = \delta_t(\mu_1 - \mu_j)$ for each t, so that the differences in means shrink to zero at the rate governed by δ_t; as a slight abuse of notation, we say that each configuration $\mathbf{F}^t$ is parametrized by δ_t. We write $\mathsf{E}_t[\cdot]$ and $\mathsf{P}_t(\cdot)$ to denote the expectation and probability under the product measure using $\mathbf{F}^t$. For brevity, we write $\mathsf{E}[\cdot] = \mathsf{E}_1[\cdot]$ and $\mathsf{P}(\cdot) = \mathsf{P}_1(\cdot)$.

THEOREM 3 **(Asymptotic validity of the approximation)**. *Consider a sequence S of system configurations $\{\mathbf{F}^t\}_{t=1}^{\infty}$, each of which is parametrized by δ_t. Suppose Assumption 1 is satisfied for the initial configuration $\mathbf{F}^1$. If $t\delta_t^2 \to \infty$ and $\delta_t \to 0$ as $t \to \infty$, then for any static policy $\boldsymbol{\pi}^{\boldsymbol{\alpha}}$ with $\boldsymbol{\alpha} \in \Delta^{k-1}$*

$$\frac{1}{t\delta_t^2} \log \mathsf{P}_t(\mathrm{FS}_t(\boldsymbol{\pi}^{\boldsymbol{\alpha}})) \to -\rho^G(\boldsymbol{\alpha}) \text{ as } t \to \infty, \tag{24}$$

where $\rho^G(\boldsymbol{\alpha})$ is the two-moment approximation (12) *under the configuration $\mathbf{F}^1$.*

Heuristically, the preceding proposition implies that the behavior of $\mathsf{P}_t(\mathrm{FS}_t(\boldsymbol{\pi}))$ for a large value of t is related to δ_t^2; that is, $\mathsf{P}_t(\mathrm{FS}_t(\boldsymbol{\pi})) \approx \exp\{-\rho^G(\boldsymbol{\alpha})t\delta_t^2\}$. This asymptotic regime, where the sampling budget is increasing and the distance between means is diminishing, the two-moment approximation is rigorously justified relative to the probability of false selection objective.

5.2. Qualitative Insights

Let $\delta = \mu_1 - \max_{j\neq 1}\{\mu_j\}$ and write the loss in relative asymptotic efficiency under policy $\boldsymbol{\pi}$ as $\ell(t;\delta) = 1 - \rho(\boldsymbol{\alpha}_t(\boldsymbol{\pi}))/\rho^*$, which can be decomposed into two parts: (i) an approximation error due to the loss of maximizing $\rho^G(\cdot)$ instead of $\rho(\cdot)$; and (ii) an estimation error due to the loss incurred by noisy estimates of means and variances. Formally, for any $t \geq k$

$$\begin{aligned}
\ell(t;\delta) &= 1 - \frac{\rho(\boldsymbol{\alpha}_t(\boldsymbol{\pi}))}{\rho^*} && (25)\\
&= \left(1 - \frac{\rho(\boldsymbol{\alpha}^G)}{\rho^*}\right) + \left(\frac{\rho(\boldsymbol{\alpha}^G) - \rho(\boldsymbol{\alpha}_t(\boldsymbol{\pi}))}{\rho^*}\right) && (26)\\
&= \ell_1(\delta) + \ell_2(t), && (27)
\end{aligned}$$

where $\ell_1(\delta)$ corresponds to the loss due to the approximation error which depends critically on the difference in means δ and $\ell_2(t)$ corresponds to the loss due to sampling (noise). Note that $\ell_1(\delta)$ is independent of t and that $\ell_2(t)$ decreases to 0 almost surely as $t \to \infty$ under the WD policy since $\boldsymbol{\alpha}_t \to \boldsymbol{\alpha}^G$. (Technically, $\ell_2(t)$ also depends on δ but we suppress that dependence.)

For the regime with small t, the latter error dominates the former, i.e., $\ell_1(\delta) \ll \ell_2(t)$, and hence, learning the mean of each system is more important than obtaining a precise approximation of

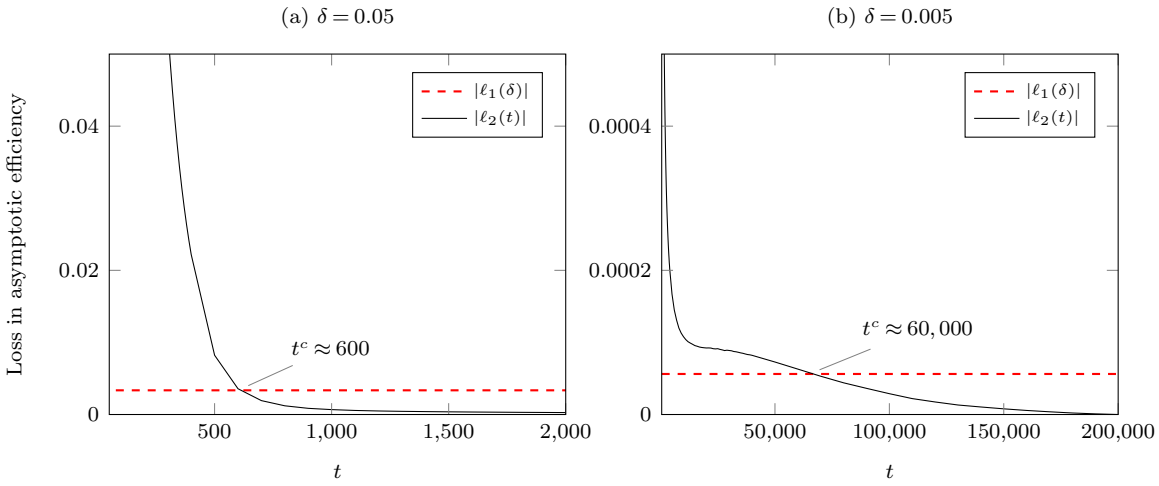

Figure 3 **Bias-variance decomposition of the efficiency loss.** The graphs show results for the WD policy, where each configuration consists of two Bernoulli systems. The graph (a) is for the configuration with $(\mu_1, \mu_2) = (0.9, 0.85)$ and the graph (b) is for the configuration with $(\mu_1, \mu_2) = (0.9, 0.895)$. Each $\ell_2(t)$, the stochastic error, is estimated by taking average of 10^6 generated values of $(\rho(\boldsymbol{\alpha}^G) - \rho(\boldsymbol{\alpha}_t(\boldsymbol{\pi})))/\rho^*$. The standard error of each estimate for $\ell_2(t)$ is less than 10^{-7} in graph (a) and is less than 10^{-9} in graph (b). The two graphs suggest that the crossing point t^c is proportional to δ^2, i.e., at that scale the approximation and stochastic errors are balanced.

$\rho(\cdot)$. Therefore, in this regime $\rho^G(\cdot)$ is a reasonable proxy for $\rho(\cdot)$. When t is large, we have that $\lim_t \ell_2(t) = 0$ from Theorem 1, and therefore, $\liminf_t \ell(t) = \ell_1(\delta)$; that is, in the non-Gaussian case, $\ell_1(\delta)$ is strictly positive and the WD policy cannot recover the full asymptotic efficiency eventually. However, it is important to note that the magnitude of ℓ_1 is not substantial as long as the gap in means is small enough (Theorem 2).

Figure 3 illustrates the magnitudes of $\ell_1(\delta)$ and $\ell_2(t)$ in two simple configurations, each with two Bernoulli systems; one with $\delta = 0.05$ and another one with $\delta = 0.005$, where $\delta = \mu_1 - \mu_2$. For both configurations, there exists $t^c < \infty$ such that $\ell_1(\delta) = \ell_2(t^c)$. For $t \ll t^c$, maximizing $\rho^G(\cdot)$ does not lead to a significant loss in efficiency because $\ell_1(\delta) \ll \ell_2(t)$. Since t^c is roughly proportional to $1/\delta^2$, when δ becomes 10 times smaller, it requires approximately 100 times more samples to reduce $\ell_2(t)$ to the level of $\ell_1(\delta)$. This is consistent with the result of Theorem 3 that the behavior of $\mathsf{P}(\mathrm{FS}_t)$ is closely related to δ^2 for large t.

Algorithm 2 AWD(n_0, m)

(Initialization) For each j, take n_0 samples or until $S_{jt}^2 > 0$

While $t \leq T$ **do**

Let $b = \arg\max_j \{\bar{X}_{jt}\}$. If

$$\frac{\alpha_{bt}}{S_{bt}} < \sqrt{\sum_{j \neq b} \frac{\alpha_{jt}^2}{S_{jt}^2}}, \tag{28}$$

set $\pi_{t+1} = b$. Otherwise, set

$$\pi_{t+s} = \arg\min_{j \neq b} \left\{ \frac{(\bar{X}_{bt} - \bar{X}_{jt})^2}{S_{bt}^2/\alpha_{bt} + S_{jt}^2/\alpha_{jt}} \right\} \tag{29}$$

for $s = 1, \ldots, m$ and let $t = t + m$

End While

6. An Adaptive Welch Divergence Policy

From an implementation standpoint, WD may entail a heavy computational burden as one needs to solve the convex optimization (17) in every stage. In particular, the computation time for solving (17) increases in proportion to the number of systems. Hence, it is desired to construct an alternative policy that is implementable in large-scale problem instances, while retaining the (asymptotic) performance of WD.

6.1. Derivation of the Heuristic

We propose the Adaptive WD (AWD) policy summarized in Algorithm 2. To provide some intuition behind AWD, note that the objective function of (17) is equivalent to $\rho_G(\boldsymbol{\alpha})$ in (12) with the exception that μ_j and σ_j are replaced with their sample estimates, $\bar{X}_{jt}$ and S_{jt}, respectively. Therefore, from Proposition 2 it can be easily seen that the first-order condition for (17) can be written as

$$\frac{(\bar{X}_{bt} - \bar{X}_{it})^2}{S_{bt}^2/\hat{\alpha}_{bt}^G + S_{it}^2/\hat{\alpha}_{it}^G} = \frac{(\bar{X}_{bt} - \bar{X}_{jt})^2}{S_{bt}^2/\hat{\alpha}_{bt}^G + S_{jt}^2/\hat{\alpha}_{jt}^G}, \quad \text{for all} \quad i, j \neq b \tag{30}$$

$$\frac{\hat{\alpha}_{bt}^G}{S_{bt}} = \sqrt{\sum_{j \neq b} \frac{(\hat{\alpha}_{jt}^G)^2}{S_{jt}^2}}. \tag{31}$$

While WD seeks an exact solution for (30) – (31) in every stage, AWD balances the left and right sides of (30) – (31) so that they are satisfied asymptotically as $t \to \infty$.

