Adaptive Search Algorithms for Discrete Stochastic Optimization: A Smooth Best-Response Approach

Omid Namvar Gharehshiran, Vikram Krishnamurthy, *Fellow, IEEE*, and George Yin, *Fellow, IEEE*

***Abstract*—This paper considers simulation-based optimization of the performance of a regime-switching stochastic system over a finite set of feasible configurations. Inspired by the stochastic fictitious play learning rules in game theory, we propose an adaptive random search algorithm that uses a smooth best-response sampling strategy and tracks the set of global optima, yet distributes the search so that most of the effort is spent on simulating the system performance at the global optima. The algorithm responds properly to the random unpredictable jumps of the global optimum even when the observations data are temporally correlated as long as a weak law of large numbers holds. Numerical examples show that the proposed scheme yields faster convergence and superior efficiency for finite sample lengths compared with several existing random search and pure exploration methods in the literature.**

***Index Terms*—Discrete stochastic optimization, Markov chain, randomized search, simulation-based optimization, stochastic approximation, time-varying optima.**

I. Introduction

DISCRETE stochastic optimization problems arise in operations research [1], [2], manufacturing engineering [3], and communication networks [4], [5]. These problems are intrinsically more difficult to solve than their deterministic counterparts due to the unavailability of an explicit relation between the objective function and the underlying decision variables. It is therefore necessary to use stochastic simulation to estimate the objective function in such problems.

A. Problem

The simplest setting of a discrete stochastic optimization problem is as follows:

$$\min_{\alpha\in\mathcal{S}} \mathbf{e}_{\alpha}\overline{\mathbf{X}} \text{ where } \overline{\mathbf{X}}:=\mathbb{E}\{\mathbf{X}_n\},\ \mathbf{X}_n=\mathrm{col}\,(X_n(1),\dots,X_n(S)). \tag{1}$$

Here, the search space $\mathcal{S}=\{1,2,\dots,S\}$ is finite, α is the decision variable, and $\mathbf{e}_{\alpha}$ denotes the unit row vector in S-dimensional Euclidean space with the αth element being equal to one. Further, $X_n(\alpha)$ is a bounded real-valued stationary stochastic process that represents the sequence of observations of the objective function value at candidate solution α. Note that the distribution of $X_n(\alpha)$ may depend on the decision variable α; however, for each α, $X_n(\alpha)$ is a stationary process.

In discrete-event systems of practical interest, the goal is often to optimize the expected performance over a set of feasible configurations. However, due to the unexplained randomness and complexity involved, there typically exists no explicit relation between the performance measure ofinterest and the chosen configuration. In such cases, one can only observe "samples" of the system performance through a simulation process [6]. Each stochastic process $X_n(\alpha)$ in (1) can be interpreted as the sequence of such samples at a particular configuration α of the stochastic system under study. In practice, $X_n(\alpha)$ has finite variance but unknown distribution; therefore, the expectation in (1) cannot be evaluated analytically. The problem of finding the global optimum can then be seen as getting stochastic simulation and optimization effectively combined.

It is normally assumed that for each $\alpha\in\mathcal{S}$, the sample mean of $\widehat{X}_N(\alpha)$ is a strongly consistent estimator. That is, $\overline{X}(\alpha)=\lim_{N\to\infty}\widehat{X}_N(\alpha)$ a.s., where

$$\widehat{X}_N(\alpha):=\frac{1}{N}\sum_{n=1}^{N}X_n(\alpha). \tag{2}$$

The above assumption is satisfied by most simulation outputs [7]. It becomes the Kolmogorov's strong law of large numbers if $\{X_n(\alpha)\}$ is independent and identically distributed (i.i.d.), and the ergodicity if $\{X_n(\alpha)\}$ is stationary ergodic. A brute force method of solving (1) involves an exhaustive enumeration: For each $\alpha\in\mathcal{S}$, compute $\widehat{X}_N(\alpha)$ in (2) via simulation for large N. Then, pick $\alpha^*=\arg\min_{\alpha\in\mathcal{S}}\widehat{X}_N(\alpha)$. This is highly inefficient since one has to perform a large number of simulations at each feasible alternative; however, those performed on nonpromising feasible solutions do not contribute to finding the global optimum and are wasted. The main idea is thus to develop a search scheme that is both *attracted* to the global optima set and *efficient*, in the sense that it spends less time simulating the nonpromising alternatives [8, Ch. 5].