REMARK 3 (WD AND AWD IN THE TWO-SYSTEM CASE). In the case with $k = 2$ systems, it is straightforward to see that WD and AWD are identical. Specifically, note that, when (28) is satisfied for some stage t, AWD takes a sample from system b. Further, for the WD policy, it can be seen that (28) implies $\hat{\alpha}_{bt}^G > \alpha_{bt}$ and $\hat{\alpha}_{jt}^G < \alpha_{jt}$ for $j \neq b$, and therefore, WD also samples system b. A similar argument holds for the case where (28) is violated.

6.2. Comparison Between WD and AWD

Figure 4 compares the performances of WD and AWD in the case with $k = 10$ normally distributed systems with $\boldsymbol{\mu} = (1, 0.9, \ldots, 0.9)$ and $\boldsymbol{\sigma} = (1, 1, \ldots, 1)$. The parameters for the two policies are set as $n_0 = 200$ and $m = 10$. The two right panels show the frequencies of α_{1t} for $t = 1.6 \cdot 10^4, 1.6 \cdot 10^5$, from which it is clear to see that $\alpha_{1t} \to \alpha_1^G = 0.25$ as $t \to \infty$. The left panel shows that AWD performs slightly better than WD in terms of the probability of false selection; indeed, this is what we observed in most numerical experiments. We note that for a finite sampling budget, an optimal policy should judiciously balance exploring unknown characteristics of the probability distributions and minimizing the probability of false selection. Neither WD nor AWD is optimized for finite-budget performance, but numerical experiments on many different examples suggest that AWD tends to be more efficient in achieving the balance than WD.

From the perspective of implementation, AWD is far more efficient and scalable than WD. In Table 1 we estimate CPU times per stage under the two policies. For the implementation of WD, we use CVX, a package for specifying and solving convex problems (Grant and Boyd 2008, 2014) on a Windows 7 operating system with 32GB RAM and Intel core i7 2.7GHz CPU. Although the implementation of each algorithm was not optimized for CPU time, the relative CPU times illustrate the dramatic difference between the policies, in particular scalability of the AWD policy.

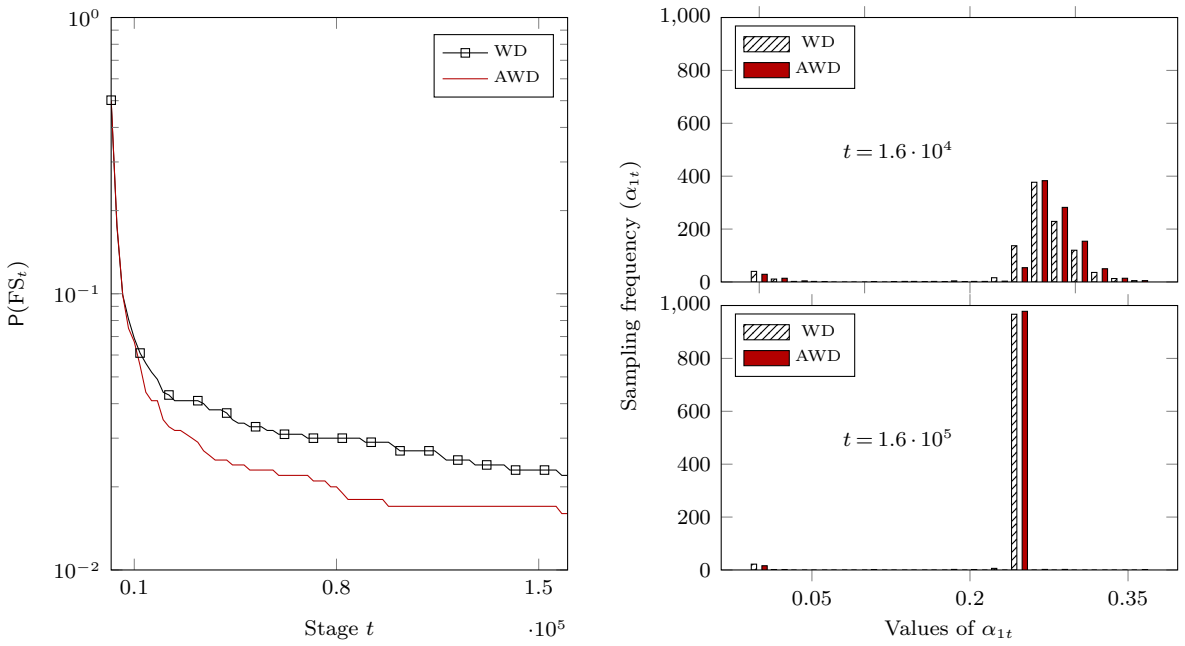

Figure 4 **Performance of the Adaptive WD policy.** Probability of false selection (left) and sampling frequencies, α_{1t}, (right) under the WD and AWD policies. The configuration is characterized by $k = 10$ normally distributed systems with $\boldsymbol{\mu} = (1, 0.9, \ldots, 0.9)$ and $\boldsymbol{\sigma} = (1, 1, \ldots, 1)$. The estimates in the two panels are generated from 1000 simulation trials. The two panels on the right show that $\alpha_{1t} \to \alpha_1^G = 0.25$ as $t \to \infty$ under both policies, while the left panel shows the performance of AWD is slightly better than WD in terms of the probability of false selection.

7. Numerical Testing and Comparison with Other Policies

7.1. Different Sampling Policies

We compare the AWD policy against three other policies: the Optimal Computing Budget Allocation (OCBA) dynamic algorithm (Chen et al. 2000); the Knowledge Gradient (KG) algorithm (Frazier et al. 2008); and the Equal Allocation (EA) policy; see details below.

- *Optimal Computing Budget Allocation (OCBA).* Chen et al. (2000) assumed that the underlying distributions are normal and formulated the problem as follows:

$$\min_{\boldsymbol{\alpha} \in \Delta^{k-1}} \sum_{j \neq b} \bar{\Phi} \left(- \frac{\bar{X}_{bt} - \bar{X}_{jt}}{\sqrt{S_{bt}^2/(\alpha_b t) + S_{jt}^2/(\alpha_j t)}} \right), \tag{32}$$

where $b = \arg\max_j \{\bar{X}_{jt}\}$, $\Phi(\cdot)$ is the cumulative standard normal distribution function, and $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$. They suggested a dynamic policy based on the first-order conditions for (32): n_0 initial

samples are taken from each system, which are then used to allocate m incremental samples. Specifically, compute $(\hat{\alpha}_1, \ldots, \hat{\alpha}_k)$ by solving

$$\frac{\hat{\alpha}_j}{\hat{\alpha}_i} = \left(\frac{S_{jt}/(\bar{X}_{bt} - \bar{X}_{jt})}{S_{it}/(\bar{X}_{bt} - \bar{X}_{it})} \right)^2 \text{ for } i, j \neq b \tag{33}$$

$$\frac{\hat{\alpha}_b}{S_{bt}} = \sqrt{\sum_{j \neq b} \frac{\hat{\alpha}_j^2}{S_{jt}^2}}, \tag{34}$$

which are derived from the first-order conditions of (32). Then allocate $m\hat{\alpha}_j$ samples to system j, ignoring non-integrality of $m\hat{\alpha}_j$. This procedure is repeated until the sampling budget is exhausted.

- *Knowledge Gradient (KG).* We briefly describe the KG policy; further details can be found in Frazier et al. (2008). Let $\tilde{\mu}_j^t$ and $\tilde{\sigma}_j^t$ denote the mean and variance of the posterior distribution for the unknown performance of system j in stage t. Then KG policy maximizes the single-period expected increase in value, $\mathsf{E}_t[\max_j \tilde{\mu}_j^{t+1} - \max_j \tilde{\mu}_j^t]$, where $\mathsf{E}_t[\cdot]$ indicates the conditional expectation with respect to the filtration up to stage t. Although Frazier et al. (2008) mainly discuss the KG policy for the known variance case, we use its version with unknown variance (equivalent to the LL_1 policy in Chick et al. (2007)) in order to be consistent with the other benchmark policies that we test. This policy requires users to set parameters for prior distributions and we use n_0 initial sample estimates to set those. Also, for the purpose of comparison with our policy, we allow the KG policy to take m samples in each stage.

- *Equal Allocation (EA).* This is a simple static benchmark policy, where all systems are equally sampled, implemented as $\pi_t = \arg\min_j N_{jt}$, breaking ties by selecting the system with the smallest index.

Note that the benchmark policies, except for EA, take two parameters; the initial number of samples n_0 and the batch size m in each stage. For comparison with our procedure, we use the same parameters across the policies in each experiment in the following subsection. These are by no means optimal choices, but we have found that the key qualitative conclusions hold for other choices of n_0 and m.

Table 1 CPU times per 100 stages under different policies. The values are estimated by taking an average of 100 simulation trials. The configurations consist of normal distributions with means $\mu_1 = 1, \mu_j = 1.05 - 0.05j$ for $j = 2, \ldots, k$, and standard deviations all set to one.

CPU times (sec)	$k = 10$	$k = 50$	$k = 100$	$k = 500$
WD	2.19	17.40	53.21	426.28
AWD	0.12	0.52	1.64	11.70
KG	0.60	1.15	1.62	16.71
OCBA	0.12	0.53	0.98	13.02
EA	0.03	0.22	0.41	9.52

7.2. Numerical Experiments

The numerical experiments include a series of tests using normal, exponential, and Student-t distributions. $\mathsf{P}(\mathrm{FS}_T)$ is estimated by counting the number of false selections out of M simulation trials, which is chosen so that:

$$\sqrt{\frac{P_T(1-P_T)}{M}} \leq \frac{P_T}{10}, \tag{35}$$

where P_T is the order of magnitude of $\mathsf{P}(\mathrm{FS}_T)$ for a given budget T. This implies the standard error for each estimate of $\mathsf{P}(\mathrm{FS}_T)$ is at least ten times smaller than the value of $\mathsf{P}(\mathrm{FS}_T)$. For example, in Figure 5, the minimum value of $\mathsf{P}(\mathrm{FS}_T)$ is 10^{-2}, so we choose M to be at least 10^4 by this argument. Consequently, we have sufficiently high confidence that the results are not attributed to simulation error.

In Table 1, we compare the computation times of different algorithms discussed in this section. The algorithms are implemented in MATLAB on a Windows 7 operating system with 32GB RAM and Intel core i7 2.7GHz CPU. Although the implementation of each algorithm is not optimized for performance, the relative CPU times illustrate complexity between the algorithms as the number of systems increases. Note that the computation time of AWD is comparable to that of OCBA which is one of the most widely cited policies in literature due to its simple allocation rule.