This paper considers two extensions of the above problem: First, we allow for $X_n(\alpha)$, $n=1,2,\dots$, to be temporally correlated for each α as long as it satisfies a *weak* law of large numbers. Second, we focus on a nonstationary variant where the global optimum undergoes random unpredictable jumps due to random evolution of the profile of the stochastic events or the objective function (or both). Such problems arise in a broad range of practical applications where the goal is to track the optimal operating configuration of a stochastic system subject

Manuscript received February 13, 2014; revised November 24, 2014, December 3, 2015, and February 22, 2016; accepted March 1, 2016. Date of publication March 8, 2016; date of current version December 26, 2016. The work of V. Krishnamurthy was supported by a NSERC Strategic grant and the Canada Research Chairs program. The work of G. Yin was supported in part by the Army Research Office under grant W911NF-15-1-0218. Recommended by Associate Editor A. Nedich.

O. Namvar Gharehshiran is with the Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada (e-mail: omidn@utstat.toronto.edu).

V. Krishnamurthy is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: vikramk@ece.ubc.ca).

G. Yin is with the Department of Mathematics, Wayne State University, Detroit, MI 48202 USA (e-mail: gyin@math.wayne.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2016.2539225

to time inhomogeneity. More precisely, we consider discrete stochastic optimization problems of the form

$$\min_{\alpha\in\mathcal{S}} \mathbb{E}\left\{\mathbf{e}_\alpha \mathbf{X}_n(\theta_n)\right\} = \min_{\alpha\in\mathcal{S}} \sum_{i=1}^{\Theta} \mathbb{E}\left\{\mathbf{e}_\alpha \mathbf{X}_n(i) I(\theta_n = i)\right\}$$
$$\text{where } \mathbf{X}_n(\theta_n) = \operatorname{col}\left(X_n(1,\theta_n), \ldots, X_n(S,\theta_n)\right). \quad (3)$$

Here, $\{\theta_n\}$ is an integer-valued process representing the random changes in the state of the stochastic system, and is unobservable with unknown dynamics. Note that each stochastic process $X_n(\alpha,\theta_n)$ in (3) is no longer stationary even though $X_n(\alpha,i)$ is stationary for each i. Denote the state space of θ_n by $\mathcal{M} = \{1,\ldots,\Theta\}$. Intuitively, in lieu of one sequence, we have Θ sequences $\{X_n(\alpha,i) : i \in \mathcal{M}\}$ for each feasible solution α. One may imagine that there is a random environment that is dictated by the process θ_n. If at time n, θ_n takes a value, say, i, then $X_n(\alpha,i)$ will be the sequence of data output by simulation. Equation (3) faithfully reflects this fact.

B. Main Results

Inspired by fictitious play learning rules in game theory [9], we propose a class of sampling-based adaptive search algorithms, which proceeds as follows: At each iteration, a sample is taken from the search space according to a randomized strategy (a probability distribution on the search space) that minimizes some perturbed variant of the expected objective function based on current beliefs. This randomized strategy is referred to as *smooth best-response sampling strategy*. The perturbation term in fact simulates the search or exploration functionality essential in learning the expected stochastic behavior. The system performance is then simulated at the sampled solution, and fed into a stochastic approximation algorithm to update beliefs and, hence, the sampling strategy.

The convergence analysis shows that, if θ_n evolves on the same timescale as the adaptive search algorithm, the proposed algorithm can properly track the global optimum as it undergoes random unpredicted jumps by showing that a *regret* measure can be made and kept arbitrarily small. The regret is defined as the *opportunity loss*, and compares the performance of a course of feasible solutions sampled by the adaptive search algorithm to the performance of the global optimum. Further, if θ_n evolves on a slower timescale, the most frequently sampled feasible solution tracks the global optima set as it undergoes random unpredictable jumps over time. This in turn implies that the proposed scheme exhausts most of its simulation budget on the global optimum. This is desirable since, in many practical applications, the system has to be operated in the sampled configuration to measure performance. The proposed algorithm assumes no functional properties such as submodularity, symmetry, or exchangeability on the objective function. It can as well be deployed in static discrete stochastic optimization problems, i.e., when θ_n is fixed.

The main features of this work are:

1) *Correlated simulation data*: It allows for temporal correlation in the collected data via simulation, that is more realistic, whereas most existing schemes assume that the simulation data are i.i.d.
2) *Adaptive search*: The proposed algorithm tracks the random unpredictable jumps of the global optimum. This is in contrast to most existing algorithms that are designed to locate static optimum.
3) *Matched timescale*: It is well known in the literature of stochastic approximation schemes that, if θ_n changes too drastically, there is no chance one can track the time-varying optima. (Such a phenomenon is known as trackability; see [10] for related discussions.) On the other hand, if θ_n evolves on a slower timescale as compared to the updates of the adaptive search algorithm, it can be approximated by a constant and its variation is ignored. In this work, we consider the nontrivial case where θ_n evolves on the *same* timescale as the adaptive search algorithm, and prove that the proposed scheme properly responds to the jumps that the global optimum undergoes.

The tracking analysis proceeds as follows: First, by a combined use of weak convergence methods [11] and treatment on Markov switched systems [12], [13], we show that the limit system for the discrete-time iterates of the proposed algorithm is a switched ordinary differential equations (ODE). (This is in contrast to the standard treatment of stochastic approximation algorithms, where the limiting dynamics converge to a deterministic ODE.) By using multiple Lyapunov function methods for randomly switched systems [14], [15], the stability analysis then shows that the limit dynamical system is asymptotically stable almost surely. This in turn establishes that the limit points of the switched ODE and the discrete-time iterates of the algorithm coincide. Finally, we conclude the tracking and efficiency results by characterizing the global attractor set of the derived limit system.

C. Literature

This work is closely connected to the literature on random search methods; see [16] for a discussion. Some random search methods spend significant effort to simulate each newly visited state at the initial stages to obtain an estimate of the objective function. Then, deploying a deterministic optimization mechanism, they search for the global optimum; see [17]–[20]. Another class, namely, discrete stochastic approximation methods [11], [21], distributes the simulation effort through time, and proceeds cautiously based on the limited information available at each time. Algorithms from this class primarily differ in the choice of the sampling strategy, which can be classified as: i) *point-based*, such as simulated annealing [22], [23], tabu search [24], stochastic ruler [25], stochastic comparison and descent algorithms [26]–[29], ii) *set-based*, such as branch-and-bound [30], nested partitions [31], stochastic comparison and descent algorithms [7], iii) *population-based*, such as genetic algorithms. The above methods are categorized under *instance-based* random search schemes since they generate new candidate solutions using solely the current (population of) solution. The adaptive search algorithm in this paper is related to another category, namely, *model-based* random search schemes [32], in which new solutions are generated via an intermediate probabilistic model that is updated from the previously seen solutions in such a way that the search will concentrate on optima. Examples of such schemes include the cross-entropy method [33], [34], ant colony optimization [35], [36], model reference adaptive search [37], gradient-based adaptive stochastic search [38], and estimation of distribution methods [39], [40]. Another related body of research pertains to the multiarmed bandit problem [41], which is concerned with optimizing the cumulative objective function values realized over a period of time, and the pure exploration problem [42], which involves finding the best arm after a given number of arm pulls.