7.2.1. Normal distribution. First we consider configurations with normally distributed systems with monotonically decreasing means where $\mu_j = 1.05 - 0.05j$. Three sets of standard deviations are considered: constant variance $\boldsymbol{\sigma} = (1, \ldots, 1)$; increasing variance $\boldsymbol{\sigma} = (0.4, \ldots, 1.6)$; and decreasing variance $\boldsymbol{\sigma} = (1.6, \ldots, 0.4)$. We examine performance for both small ($k = 10$) and large-scale ($k = 500$) configurations. For all cases, we set $m = 10$ and the number of initial samples is $n_0 = 0.1T/k$; that is, 10% of the total budget is allocated to the initilization of sampling.

In Figure 5, the estimated $\mathsf{P}(\mathrm{FS}_t)$ is shown as a function of stage t. Poor performance of the KG policy, especially in the case with large number of systems, is anticipated as it is not designed to minimize the probability of false selection. In the case with $k = 10$ systems, both the AWD and OCBA policies perform comparably. However, in the case with $k = 500$ systems, the OCBA policy exhibits relatively poor performance compared to AWD. We find that when the sample average of the true best system (i.e., system 1) is far smaller than those of the others, OCBA myopically allocates too few additional samples to system 1, making it appear to be "non-best" for the rest of the sampling horizon. Further, despite the drastic difference in performance between the AWD and OCBA policies, the computational costs of these are comparable as can be seen from Table 1.

7.2.2. Exponential distribution. As an example of a non-Gaussian distribution, we consider configurations where system performances are exponentially distributed. In this experiment, in addition to the benchmark policies above, we consider another variant of the WD policy, with $\rho^G(\cdot)$ in Algorithm 1 replaced with:

$$\rho(\boldsymbol{\alpha}) = \min_{j \neq 1} \left\{ -\alpha_1 \log \left(\frac{(\alpha_1 + \alpha_j)/\mu_1}{\alpha_1/\mu_1 + \alpha_j/\mu_j} \right) - \alpha_j \log \left(\frac{(\alpha_1 + \alpha_j)/\mu_j}{\alpha_1/\mu_1 + \alpha_j/\mu_j} \right) \right\}, \tag{36}$$

which is the rate function when the underlying probability distributions are exponential. We call this policy modified GJ, since it is adapted from the sequential procedure in Glynn and Juneja (2004) but takes as primitive knowledge of the rate function $\rho(\cdot)$. Of course, the knowledge of $\rho(\cdot)$ is an unrealistic assumption in practice, but this serves as an aggressive benchmark.

We consider two configurations, each with ten exponentially distributed systems. In the first configuration, the gap between the best and non-best systems is large: $\boldsymbol{\mu} = (1, 0.9, 0.8, \ldots, 0.5, 0.5)$. In the second configuration, the gap is smaller than the previous one: $\boldsymbol{\mu} = (1, 0.95, 0.9, \ldots, 0.5, 0.5)$.

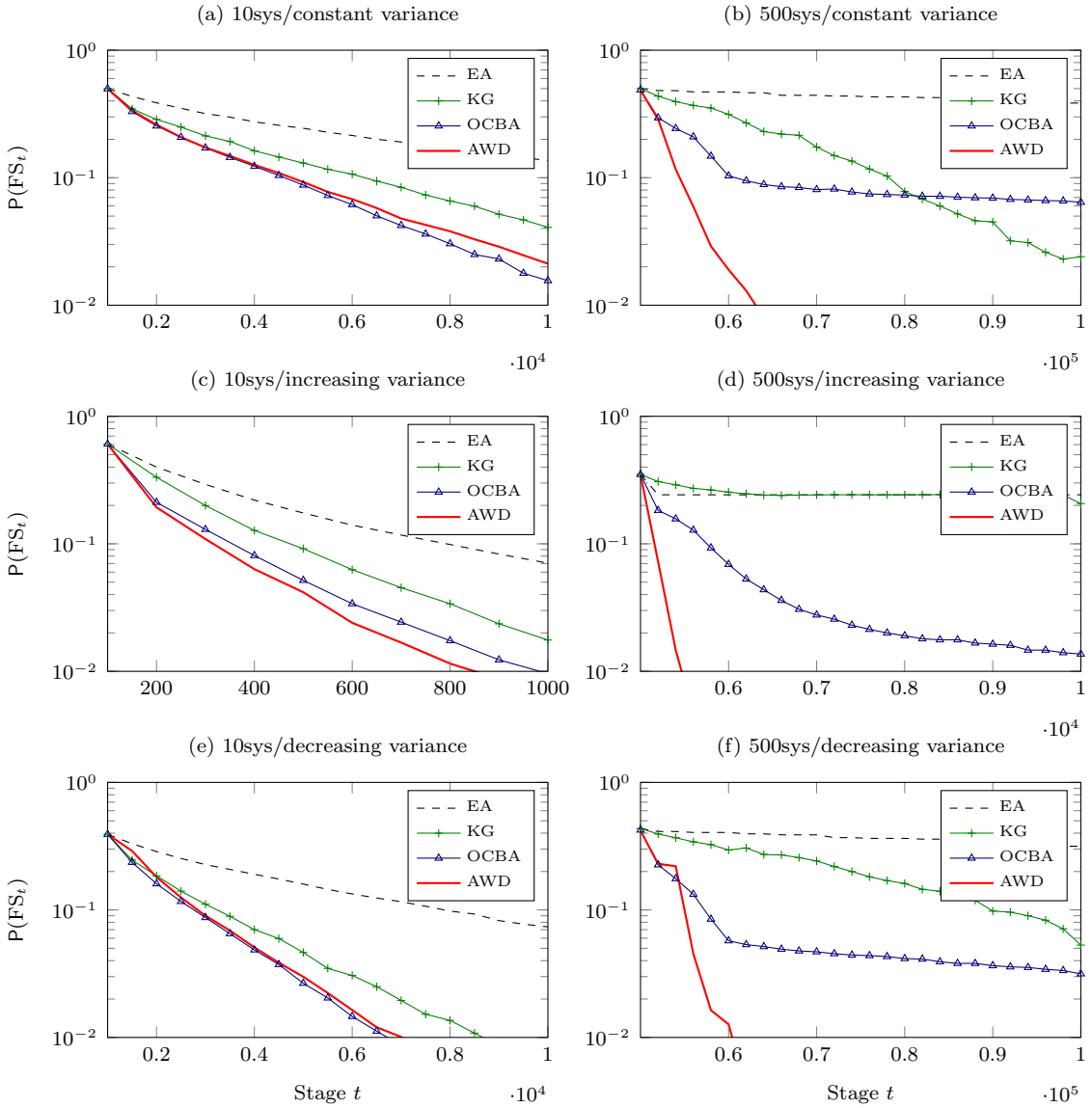

Figure 5 **Probability of false selection in normal distribution case.** $\mathsf{P}(\mathrm{FS}_t)$ is plotted as a function of stage t in log-linear scale. The four sampling policies are: equal allocation (EA); Knowledge Gradient (KG); Optimal Computing Budget Allocation (OCBA); and the AWD policy. For all cases, $\mu_j = 1.05 - 0.05j$ for $j = 1, \ldots, k$. Three sets of standard deviations are considered: constant variance $\boldsymbol{\sigma} = (1, \ldots, 1)$; increasing variance $\boldsymbol{\sigma} = (0.4, \ldots, 1.6)$; and decreasing variance $\boldsymbol{\sigma} = (1.6, \ldots, 0.4)$.

In the first configuration, the modified GJ policy outperforms the others due to the prior information on the rate function, which is "baked into" their algorithm. In the second configuration, however, the modified GJ and AWD policies perform similarly. This observation is consistent with Theorem 2; specifically, when the gap between the best and "second-best" systems is small enough,

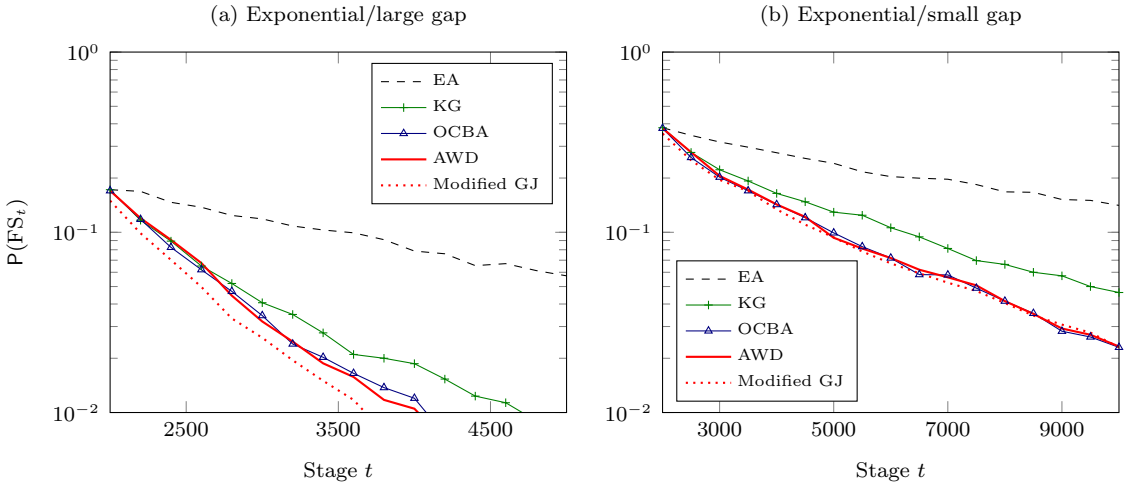

Figure 6 **Probability of false selection in exponential distribution case.** $\mathsf{P}(\mathrm{FS}_t)$ is plotted as a function of t in log-linear scale. The five sampling policies are: equal allocation (EA); Knowledge Gradient (KG); Optimal Computing Budget Allocation (OCBA); the AWD policy; and the modified GJ policy. The two configurations are characterized by (a) $\boldsymbol{\mu} = (1, 0.9, 0.8, \ldots, 0.5, 0.5)$ and (b) $\boldsymbol{\mu} = (1, 0.95, 0.9, \ldots, 0.5, 0.5)$, each with $k = 10$ systems.

it is safe to maximize $\rho^G(\cdot)$ instead of $\rho(\cdot)$. In other words, one does not lose much by the lack of information that the underlying distribution is exponential. Also, observe that the OCBA policy performs competitively with the modified GJ and AWD policies. Although the OCBA policy is developed based on the premise of underlying normal distributions, its application in the exponential case is well-justified when the gap in means is sufficiently small.