All the above works assume static problems—fixed arms' reward distributions in the case of multiarmed bandit and pure exploration schemes—with i.i.d. simulation data/sampled payoffs.[1] In contrast, this paper addresses nonstationary discrete stochastic optimization problems with correlated data. The proposed scheme further differs from the above works in both the randomized sampling method as well as the belief updating mechanism. They use decreasing step-sizes, which prevents tracking time variations, however allows them to achieve almost sure convergence results (in contrast to weak convergence results in this paper). Finally, numerical studies in Section VII reveal that, in contrast to the proposed scheme, bandit-based algorithms such as upper confidence bound (UCB) [41], [43] exhibit reasonable efficiency only when the size of the search space is relatively small.

This work also connects well with earlier efforts [44], [45], which view optimizing the performance of a system comprising several decision variables as a noncooperative game with identical payoff functions, and propose heuristics based on fictitious play to reach the Nash equilibrium. These schemes, however, ensure achieving the optimal game outcome (in a distributed fashion), which could potentially differ from the optimal outcome for individual players in isolation as targeted in this paper. They further assume full knowledge of the players' payoff functions. This is in contrast to the setting in this paper which only requires the stream of realized payoffs.

D. Organization

The rest of the paper is organized as follows. Section II formalizes the main assumption posed on the problem. The proposed adaptive search scheme is then presented in Section III followed by the theorem entailing the tracking and efficiency properties in Section IV. Section V gives the proof of the main theorem. Subsequently, Section VI analyzes the proposed scheme under slow random time variations. Finally, numerical examples are provided in Section VII before the concluding remarks in Section VIII. The proofs are relegated to the Appendix for clarity of presentation.

II. Main Assumption

This section formalizes the main assumptions posed on the sequence of data collected via simulation. Denote by $\mathbb{E}_\ell$ the conditional expectation given $\mathcal{F}_\ell$, the σ-algebra generated by $\{X_n(\alpha, i), \theta_n : i \in \mathcal{M}, \alpha \in \mathcal{S}, n < \ell\}$. We make the following assumption on the sequence of simulation data.

Assumption 1: For each $\alpha \in \mathcal{S}$ and $i \in \mathcal{M}$,

(1) $\{X_n(\alpha, i)\}$ is a bounded, stationary, and real-valued sequence of random variables;

(2) for any $\ell \geq 0$, there exists $\overline{X}(\alpha, i)$ such that

$$\frac{1}{N}\sum_{n=\ell}^{N+\ell-1} \mathbb{E}_\ell\{X_n(\alpha, i)\} \to \overline{X}(\alpha, i) \text{ in probability as } N \to \infty.$$

The above condition allows us to work with correlated sequences of simulation data (on a particular feasible solution) whose remote past and distant future are asymptotically independent. For simplicity, we state the assumption in terms of the sequence $\{X_n(\alpha, i)\}$. The result follows from the mixing condition with appropriate mixing rates. Examples include sequences of i.i.d. random variables with bounded variance, martingale difference sequences with finite second moments, moving average processes driving by a martingale difference sequence, mixing sequences in which remote past and distant future are asymptotically independent, certain nonstationary sequences such as a function of a Markov chain, etc.; see, e.g., [7], [46] for further discussions.

The following assumption is made on the random changes underlying the discrete stochastic optimization problem (3).

Assumption 2: The sequence $\{\theta_n\}$ is unobservable and its dynamics are unknown.

The above assumption implies that the explicit dependence of $X_n(\alpha, \theta_n)$ on θ_n is unknown. This allows us to simplify the notation to $X_n(\alpha)$ with all the time variations captured by the subscript n. We need to emphasize that, although this notation is similar to the one used in the static problem in (1), the main difference is that the stochastic process $\{X_n(\alpha)\}$ is now *nonstationary*.

III. Smooth Best-Response Adaptive Search

A random search algorithm repeatedly takes samples from the search space according to a randomized sampling strategy (a probability distribution over the search space). These samples are then evaluated via real-time simulation experiments, using which the estimate of the global optimum is revised. The adaptive random search algorithm proposed in this section can be simply described as an adaptive sampling scheme. The appeal of sampling-based methods is because they often approximate well, with a relatively small number of samples, problems with a large number of scenarios; see [47] and [48] for numerical reports. Define by

$$\Delta\mathcal{S} = \left\{\mathbf{p} \in \mathbb{R}^S; p_i \geq 0, \sum_{i \in \mathcal{S}} p_i = 1\right\}. \tag{4}$$

the simplex of all probability distributions on the search space $\mathcal{S}$. The sampling strategy that we propose, namely, *smooth best-response sampling strategy* $\boldsymbol{b}^\gamma(\cdot)$, is inspired by learning algorithms in games [9], [49], and is formally defined below.

Definition 3.1: Choose a perturbation function

$$\rho(\boldsymbol{\sigma}) : \text{int}(\Delta\mathcal{S}) \to \mathbb{R}$$

where $\text{int}(\Delta\mathcal{S})$ represents the interior of the simplex $\Delta\mathcal{S}$, defined in (4), such that:

1) $\rho(\cdot)$ is $\mathcal{C}^1$ (i.e., continuously differentiable), strictly concave, and $|\rho| \leq 1$;

2) $\|\nabla\rho(\boldsymbol{\sigma})\| \to \infty$ as $\boldsymbol{\sigma}$ approaches the boundary of $\Delta\mathcal{S}$, i.e.,

$$\lim_{\boldsymbol{\sigma} \to \partial(\Delta\mathcal{S})} \|\nabla\rho(\boldsymbol{\sigma})\| = \infty$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\partial(\Delta\mathcal{S})$ represents the boundary of simplex $\Delta\mathcal{S}$.

Given any vector $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_S) \in \mathbb{R}^S$ of beliefs about objective values at different candidate solutions, the *smooth best-response sampling* strategy is then given by

$$\boldsymbol{b}^\gamma(\boldsymbol{\psi}) := \operatorname*{arg\,min}_{\boldsymbol{\sigma} \in \Delta\mathcal{S}} (\boldsymbol{\sigma} \cdot \boldsymbol{\psi} - \gamma\rho(\boldsymbol{\sigma})), \quad 0 < \gamma < \Delta\widetilde{D} \tag{7}$$

where γ determines the exploration weight, and $\Delta\widetilde{D}$ denotes the greatest difference in objective function values among the feasible solutions.

Let $I(X)$ denote the indicator function: $I(X) = 1$ if the statement X is true, and $I(X) = 0$ otherwise. Let further $\mathbf{0}_S$ denote a vector of zeros of size S. The adaptive random search scheme can then be summarized as in Algorithm 1. Based on the belief $\boldsymbol{\psi}_n$, Algorithm 1 prescribes to take a sample s_n from the search space according to the sampling strategy $\boldsymbol{b}^\gamma(\boldsymbol{\psi}_n)$,

[1] See [43] for upper confidence bound policies for nonstationary bandit problems.

given in Definition 3.1. A simulation experiment is then performed to evaluate the sample and obtain $X_n(s_n)$. (Recall that all time variations underlying (3) are captured in the subscript n since θ_n, which drives the time variations, is unobservable.) The simulation output is then fed into a stochastic approximation algorithm to update the belief and, hence, the sampling strategy. This is in contrast to most existing algorithms that use stationary sampling strategies. The proposed algorithm relies only on the simulation data and requires minimal computational resources per iteration: It needs only one simulation experiment per iteration as compared to two in [26] or S (size of the search space) in [8, Ch. 5.3]. Yet, as evidenced by the numerical evaluations in Section VII, it guarantees performance gains in terms of the tracking speed.