7.2.3. t **distribution.** As another example of non-Gaussian distributions, we consider configurations where systems follow the heavy-tailed Student-t distribution. Each system is parameterized by the location parameter μ_j and the degrees of freedom ν_j; the variance for system j is given by $\nu_j/(\nu_j - 2)$. We take two configurations, each with $k = 10$ systems: The first configuration is characterized by $\nu_j = 2.13$ with means $\mu_j = 1.2 - 0.2j$ for $j = 1, \ldots, k$. The second configuration is characterized by $\nu_j = 4$ with means $\mu_j = 1.07 - 0.07j$ for $j = 1, \ldots, k$. The values of ν_j are chosen so that the gap in means between systems j and $j+1$, in terms of the number of standard deviations, is identical in the two configurations, but the distributions in the first configuration are more heavy-tailed than those in the second.

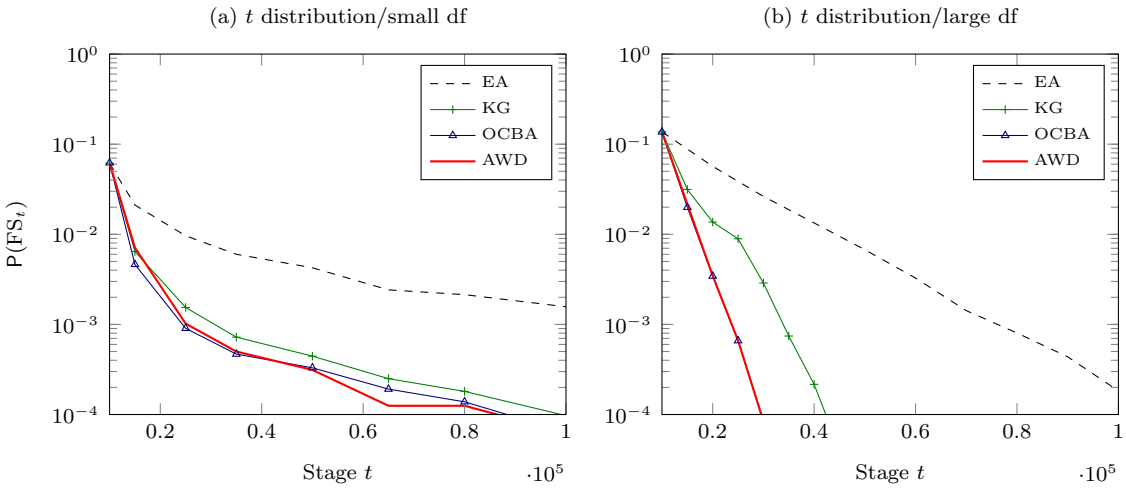

Figure 7 **Probability of false selection in t distribution case.** $\mathsf{P}(\mathrm{FS}_t)$ is plotted as a function of stage t in log-log scale. The four sampling policies are: equal allocation (EA); Knowledge Gradient (KG); Optimal Computing Budget Allocation (OCBA); and the AWD policy. The two configurations are characterized by (a) $\nu_j = 2.13$ and $\mu_j = 1.2 - 0.2j$ and (b) $\nu_j = 4$ and $\mu_j = 1.07 - 0.7j$, respectively, for $j = 1, \ldots, 10$.

Since the theoretical results in this paper do not hold in the heavy-tailed case, the application of AWD can be viewed as a heuristic. Further, unlike the exponential convergence of the probability of false selection in light-tailed environments, Broadie et al. (2007) show that, under a static policy characterized by an allocation vector $\boldsymbol{\alpha} \in \Delta^{k-1}$, the probability of false selection converges to zero at a polynomial rate that does not depend on $\boldsymbol{\alpha}$. This implies the policies discussed in this section may exhibit identical performance in terms of the convergence rate of the probability of false selection. However, the results in Figure 7 indicate that, despite the presence of heavy tails, it is still important to judiciously allocate samples based on means and variances in order to minimize the probability of false selection. In particular, one can observe that the AWD and OCBA policies outperform the others in both cases of Figure 7.

8. Concluding Remarks

By analyzing the asymptotics of the probability of false selection, we were able to obtain a close approximation to the large deviations rate function, and leverage it to design well-performing dynamic sampling policies. An important conclusion from our results is that the rate function

for general distributions can be closely approximated by using only the first two moments, which allowed us to construct the WD and AWD policies that are tractable for implementation purposes.

Although numerical analyses in §6 provide evidence that AWD possesses the asymptotic properties of WD as in Theorem 1, the performance of AWD is not analyzed in a precise mathematical manner except for the case with $k = 2$ systems (Remark 3); the case with $k > 2$ remains an open question.

While we primarily focused on the probability of false selection for a single best system, the methodology developed in this paper can be extended to other criteria. For example, it can be applied to the problem of selecting n (> 1) best systems out of k. If $\mathsf{P}(\mathrm{FS}_t)$ is defined as the probability of false selection of the $(k - n)$ inferior systems, one can track the behavior of $\mathsf{P}(\mathrm{FS}_t)$ for large t using the relationship (7), from which it is possible to develop a tractable policy using the mechanics used in this paper.

Acknowledgments

The authors are grateful to Stephen Chick, the area editor, for his feedback in helping improve the paper. They also thank the associate editor and the three referees for their thoughtful and detailed comments on the earlier version of the manuscript; their suggestions greatly improved the quality and the presentation of the authors' work. Assaf Zeevi wishes to thank the financial support from grants BSF 2010466 and NSF CNS-0964170.

References

Audibert J, Bubeck S (2010) Best arm identification in multi-armed bandits. *23th Conf. on Learn. Theory*, 41–53.

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

Bechhofer R (1954) A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.* 25:16–39.

Blanchet J, Liu J, Zwart B (2008) Large deviations perspective on ordinal optimization of heavy-tailed systems. Mason SJ, Hill RR, Mönch L, O Rose JWF T Jefferson, eds., *Proc. 2008 Winter Simulation Conf.*, 489–494 ((IEEE, Piscataway, NJ)).

Broadie M, Han M, Zeevi A (2007) Implications of heavy tails on simulation-based ordinal optimization. Henderson SG, Biller B, Hsieh MH, Tew JD, Barton RR, eds., *Proc. 2007 Winter Simulation Conf.*, 439–447 ((IEEE, Piscataway, NJ)).

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

Bubeck S, Munos R, Stoltz G (2009) Pure exploration in multi-armed bandits problems. *International Conf. on Alg. Learn. Theory*, 23–37 (Springer).

Bubeck S, Wang T, Viswanathan N (2013) Multiple identifications in multi-armed bandits. *Proc. 30th International Conf. on Machine Learn.*, 258–265.

Chen CH (1995) An effective approach to smartly allocate computing budget for discrete event simulation. *Proc. 34th IEEE Conf. on Decision and Control*, 2598–2603.

Chen CH, Chen HC, Dai L (1996) A gradient approach for smartly allocating computing budget for discrete event simulation. Charnes JM, Morrice DJ, Brunner DT, Swain JJ, eds., *Proc. 1996 Winter Simulation Conf.*, 398–405 ((IEEE, Piscataway, NJ)).

Chen CH, Donohue K, Yücesan E, Lin J (2003) Optimal computing budget allocation for Monte Carlo simulation with application to product design. *Simulation Modeling Practice and Theory* 11:57–74.

Chen CH, Lin J, Yücesan E, Chick SE (2000) Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems* 10:251–270.

Chen HC, Dai L, Chen CH, Yücesan E (1997) New development of optimal computing budget allocation for discrete event simulation. Andradóttir S, Healy KJ, Withers DH, Nelson BL, eds., *Proc. 1997 Winter Simulation Conf.*, 334–341 ((IEEE, Piscataway, NJ)).

Chernoff H (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* 23:493–507.

Chick SE, Branke J, Schmidt C (2007) New greedy myopic and existing asymptotic sequential selection procedures: preliminary empirical results. Henderson SG, Biller B, Hsieh MH, Shortle J, Tew JD, Barton RR, eds., *Proc. 2007 Winter Simulation Conf.*, 289–296 ((IEEE, Piscataway, NJ)).

Chick SE, Inoue K (2001) New two-stage and sequential procedures for selecting the best simulated system. *Oper. Res.* 49:732–743.

Dembo A, Zeitouni O (2009) *Large deviations techniques and applications* (New York: Springer Science & Business Media).

Even-Dar E, Mannor S, Y M (2002) Pac bounds for multi-armed bandit and markov decision processes. Kivinen J, Sloan RH, eds., *Prof. 2002 Computational Learn. Theory Conf.*, 255–270 (Springer).

Even-Dar E, Mannor S, Y M (2006) Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. of Machine Learn. Res.* 7:1079–1105.

Fan W, Hong J, Nelson BL (2016) Indifference-zone-free selection of the best. *Oper. Res.* 64:1499–1514.

Frazier P, Powell WB, Dayanik S (2008) A knowledge-gradient policy for sequential information collection. *SIAM J. on Control and Optim.* 47:2410–2439.

Glynn P, Juneja S (2004) A large deviations perspective on ordinal optimization. Ingalls RG, Rossetti MD, Smith JS, Peters BA, eds., *Proc. 2004 Winter Simulation Conf.*, 577–585 ((IEEE, Piscataway, NJ)).

Glynn P, Juneja S (2015) Ordinal optimization-empirical large deviations rate estimators, and stochastic multi-armed bandits. *arXiv:1507.04564* .

Goldsman D, Kim SH, Marshall WS, Nelson BL (2002) Ranking and selection for steady-state simulation: Procedures and perspectives. *INFORMS J. on Comput.* 14:2–19.

Grant M, Boyd S (2008) Graph implementations for nonsmooth convex programs. Blondel V, Boyd S, Kimura H, eds., *Recent Advances in Learn. and Control*, 95–110, Lecture Notes in Control and Information Sciences (Springer-Verlag Limited), `http://stanford.edu/~boyd/graph_dcp.html`.

Grant M, Boyd S (2014) CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`.

Hong LJ (2006) Fully sequential indifference-zone selection procedures with variance-dependent sampling. *Naval Research Logistics* 53:464–476.

Jamieson KG, Malloy M, Nowak RD, Bubeck S (2014) lil'UCB: An optimal exploration algorithm for multi-armed bandits. *J. of Machine Learn. Res.* 35:423–439.

Kalyanakrishnan S, Tewari A, Auer P, Stone P (2012) Pac subset selection in stochastic multi-armed bandits. *Proc. 29th International Conf. on Machine Learn.*, 655–662.