Algorithm 1 Smooth Best-Response Adaptive Search

***Step 0*. Initialization**. Choose $\rho(\cdot)$ to satisfy the conditions of Definition 3.1. Set the adaptation rate $0 < \mu \ll 1$ and the exploration parameter $0 < \gamma < \Delta\widetilde{D}$, where $\Delta\widetilde{D}$ is an upper bound on the greatest difference in objective function values between any two feasible solutions. Initialize

$$\boldsymbol{\psi}_0 = \mathbf{0}_S.$$

***Step 1*. Sampling**. Sample s_n according to the randomized strategy

$$\boldsymbol{b}^{\gamma}(\boldsymbol{\psi}_n) = [b_1^{\gamma}(\boldsymbol{\psi}_n), \ldots, b_S^{\gamma}(\boldsymbol{\psi}_n)] \in \Delta\mathcal{S}$$

where $\boldsymbol{b}^{\gamma}(\boldsymbol{\psi}_n)$ is given in Definition 3.1.

***Step 2*. Evaluation**. Perform simulation to obtain $X_n(s_n)$.

***Step 3*. Belief Update**.

$$\boldsymbol{\psi}_{n+1} = \boldsymbol{\psi}_n + \mu \left[\mathbf{f}\left(s_n, \boldsymbol{\psi}_n, X_n(s_n)\right) - \boldsymbol{\psi}_n\right] \tag{5}$$

where $\mathbf{f} = \mathrm{col}(f_1, \ldots, f_S)$ and

$$f_i\left(s_n, \boldsymbol{\psi}_n, X_n(s_n)\right) = \frac{X_n(s_n)}{b_i^{\gamma}(\boldsymbol{\psi}_n)} \cdot I(s_n = i). \tag{6}$$

***Step 4*. Recursion**. Set $n \leftarrow n + 1$ and go to Step 1.

The conditions imposed on the perturbation function $\rho(\cdot)$ leads to the following distinct properties of the resulting strategy:

1) The strict concavity condition ensures uniqueness of $\boldsymbol{b}^{\gamma}(\cdot)$;
2) The boundary condition implies $\boldsymbol{b}^{\gamma}(\cdot)$ belongs to the interior of the simplex $\Delta\mathcal{S}$.

The smooth best-response sampling strategy (7) approaches the pure (unperturbed) best-response strategy

$$\arg\min_{\alpha \in \mathcal{S}} \mathbf{e}_\alpha \boldsymbol{\psi} = \arg\min_{\alpha \in \mathcal{S}} \psi_\alpha$$

when $\gamma \to 0$, and the uniform randomization, when $\gamma \to \infty$. Here, as in (1) and (3), $\mathbf{e}_\alpha$ denotes the row unit vector in S-dimensional Euclidean space with the αth element being equal to one.

It exhibits exploration using the idea of adding random values to the beliefs.[2] Such exploration is natural in any learning scenario. Let $\boldsymbol{v}$ denote a random column vector that takes values in $\mathbb{R}^S$ according to some strictly positive distribution. Given a belief vector $\boldsymbol{\psi}$, suppose one samples each feasible solution i following the *choice probability function*:

$$C_i(\boldsymbol{\psi}) = P\left(\arg\min_{\alpha\in\mathcal{S}} \mathbf{e}_\alpha[\boldsymbol{\psi} + \boldsymbol{v}] = i\right). \tag{8}$$

Using Legendre transformation, [49, Th. 2.1] shows that, regardless of the distribution of the random vector $\boldsymbol{v}$, a deterministic representation of the form (7) can be obtained for the best response strategy with added random disturbances $\boldsymbol{v}$. That is, the sampling strategy (7) with deterministic perturbation is equivalent to a strategy that minimizes the sum of the belief and the random disturbance.

The smooth best-response strategy constructs a genuine randomized strategy due to the existing exploration captured by the parameter γ. This is an appealing feature since it circumvents the discontinuity inherent in algorithms of pure best-response type [$\gamma = 0$ in (7)]; in such algorithms, small changes in the belief $\boldsymbol{\psi}$ can lead to an abrupt change in the behavior of the algorithm. Such switching behavior in the dynamics of the algorithm complicates the convergence analysis. The choice of the exploration parameter γ depends on the range of the objective function.

Remark 3.1: An example of the function $\rho(\cdot)$ in Definition 3.1 is the *entropy function* [9], [50]

$$\rho(\boldsymbol{\sigma}) = -\sum_{i\in\mathcal{S}} \sigma_i \ln(\sigma_i)$$

which gives rise to the well-known Boltzmann exploration strategy [51] with constant temperature

$$b_i^{\gamma}(\boldsymbol{\psi}) = \frac{\exp(-\psi_i/\gamma)}{\sum_{j\in\mathcal{S}} \exp(-\psi_j/\gamma)}. \tag{9}$$

Such a strategy is also used in the context of learning in games, known as logistic fictitious-play [52] or logit choice function [49].

We now proceed to describe the belief vector and its update mechanism. The belief is a vector

$$\boldsymbol{\psi}_n = \mathrm{col}\left(\psi_{1,n}, \ldots, \psi_{S,n}\right) \in \mathbb{R}^S$$

where each element $\psi_{i,n}$ is the belief up to time n about the objective function value at alternative i, which can be directly defined using the output of simulation experiments up to time n as follows:

$$\psi_{i,n} = (1-\mu)^{n-1} \left[\frac{X_1(s_1)}{b_i^{\gamma}(\boldsymbol{\psi}_1)}\right] I(s_1 = i) + \mu \sum_{2\le\tau\le n} (1-\mu)^{n-\tau} \left[\frac{X_\tau(s_\tau)}{b_i^{\gamma}(\boldsymbol{\psi}_\tau)}\right] I(s_\tau = i). \tag{10}$$

Here, $b_i^{\gamma}(\boldsymbol{\psi}_\tau)$ is the weight that the smooth best-response sampling strategy in Definition 3.1 places on sampling candidate i at time τ, and $I(X)$ denotes the indicator function. The normalization factor $1/b_i^{\gamma}(\boldsymbol{\psi}_\tau)$ makes the length of the periods that each element i is chosen comparable to other elements. The discount factor $0 < \mu \ll 1$ is further a small parameter that places more weight on recent simulation experiments and is necessary as the algorithm is deemed to track time-varying minima. The nonrecursive expression (10) verifies that components of the belief vector can be interpreted as discounted average system performance at each feasible configuration. Note that (10) relies only on the data obtained from simulation experiments; it does not require the system model nor the realizations of θ_n, which represents the random changes in the parameters underlying the discrete stochastic optimization problem. Recall that the effect of θ_τ on the random changes underlying the problem is captured in the subscript τ in $X_\tau(s_\tau)$.

Remark 3.2: The choice of step-size μ is of critical importance. Ideally, one would want μ to be small when the belief is

[2] This is in contrast to picking states at random with a small probability, as is common in game-theoretic learning and multiarmed bandit algorithms.

close to the vector of true objective values, and large otherwise. Choosing the best μ is, however, not straightforward as it depends on the dynamics of the time-varying parameters (which is unknown). One enticing solution is to use a time-varying step-size μ_n, and an algorithm which recursively updates it as the underlying parameters evolve. The updating algorithm for μ_n will in fact be a separate stochastic approximation algorithm which works in parallel with the adaptive search algorithm. Such adaptive step-size stochastic approximation algorithms are studied in [10], [11, Sec. 3.2], and [53], and are out of the scope of this paper.