Kim SH, Nelson BL (2001) A fully sequential procedure for indifference-zone selection in simulation. *ACM TOMACS* 11:251–273.

Kim SH, Nelson BL (2006) Selecting the best system. Henderson SG, Nelson BL, eds., *Handbooks in Operations Research and Management Science: Simulation*, volume 13, chapter 17, 501–534 ((Elsevier, Boston)).

Mannor S, Tsitsiklis JN (2004) The sample complexity of exploration in the multi-armed bandit problem. *J. of Machine Learn. Res.* 5:623–648.

Nelson BL, Swann J, Goldsman D, Song W (2001) Simple procedures for selecting the best simulated system when the number of alternatives is large. *Oper. Res.* 49:950–963.

Pasupathy R, Hunter SR, Pujowidianto NA, Lee LH, Chen CH (2015) Stochastically constrained ranking and selection via score. *ACM TOMACS* 25:1–26.

Paulson E (1964) A sequential procedure for selecting the population with the largest mean from k normal populations. *Ann. Math. Statist.* 35:174–180.

Rinott Y (1978) On two-stage selection procedures and related probability-inequalities. *Comm. Statist.* 7:799–811.

Russo D (2016) Simple bayesian algorithms for best arm identification. *arXiv:1602.08448* .

Welch BL (1947) The generalization of student's problem when several different population variances are involved. *Biometrika* 34:28–35.

Dongwook Shin is an assistant professor at the HKUST Business School. His research centers on revenue management with applications in e-commerce and sequential decision making in operations research context.

Mark Broadie is the Carson Family Professor of Business at the Graduate School of Business, Columbia University. His research addresses issues in quantitative finance and sports analytics and, more generally, methods for decision making under uncertainty.

Assaf Zeevi is the Kravis Professor of Business at the Graduate School of Business, Columbia University. His research focuses on the formulation and analysis of mathematical models of complex systems, with particular research and teaching interests that lie in the intersection of operations research, statistics, computer science, and economics. Recent application areas have been motivated by problems in healthcare analytics, dynamic pricing, recommendations engines and personalization, and the valuation and monetization of digital goods.

Appendices

A. Proofs for Main Results

In order to prove Equation (7), we need the following lemma. Proofs for all auxiliary lemmas are provided in Appendix B.

LEMMA A.1. *Fix a static allocation rule* $\boldsymbol{\pi}^{\boldsymbol{\alpha}}$ *for some* $\boldsymbol{\alpha} \in \Delta^{k-1}$. *Let* $Z_t = (\bar{X}_{1t}(\boldsymbol{\pi}^{\boldsymbol{\alpha}}), \bar{X}_{jt}(\boldsymbol{\pi}^{\boldsymbol{\alpha}}))$ *for some* $j = 2, \ldots, k$, *and denote the logarithmic moment generating function of* Z_t *by* $\Lambda_t(\lambda_1, \lambda_j) = \log \mathsf{E}[e^{\lambda_1 \bar{X}_{1t}(\boldsymbol{\pi}^{\boldsymbol{\alpha}}) + \lambda_j \bar{X}_{jt}(\boldsymbol{\pi}^{\boldsymbol{\alpha}})}]$ *for* $t \geq k$. *Then, the rate function of* Z_t, $I_{1j}(x_1, x_j)$, *equals* $\alpha_1 I_1(x_1) + \alpha_j I_j(x_j)$.

Proof for Equation (7). In this proof we fix $\boldsymbol{\pi} = \boldsymbol{\pi}^{\boldsymbol{\alpha}}$, where for clarity we suppress the function arguments. Observe that

$$\max_{j=2,\ldots,k} \mathsf{P}(\bar{X}_{1T} < \bar{X}_{jT}) \leq \mathsf{P}(\mathrm{FS}_T) \leq (k-1) \max_{j=2,\ldots,k} \mathsf{P}(\bar{X}_{1T} < \bar{X}_{jT}). \tag{A.1}$$

Further, if, for each $j = 2, \ldots, k$,

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}(\bar{X}_{1T} < \bar{X}_{jT}) = -H_j(\alpha_1, \alpha_j) \tag{A.2}$$

for some rate function $H_j(\cdot, \cdot)$, then

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}(\mathrm{FS}_T) = -\min_{j=2,\ldots,k} \{H_j(\alpha_1, \alpha_j)\}. \tag{A.3}$$

From Lemma A.1, it can be easily seen that

$$H_j(\alpha_1,\alpha_j)=\inf_{x_j\geq x_1}\{\alpha_1 I_1(x_1)+\alpha_j I_j(x_j)\}. \tag{A.4}$$

Since both $I_1(x)$ and $I_j(x)$ are decreasing in x for $x<\mu_j$ and increaseing in x for $x>\mu_1$, it suffices to search for the minimum for $\mu_j\leq x_1\leq x_j\leq\mu_1$. In this region, $I_1(x)$ is decreasing and $I_j(x)$ is increasing with x, so we establish that

$$H_j(\alpha_1,\alpha_j)=\inf_{x}\{\alpha_1 I_1(x)+\alpha_j I_j(x)\}, \tag{A.5}$$

which is identical to $G_j(\boldsymbol{\alpha})$ defined in (8). This completes the proof of Equation (7). $\square$

Additional notation is introduced for the proof of Theorem 1. Fix $\boldsymbol{x}=(x_1,\ldots,x_k)\in\mathbb{R}^k$ and $\boldsymbol{y}=(y_1,\ldots,y_k)\in\mathbb{R}^k_+$. Define $B(\boldsymbol{x})$ as

$$B(\boldsymbol{x})\coloneqq\{i\in\{1,\ldots,k\}\mid x_i=\max_j\{x_j\}\} \tag{A.6}$$

Note that we allow the case with $|B(\boldsymbol{x})|>1$, where $|B(\boldsymbol{x})|$ represents the cardinality of the set $B(\boldsymbol{x})$. Let Θ be the set of all $(\boldsymbol{x},\boldsymbol{y})\in\mathbb{R}^k\times\mathbb{R}^k_+$ with $|B(\boldsymbol{x})|<k$. For each $(\boldsymbol{x},\boldsymbol{y})\in\Theta$ and $\boldsymbol{\alpha}\in\Delta^{k-1}$, define

$$H(\boldsymbol{\alpha};\boldsymbol{x},\boldsymbol{y})=\min_{i\in B(\boldsymbol{x}),j\notin B(\boldsymbol{x})}\frac{(x_i-x_j)^2}{(y_i^2/\alpha_i+y_j^2/\alpha_j)}. \tag{A.7}$$

Note that $\rho(\boldsymbol{\alpha})=H(\boldsymbol{\alpha};\boldsymbol{\mu},\boldsymbol{\sigma})$. Define $\boldsymbol{\alpha}(\boldsymbol{x},\boldsymbol{y})$ to be the maximizer of $H(\boldsymbol{\alpha};\boldsymbol{x},\boldsymbol{y})$, i.e.,

$$\boldsymbol{\alpha}(\boldsymbol{x},\boldsymbol{y})=\arg\max_{\boldsymbol{\alpha}\in\Delta^{k-1}}H(\boldsymbol{\alpha};\boldsymbol{x},\boldsymbol{y}). \tag{A.8}$$

Note that $H(\boldsymbol{\alpha};\boldsymbol{x},\boldsymbol{y})$ is a strictly concave function of $\boldsymbol{\alpha}$ since it is the minimum of the strictly concave functions. Hence, $\boldsymbol{\alpha}(\boldsymbol{x},\boldsymbol{y})$ in (A.8) is well defined. The proof of Theorem 1 requires the following lemma.

LEMMA A.2. *Fix $(\boldsymbol{x},\boldsymbol{y})\in\Theta$. Consider a sequence of parameters $(\boldsymbol{x}^t,\boldsymbol{y}^t)\in\Theta$ such that $x_j^t\to x_j$ and $y_j^t\to y_j$ as $t\to\infty$ for each j. Then, $\boldsymbol{\alpha}(\boldsymbol{x}^t,\boldsymbol{y}^t)\to\boldsymbol{\alpha}(\boldsymbol{x},\boldsymbol{y})$ as $t\to\infty$ and the vector $\boldsymbol{\alpha}(\boldsymbol{x},\boldsymbol{y})$ is strictly positive.*

Proof of Theorem 1. In this proof, we first show that the WD policy is consistent and then show that it is asymptotically optimal. We fix $\boldsymbol{\pi} = \text{WD}$, where for clarity we suppress the function arguments.

Consistency. Fix a sequence of samples, $\{(X_{j1}, X_{j2}, \ldots)\}_{j=1}^{k}$, and let $F \subset \{1, 2, \ldots, k\}$ be the set of systems that are sampled only finitely many times and let $I \coloneqq \{1, 2, \ldots, k\} \setminus F$. Suppose, towards a contradiction, that F is non-empty and let $\tau < \infty$ be the last time that the systems in F are sampled. It can be seen from Proposition 1 that, for each system $j \in I$, $\bar{X}_{jt}$ and S_{jt}^2 converges to μ_j and σ_j^2 almost surely as $t \to \infty$, respectively. Also, for any system $j \in F$, $\bar{X}_{jt}$ is constant subsequent to the stage τ. Formally, define $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_k)$ and $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_1, \ldots, \hat{\sigma}_k)$ with

$$(\hat{\mu}_j, \hat{\sigma}_j^2) = \begin{cases} (\max_i\{\bar{X}_{i\tau}\} - (\epsilon/N_{j\tau})^{0.5}, S_{j\tau}^2) \text{ for } j \in F \text{ with } \bar{X}_{j\tau} = \max_i\{\bar{X}_{i\tau}\} \\ (\bar{X}_{j\tau}, S_{j\tau}^2) \text{ for } j \in F \text{ with } \bar{X}_{j\tau} < \max_i\{\bar{X}_{i\tau}\} \\ (\mu_j, \sigma_j^2) \text{ for } j \in I, \end{cases} \tag{A.9}$$

so that $\bar{X}_{jt} \to \hat{\mu}_j$ and $S_{jt}^2 \to \hat{\sigma}_j^2$ for each $j = 1, \ldots, k$.

It can be easily seen that $|B(\hat{\boldsymbol{\mu}})| < k$ by the definition of $\hat{\boldsymbol{\mu}}$, where $B(\cdot)$ is defined in (A.6) and $|B(\cdot)|$ represents the cardinality of the set $B(\cdot)$. Note that $\hat{\boldsymbol{\alpha}}_t^G = \boldsymbol{\alpha}(\bar{\boldsymbol{X}}_t, \boldsymbol{S}_t)$ for each $t \geq k$ and define $\hat{\boldsymbol{\alpha}} \coloneqq \boldsymbol{\alpha}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}})$, where $\hat{\boldsymbol{\alpha}}_t^G$ is defined in (17) and the function $\boldsymbol{\alpha}(\cdot, \cdot)$ is defined in (A.8). Applying Lemma A.2 with $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ replaced with $(\bar{\boldsymbol{X}}_t, \boldsymbol{S}_t)$ and $(\boldsymbol{x}, \boldsymbol{y})$ replaced with $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}})$, it follows that $\hat{\boldsymbol{\alpha}}_t^G \to \hat{\boldsymbol{\alpha}}$ as $t \to \infty$ with $\tilde{\boldsymbol{\alpha}} > 0$. Further, by construction of our policy, $\boldsymbol{\alpha}_t(\boldsymbol{\pi}) - \hat{\boldsymbol{\alpha}}_t^G \to 0$ as $t \to \infty$, and therefore, $\boldsymbol{\alpha}_t(\boldsymbol{\pi}) \to \hat{\boldsymbol{\alpha}} > 0$ as $t \to \infty$. However, this contradicts our assumption because each system $j \in F$ is sampled only finitely many times so that $\alpha_{jt}(\boldsymbol{\pi}) \to 0$. Consequently, F is empty with probability 1 and the proposed policy is consistent.

Asymptotic performance. Since the proposed policy is consistent, $\bar{X}_{jt} \to \mu_j$ and $S_{jt}^2 \to \sigma_j^2$ almost surely for each $j = 1, \ldots, k$. Therefore, applying Lemma A.2 with $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ replaced with $(\bar{\boldsymbol{X}}_t, \boldsymbol{S}_t)$ and $(\boldsymbol{x}, \boldsymbol{y})$ replaced with $(\boldsymbol{\mu}, \boldsymbol{\sigma})$, it follows that $\hat{\boldsymbol{\alpha}}_t^G \to \boldsymbol{\alpha}^G = \arg\max_{\boldsymbol{\alpha} \in \Delta^{k-1}}\{\rho^G(\boldsymbol{\alpha})\}$. Moreover, by construction of the policy it is not difficult to see that the term $\boldsymbol{\alpha}_t(\boldsymbol{\pi}) - \hat{\boldsymbol{\alpha}}_t^G \to 0$ as $t \to \infty$. Consequently, it follows that $\boldsymbol{\alpha}_t(\boldsymbol{\pi}) \to \boldsymbol{\alpha}^G$ as $t \to \infty$, which completes the proof of the theorem. $\square$

Proof of Theorem 2. In this proof, we consider a sequence of system configurations and, as a slight abuse of notation, each configuration is uniquely parametrized by $\delta > 0$. We assume without loss of generality that $\mu_1 > \mu_2 = \max_{j>2}\{\mu_j\}$ for every system configuration.

Case 1. Consider a sequence of system configurations where $\rho(\boldsymbol{\alpha}) = G_2(\boldsymbol{\alpha})$ for each configuration. Recalling the definition of $\rho(\boldsymbol{\alpha})$ in (7), the preceding condition implies that the probability of falsely selecting system $j = 3, \ldots, k$ is dominated by the probability of falsely selecting system 2. In this case, we can write, for sufficiently small $\delta > 0$,

$$\rho(\boldsymbol{\alpha}) = \inf_x \{\alpha_1 I_1(x) + \alpha_2 I_2(x)\}. \tag{A.10}$$

Since both $I_1(x)$ and $I_2(x)$ are decreasing in x for $x < \mu_2$ and increasing in x for $x > \mu_1$, it suffices to search for the infimum for $x \in [\mu_2, \mu_1]$. In this region, $I_1(x)$ is decreasing and $I_2(x)$ is increasing.

Define $\theta_j(x)$, $j = 1, 2$, as the value of θ that satisfies $\Lambda_j'(\theta) = x$, which exist for all $x \in [\mu_2, \mu_1]$ due to Assumption 1. Observe that $I_j(x) = x\theta_j(x) - \Lambda_j(\theta_j(x))$ and

$$I_j'(x) = \theta_j(x) + x\theta_j'(x) - \Lambda_j'(\theta_j(x))\theta_j'(x) = \theta_j(x), \tag{A.11}$$

where the second equality follows from the definition of $\theta_j(x)$. From the above equation we also obtain $I_j''(x) = \theta_j'(x)$. Both $I_j'(x)$ and $I_j''(x)$ are well defined because $I_j(\cdot)$ is continuously differentiable for $x \in [\mu_2, \mu_1] \in \mathscr{H}_j^0$ (e.g., see Lemma 2.2.5 in Dembo and Zeitouni 2009). Applying a second-order Taylor expansion at $x = \mu_j$ with $I_j(\mu_j) = 0$, $I_j'(\mu_j) = 0$, and $I_j''(x) = \theta_j'(x)$, we have that

$$I_j(x) = \frac{(x - \mu_j)^2}{2}\theta_j'(\tilde{x}), \tag{A.12}$$

where $\tilde{x}$ is in the interval between x and μ_j. It can be easily seen that $\mathsf{E}[X_j^2] < \infty$ by Assumption 1, which in turn implies that $\theta_j'(\mu_j) = 1/\sigma_j^2$ by Lemma 7 of Chernoff (1952). Also, note that $I_j(\cdot) \in C^\infty$, and hence, so is $\theta_j(\cdot)$. Combining these observations with the fact that $\sigma_j \geq \sigma_{\min}$, (A.12) can be written as

$$I_j(x) = \frac{(x - \mu_j)^2}{2\sigma_j^2} + o\left((x - \mu_j)^2\right) \tag{A.13}$$

for each $j = 1, 2$, where $f(x) = o(g(x))$ implies $|f(x)|/|g(x)| \to 0$ as $x \to 0$.

Fix $\boldsymbol{\alpha} \in \Delta^{k-1}$ and recall that $\rho(\boldsymbol{\alpha}) = \inf_x\{\alpha_1 I_1(x) + \alpha_2 I_2(x)\}$. Define $I(x) \coloneqq \alpha_1 I_1(x) + \alpha_2 I_2(x)$ and

$$\begin{aligned}\bar{I}(x) &\coloneqq \alpha_1 \bar{I}_1(x) + \alpha_2 \bar{I}_2(x) \\ &= \alpha_1 \frac{(x-\mu_1)^2}{2\sigma_1^2} + \alpha_2 \frac{(x-\mu_2)^2}{2\sigma_2^2}. \end{aligned} \tag{A.14}$$

From (A.13) and the fact that $(x-\mu_j)^2 \le (\mu_1-\mu_2)^2$ for any $x \in [\mu_2, \mu_1]$, it can be easily seen that

$$I(x) = \bar{I}(x) + o\left((\mu_1-\mu_2)^2\right) \quad \text{as } \mu_2 \to \mu_1 \tag{A.15}$$

Let x^* be the minimizer of $I(x)$ and $\bar{x}$ be the minimizer of $\bar{I}(x)$. Using first-order conditions, we obtain that

$$\bar{x} = \left(\frac{\alpha_1/\sigma_1^2}{\alpha_1/\sigma_1^2 + \alpha_2/\sigma_2^2}\right)\mu_1 + \left(\frac{\alpha_2/\sigma_2^2}{\alpha_1/\sigma_1^2 + \alpha_2/\sigma_2^2}\right)\mu_2 \tag{A.16}$$

and

$$\bar{I}(\bar{x}) = \frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2/\alpha_1 + \sigma_2^2/\alpha_2)}. \tag{A.17}$$

Observe that $|I(x^*) - \bar{I}(\bar{x})| \le |I(x^*) - \bar{I}(x^*)| + |\bar{I}(x^*) - \bar{I}(\bar{x})|$ and by (A.15),

$$|I(x^*) - \bar{I}(x^*)| = o\left((\mu_1-\mu_2)^2\right). \tag{A.18}$$

Further,

$$\begin{aligned}|\bar{I}(x^*) - \bar{I}(\bar{x})| &= \left|\left(\alpha_1 \frac{\tilde{x}-\mu_1}{\sigma_1^2} + \alpha_2 \frac{\tilde{x}-\mu_2}{\sigma_2^2}\right)(x^* - \bar{x})\right| \\ &= o\left((\mu_1-\mu_2)^2\right) \quad \text{as } \mu_2 \to \mu_1 \end{aligned} \tag{A.19}$$

where the first equality follows from the first-order Taylor expansion of $\bar{I}(x^*)$ at $\bar{x}$ for some $\tilde{x}$ between x^* and $\bar{x}$; the second equality follows since $\sigma_j \ge \sigma_{\min}$ and $|(\tilde{x}-\mu_j)(x^*-\bar{x})| \le (\mu_1-\mu_2)^2$ for each $j = 1, 2$. Therefore, we establish that

$$\rho(\boldsymbol{\alpha}) = I(x^*) = \frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2/\alpha_1 + \sigma_2^2/\alpha_2)} + o\left((\mu_1-\mu_2)^2\right), \tag{A.20}$$

where the right-hand side of the preceding equality converges to $\rho^G(\boldsymbol{\alpha})$ as $\mu_1 - \mu_2 \to 0$. To show that $\rho(\boldsymbol{\alpha}^G)/\rho(\boldsymbol{\alpha}^*) \to 1$ as $\mu_1 - \mu_2 \to 0$, one can check that

$$|\rho(\boldsymbol{\alpha}^G) - \rho(\boldsymbol{\alpha}^*)| \leq \max\left(|\rho(\boldsymbol{\alpha}^G) - \rho^G(\boldsymbol{\alpha}^G)|, |\rho^G(\boldsymbol{\alpha}^*) - \rho(\boldsymbol{\alpha}^*)|\right) = o((\mu_1 - \mu_2)^2), \tag{A.21}$$

where the equality follows from (A.20). Further, due to the assumption that $\sigma_j \leq \sigma_{\max}$ for each $j = 1, 2$, we have that

$$\rho^G(\boldsymbol{\alpha}^G) \geq \rho^G(\boldsymbol{\alpha}^{\text{eq}}) \geq c'(\mu_1 - \mu_2)^2 \tag{A.22}$$

for some constant $c' > 0$ after some straightforward algebra. Combined with (A.20), this in turn implies $\rho(\boldsymbol{\alpha}^*) \geq c''(\mu_1 - \mu_2)^2$ for sufficient small $(\mu_1 - \mu_2)$. Hence, we establish the desired result by dividing both sides of (A.21) by $\rho(\boldsymbol{\alpha}^*)$.

Case 2. Consider a sequence of system configurations such that $\rho(\boldsymbol{\alpha}) \neq G_2(\boldsymbol{\alpha})$ for some $\delta > 0$. In this case, it suffices to consider the case where $\rho(\boldsymbol{\alpha}) = G_i(\boldsymbol{\alpha})$ for some $i \neq 1, 2$ and sufficiently small $\delta \in (0, \delta_0)$, or equivalently, $\arg\min_{j \neq 1}\{G_j(\boldsymbol{\alpha})\} \to i$ as $\delta \to 0$. (In cases where $\arg\min_{j \neq 1}\{G_j(\boldsymbol{\alpha})\}$ does not converge as $\delta \to 0$, one can consider subintervals of $(0, \delta_0)$ along which $\arg\min_{j \neq 1}\{G_j(\boldsymbol{\alpha})\}$ converges to $i' \neq 1, 2$ and then follow the same logical steps as below, which will be omitted in this proof.)

We first show that $\mu_1 - \mu_i \to 0$ as $\delta \to 0$. Observe that $G_i(\boldsymbol{\alpha}) \leq G_2(\boldsymbol{\alpha})$ for sufficiently small δ since $\rho(\boldsymbol{\alpha}) = G_i(\boldsymbol{\alpha})$ for $\delta \in (0, \delta_0)$. From the definition of $G_j(\cdot)$ in (8) and the fact that $I_j(x)$ is strictly concave with $I_j(\mu_j) = 0$, it can be easily seen that $G_2(\boldsymbol{\alpha}) \to 0$ as $\delta \to 0$, and therefore, we establish that $G_i(\boldsymbol{\alpha}) \to 0$ as $\delta \to 0$. Now, towards a contradiction, suppose that $\liminf_{\delta \to 0}(\mu_1 - \mu_i) \geq d$ for some constant $d > 0$. Recall that $G_i(\boldsymbol{\alpha}) = \inf_x\{\alpha_1 I_1(x) + \alpha_i I_i(x)\}$ and let x_i^* be the minimizer which lies in $[\mu_i, \mu_1]$. Note that $G_i(\boldsymbol{\alpha}) \to 0$ implies both $I_1(x_i^*)$ and $I_i(x_i^*)$ converges to zero. However, due to (A.13) and the assumption that $\sigma_j \leq \sigma_{\max}$ for each j, it can be seen that $I_1(x_i^*)$ and $I_i(x_i^*)$ can converge to zero only when $x_i^* - \mu_1 \to 0$ and $x_i^* - \mu_i \to 0$ as $\delta \to 0$, respectively. This is a contradiction due to the assumption that $\liminf_{\delta \to 0}(\mu_1 - \mu_i) \geq d$. Therefore, we have that $\mu_1 - \mu_i \to 0$ as $\delta \to 0$. Now, the rest of the proof can be done by following the same logical steps as the proof for *Case 1*, with variables for system 2 being replaced with those for system i; the remaining part will be omitted. □

Proof of Theorem 3. In this proof we fix $\boldsymbol{\pi} = \boldsymbol{\pi}^{\boldsymbol{\alpha}}$, where for clarity we suppress the function arguments. Note that the inequalities in (A.1) imply, on a logarithmic scale, the rate function of $\mathsf{P}_t(\mathrm{FS}_t)$ can be immediately obtained once we find the rate function of $\mathsf{P}_t(\bar{X}_{1t} < \bar{X}_{jt})$ for each j. From this point and on, we consider $k = 2$ for simplicity and without loss of generality the proof easily extends to $k \geq 3$.

Fix $\boldsymbol{\alpha} = (\alpha_1, \alpha_2) \in \Delta^1$ and observe that

$$\begin{aligned}
\mathsf{P}_t(\mathrm{FS}_t) &= \mathsf{P}_t(\bar{X}_{1t} < \bar{X}_{2t}) \\
&= \mathsf{P}_t(\bar{X}_{1t} - \mu_1^t < \bar{X}_{2t} - \mu_2^t - (\mu_1^t - \mu_2^t)) \\
&\overset{(a)}{=} \mathsf{P}(\bar{X}_{1t} - \mu_1 < \bar{X}_{2t} - \mu_2 - (\mu_1^t - \mu_2^t)) \\
&\overset{(b)}{=} \mathsf{P}(\bar{X}_{1t} - \mu_1 + \delta_t(\mu_1 - \mu_2)/2 < \bar{X}_{2t} - \mu_2 - \delta_t(\mu_1 - \mu_2)/2), \qquad \text{(A.23)}
\end{aligned}$$

where (a) follows from the change of measure from $\mathbf{F}^t$ to $\mathbf{F}$; specifically, the distribution of $\bar{X}_{jt} - \mu_j^t$ under the the probability measure $\mathsf{P}_t(\cdot)$ is identical to that of $\bar{X}_{jt} - \mu_j$ under the probability measure $\mathsf{P}(\cdot)$. Also, (b) follows from the fact that $(\mu_1^t - \mu_j^t) = \delta_t(\mu_1 - \mu_j)$. Define $W_{1t} = (\bar{X}_{1t} - \mu_1)/\delta_t + (\mu_1 - \mu_2)/2$ and $W_{2t} = (\bar{X}_{2t} - \mu_2)/\delta_t - (\mu_1 - \mu_2)/2$. Then, $\mathsf{P}_t(\mathrm{FS}_t)$ can be written as

$$\mathsf{P}_t(\mathrm{FS}_t) = \mathsf{P}(W_{1t} < W_{2t}). \qquad \text{(A.24)}$$

Set $W_t = (W_{1t}, W_{2t})$. Denote the logarithmic moment generating function of W_t by $\Lambda_t(\lambda_1, \lambda_2) = \log \mathsf{E}[e^{\lambda_1 W_{1t} + \lambda_2 W_{2t}}]$ for $(\lambda_1, \lambda_2) \in \mathscr{R}^2$. By the Gartner-Ellis Theorem (cf. See Theorem 2.3.6 in Dembo and Zeitouni 2009), if the limit

$$\Lambda(\lambda_1, \lambda_2) \coloneqq \lim_{t \to \infty} \frac{1}{t\delta_t^2} \Lambda_t(t\delta_t^2 \lambda_1, t\delta_t^2 \lambda_2) \qquad \text{(A.25)}$$

exists, then

$$\frac{1}{t\delta_t^2} \log \mathsf{P}_t(\mathrm{FS}_t) \to - \inf_{x_1 \leq x_2} \{I(x_1, x_2)\} \text{ as } t \to \infty, \qquad \text{(A.26)}$$

where

$$I(x_1, x_2) = \sup_{\lambda_1, \lambda_2} \{\lambda_1 x_1 + \lambda_2 x_2 - \Lambda(\lambda_1, \lambda_2)\}. \qquad \text{(A.27)}$$

We first verify that (A.25) holds. Observe that $\Lambda_t(t\delta_t^2\lambda_1, t\delta_t^2\lambda_2) = \log \mathsf{E}[e^{t\delta_t^2\lambda_1 W_{1t}}] + \log \mathsf{E}[e^{t\delta_t^2\lambda_2 W_{2t}}]$ by the independence of W_{1t} and W_{2t}. Then, one can check that

$$\begin{aligned}\frac{1}{t\delta_t^2}\log \mathsf{E}[e^{t\delta_t^2\lambda_1 W_{1t}}] &= \frac{1}{t\delta_t^2}\alpha_j t \log \mathsf{E}[e^{\lambda_1\delta_t(X_1-\mu_1)/\alpha_1}] + \frac{\lambda_1(\mu_1-\mu_2)}{2} \\ &\to \frac{\lambda_1^2\sigma_1^2}{2\alpha_1} + \frac{\lambda_1(\mu_1-\mu_2)}{2} \quad \text{as } t\to\infty, \end{aligned} \tag{A.28}$$

where the equality follows because the sequence, $X_{11}, X_{12}, \ldots$, is independent and identically distributed and the limit follows from the second-order Taylor expansion of the log-moment generating function. For simplicity of exposition we ignore non-integrality of $\alpha_j t$. By the same procedure one can also obtain that

$$\frac{1}{t\delta_t^2}\log \mathsf{E}[e^{t\delta_t^2\lambda_2 W_{2t}}] \to \frac{\lambda_2^2\sigma_2^2}{2\alpha_2} - \frac{\lambda_2(\mu_1-\mu_2)}{2} \quad \text{as } t\to\infty \tag{A.29}$$

Therefore, the limit in (A.25) exists and (A.26) holds by the Gartner-Ellis Theorem. To evaluate $I(x_1, x_2)$, observe that

$$\begin{aligned} I(x_1,x_2) &= \sup_{\lambda_1,\lambda_2}\left(\lambda_1 x_1 + \lambda_2 x_2 - \frac{\lambda_1^2\sigma_1^2}{2\alpha_1} - \frac{\lambda_2^2\sigma_2^2}{2\alpha_2} - \frac{\lambda_1(\mu_1-\mu_2)}{2} + \frac{\lambda_2(\mu_1-\mu_2)}{2}\right) \\ &= \sup_{\lambda_1}\left(\lambda_1 x_1 - \frac{\lambda_1^2\sigma_1^2}{2\alpha_1} - \frac{\lambda_1(\mu_1-\mu_2)}{2}\right) + \sup_{\lambda_2}\left(\lambda_2 x_2 - \frac{\lambda_2^2\sigma_2^2}{2\alpha_2} + \frac{\lambda_2(\mu_1-\mu_2)}{2}\right) \\ &= \alpha_1\frac{(x_1-(\mu_1-\mu_2)/2)^2}{2\sigma_1^2} + \alpha_2\frac{(x_2+(\mu_1-\mu_2)/2)^2}{2\sigma_2^2}. \end{aligned} \tag{A.30}$$

It can be easily seen that the infimum of $I(x_1, x_2)$ is achieved in the range, $-(\mu_1-\mu_2)/2 \le x_1 \le x_2 \le (\mu_1-\mu_2)/2$. Further, in this region, the first and second terms on the right-hand side of (A.30) are decreasing and increasing, respectively. Hence, using first-order conditions, one can show that the infimum of $I(x_1, x_2)$ is achieved at $x_1 = x_2 = x$ with

$$x = \frac{\alpha_1(\mu_1-\mu_2)/2\sigma_1^2 - \alpha_2(\mu_1-\mu_2)/2\sigma_2^2}{\alpha_1/\sigma_1^2 + \alpha_2/\sigma_2^2}. \tag{A.31}$$

Hence, from (A.26),

$$\frac{1}{t\delta_t^2}\log \mathsf{P}_t(\mathrm{FS}_t) \to -\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2/\alpha_1 + \sigma_2^2/\alpha_2)} \text{ as } t\to\infty. \tag{A.32}$$

This completes the proof. □

B. Proofs for Auxiliary Lemmas

Proof of Lemma A.1. In the proof we fix $\boldsymbol{\pi} = \boldsymbol{\pi}^{\boldsymbol{\alpha}}$, where for clarity we suppress $\boldsymbol{\pi}^{\boldsymbol{\alpha}}$ in the function arguments. First, it is easily checked that $N_{jt}/t \to \alpha_j$ as $t \to \infty$ for each $j = 1, \ldots, k$ by the definition of the static allocation rule. Observe that

$$\frac{1}{t}\Lambda_t(t\lambda_1, t\lambda_j) = \frac{N_{1t}}{t}\Lambda_1\left(\frac{\lambda_1}{N_{1t}/t}\right) + \frac{N_{jt}}{t}\Lambda_j\left(\frac{\lambda_j}{N_{jt}/t}\right) \tag{B.1}$$

for $t \geq k$ and $j = 2, \ldots, k$. Since $N_{jt}/t \to \alpha_j$, we have that

$$\frac{1}{t}\Lambda_t(t\lambda_1, t\lambda_j) \to \alpha_1\Lambda_1(\lambda_1/\alpha_1) + \alpha_j\Lambda_j(\lambda_j/\alpha_j) \text{ as } t \to \infty \tag{B.2}$$

for $j = 2, \ldots, k$. Therefore, by Gartner-Ellis Theorem (cf. See Theorem 2.3.6 in Dembo and Zeitouni 2009), $I_{1j}(x_1, x_j)$ equals

$$\sup_{\lambda_1, \lambda_j} \{\lambda_1 x_1 + \lambda_j x_j - \alpha_1\Lambda_1(\lambda_1/\alpha_1) - \alpha_j\Lambda_j(\lambda_j/\alpha_j)\} \tag{B.3}$$

$$= \sup_{\lambda_1}\{\lambda_1 x_1 - \alpha_1\Lambda_1(\lambda_1/\alpha_1)\} + \sup_{\lambda_j}\{\lambda_j x_j - \alpha_j\Lambda_j(\lambda_j/\alpha_j)\} \tag{B.4}$$

$$= \alpha_1 \sup_{\lambda_1/\alpha_1} \{(\lambda_1/\alpha_1)x_1 - \Lambda_1(\lambda_1/\alpha_1)\} + \alpha_j \sup_{\lambda_j/\alpha_j} \{(\lambda_j/\alpha_j)x_j - \Lambda_j(\lambda_j/\alpha_j)\}. \tag{B.5}$$

Recalling the definition of $I_j(\cdot)$ in (6), the proof for this lemma is complete. □

Proof of Lemma A.2. First, we show that $\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{y})$ is strictly positive. Note that the objective function of the optimization problem (A.8) is strictly positive at its optimal solution; to see this, notice that $\boldsymbol{\alpha}_e = (1/k, \ldots, 1/k)$ is a feasible solution and the objective function is positive at $\boldsymbol{\alpha}_e$. Suppose that $\alpha_j(\boldsymbol{x}, \boldsymbol{y}) = 0$ for some j. Then the objective function of the optimization problem (A.8) is 0, which contradicts the optimality of $\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{y})$.

Without loss of generality, assume that $x_1 \geq \cdots \geq x_k$ and let $b = |B(\boldsymbol{x})| < k$. Define $\epsilon \in (0, (x_1 - x_{b+1})/2)$ and let $t_0 < \infty$ such that $\max_j\{|x_j^t - x_j|\} \leq \epsilon$ and $\max_j\{|y_j^t - y_j|\} \leq \epsilon$ for all $t \geq t_0$. That is, for each $t \geq t_0$, $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is in a ball with radius ϵ, centered at $(\boldsymbol{x}, \boldsymbol{y})$. Also, note that $B(\boldsymbol{x}^t) = B(\boldsymbol{x})$ for $t \geq t_0$. In the rest of the proof, it suffices to consider $t \geq t_0$.

Suppose, towards a contradiction, that $\boldsymbol{\alpha}(\boldsymbol{x}^t, \boldsymbol{y}^t)$ does not converge to $\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{y})$. Since $\boldsymbol{\alpha}(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is a bounded sequence in Δ^{k-1}, by the Bolzano-Weierstrass theorem it has a convergent subsequence, $\{t_1, t_2, \ldots\}$, such that

$$\boldsymbol{\alpha}(\boldsymbol{x}^{t_n}, \boldsymbol{y}^{t_n}) \to \tilde{\boldsymbol{\alpha}} \neq \boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{y}) \quad \text{as } n \to \infty. \tag{B.6}$$

Since $\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{y})$ is the unique maximizer of $H(\boldsymbol{\alpha}; \boldsymbol{x}, \boldsymbol{y})$ and $\tilde{\boldsymbol{\alpha}} \neq \boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{y})$, it can be seen that

$$H(\tilde{\boldsymbol{\alpha}}; \boldsymbol{x}, \boldsymbol{y}) < H(\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{y}); \boldsymbol{x}, \boldsymbol{y}). \tag{B.7}$$

On the other hand, since $\boldsymbol{\alpha}(\boldsymbol{x}^{t_n}, \boldsymbol{y}^{t_n})$ is the unique maximizer of $H(\boldsymbol{\alpha}; \boldsymbol{x}^{t_n}, \boldsymbol{y}^{t_n})$,

$$H(\boldsymbol{\alpha}(\boldsymbol{x}^{t_n}, \boldsymbol{y}^{t_n}); \boldsymbol{x}^{t_n}, \boldsymbol{y}^{t_n}) \geq H(\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{y}); \boldsymbol{x}^{t_n}, \boldsymbol{y}^{t_n}). \tag{B.8}$$

Note that $H(\boldsymbol{\alpha}; \boldsymbol{x}, \boldsymbol{y})$ is continuous in $\boldsymbol{\alpha}$ and $(\boldsymbol{x}, \boldsymbol{y})$. Since $\boldsymbol{\alpha}(\boldsymbol{x}^{t_n}, \boldsymbol{y}^{t_n}) \to \tilde{\boldsymbol{\alpha}}$, $\boldsymbol{x}^{t_n} \to \boldsymbol{x}$, and $\boldsymbol{y}^{t_n} \to \boldsymbol{y}$, taking $n \to \infty$ on both sides of (B.8), we obtain that

$$H(\tilde{\boldsymbol{\alpha}}; \boldsymbol{x}, \boldsymbol{y}) \geq H(\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{y}); \boldsymbol{x}, \boldsymbol{y}), \tag{B.9}$$

which contradicts (B.7). Therefore, $\boldsymbol{\alpha}(\boldsymbol{x}^t, \boldsymbol{y}^t) \to \boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{y})$ almost surely and the proof is complete. □

C. Effect of the Number of Initial Samples

While the qualitative conclusions in §7 seem reasonably robust relative to the choice of n_0, the number of initial samples, it still leaves open the question of how to determine n_0. In this section we show via numerical examples the sensitivity of performance in terms of the probability of false selection for different values of n_0. In light of this, we consider two configurations with $k = 50$ normally distributed systems: The monotone decreasing means (MDM) in which $\mu_j = 1.05 - 0.05j$ for $j = 1, \ldots, k$ and the SC configuration in which $\mu_1 = 1$ and $\mu_j = 0.8$ for $j = 2, \ldots, k$. For both cases, we let $T = 2 \cdot 10^4$ and standard deviations are equally set to one.

Figure 8 illustrates the probability of false selection for different values of $n_0 = 50, 100, 200$. When the majority of systems are significantly inferior to the best system, it is desirable to set n_0 to a low

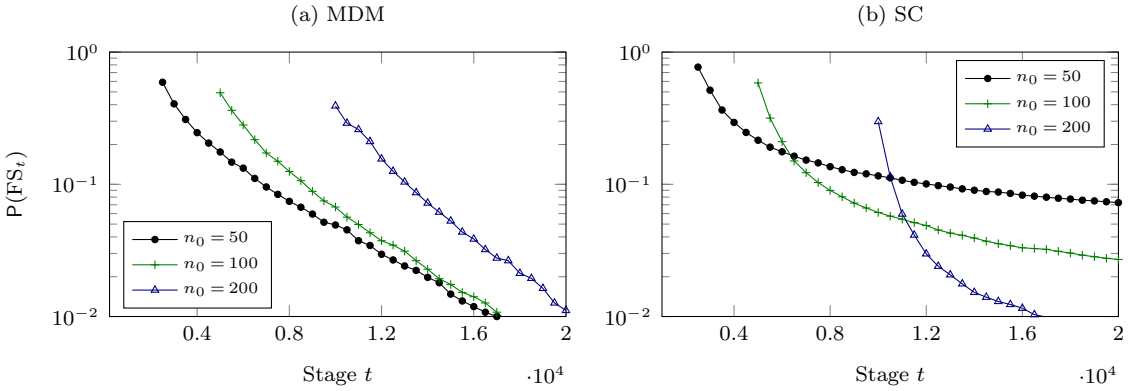

Figure 8 **Probability of false selection under the AWD policy for different values of n_0.** $\mathsf{P}(\mathrm{FS}_t)$ is plotted as a function of stage t in log-linear scale. The example considers the normal distribution case. On the left panel, $\mu_j = 1.05 - 0.05j$ for $j = 1, \ldots, k$, while on the right panel, $\mu_1 = 1$ and $\mu_j = 0.8$ for all $j \neq 1$. For both cases, $k = 50$ and standard deviations are equally set to one.

number. In the MDM configuration, only the first 20 systems have means that are within a unit distance from system 1, in terms of the number of standard deviation ($\sigma_1 = 1$). In this case, most systems may turn out to be inferior in a very early stage of the sampling horizon. So, there can be significant inefficiencies when n_0 is set too high because a significant portion of the sampling budget is wasted in estimating means of seemingly inferior systems. Observe from the left panels of Figure 8 that the probability of false selection is higher for larger values of n_0 in a range from 50 to 200.

Conversely, if non-best systems are "equally worse" than the best system, it may be desirable to set n_0 to a high number. As seen in the case of the SC configuration (right panels of Figure 8), the probability of false selection is higher for smaller values of n_0 in a range from 50 to 200. Of course, an optimal value of n_0 cannot be determined exactly because the configuration of means, as well as variances, are unknown a priori. Nevertheless, the preceding arguments can provide a qualitative insight for a "good choice" of n_0.