IV. Main Result: Tracking the Nonstationary Global Optima

This section is devoted to obtaining asymptotic properties of the above adaptive random search algorithm. To this end, we define a regret measure in Section IV-A, which play a key role in the analysis to follow. Then, a hypermodel is defined in Section IV-B for the time variations in the parameters underlying the discrete stochastic optimization problem. The main theorem entailing the tracking properties of Algorithm 1 is presented in Section IV-C.

A. Regret Measure

The *regret* for an online learning algorithm is defined as the "opportunity loss" and compares the performance of an algorithm, selecting among S alternatives, to the performance of the best of those alternatives in hindsight[3] [54]. Accordingly, we define

$$r_n := (1-\mu)^{n-1}\left[X_1(s_1) - \overline{X}_{\min,1}\right] + \mu \sum_{2\le\tau\le n} (1-\mu)^{n-\tau}\left[X_\tau(s_\tau) - \overline{X}_{\min,\tau}\right] \quad (11)$$

where $\overline{X}_{\min,\tau}$ denotes the true global optimum at time τ and $\{s_\tau\}$ represents the sequence of candidate solutions sampled and evaluated by Algorithm 1. Since the algorithm is deemed to track nonstationary global optima, the regret r_n is defined as the discounted average performance obtained using the proposed sampling scheme as compared with the true global optimum. The regret r_n is a diagnostic that quantifies the tracking capability of the algorithm, and is not required to be evaluated iteratively by the algorithm.

B. Hypermodel

A typical method for analyzing the performance of an adaptive algorithm is to postulate a hypermodel for the underlying time variations [10]. Here, we assume that all time-varying underlying parameters in the problem are finite-state and absorbed to a vector. For simplicity, we index all possible such vectors by $\mathcal{M} = \{1,\dots,\Theta\}$, and work with the indices instead. Evolution of the underlying parameters can then be captured by an integer-valued process $\theta_n \in \mathcal{M}$, whose dynamics follow a finite-state discrete-time Markov chain. Let $\mathbf{1}_r$ and $\mathbf{0}_r$ denote r-dimensional column vectors of ones and zeros, respectively. The following assumption formally characterizes the hypermodel.

Assumption 3: Let $\{\theta_n\}$ be a discrete-time Markov chain with finite state space $\mathcal{M} = \{1,2,\dots,\Theta\}$, and transition probability matrix

$$P^\varepsilon := I + \varepsilon Q, \quad 0 < \varepsilon < 1. \quad (12)$$

Here, I denotes the $\Theta\times\Theta$ identity matrix and $Q = [q_{ij}] \in \mathbb{R}^{\Theta\times\Theta}$ is the generator of a continuous-time Markov chain satisfying

$$q_{ij} \ge 0 \text{ for } i \ne j,\ |q_{ij}| \le 1\ \forall i,j \in \mathcal{M},\ Q\mathbf{1}_\Theta = \mathbf{0}_\Theta. \quad (13)$$

Further, Q is irreducible.

Note in the above assumption that, due to the dominating identity matrix in (12), $\{\theta_n\}$ varies slowly with time.

Remark 4.1: It is important to stress that the Markov chain model is only a hypermodel for time variations underlying the optimization problem to analyze the tracking capability of the proposed adaptive random search scheme. The end-user running the algorithm is in fact oblivious to the time variations underlying the problem. In other words, θ_n does not enter implementation of the algorithm, nor are its dynamics accounted for in the algorithm design; see Assumption 2.

C. Main Tracking Result

The step-size μ in (5) determines how fast the sampling strategy is updated. The impact of switching rate of the hypermodel θ_n on convergence properties of the proposed adaptive search scheme is captured by the relationship between μ and ε in the transition probability matrix (12). Here, we assume:

Assumption 4: $\varepsilon = O(\mu)$ in the transition probability matrix P^ε in Assumption 3.

The above condition states that the time-varying parameters underlying the discrete stochastic optimization problem evolve with time on a scale that is commensurate with that determined by the step-size of Algorithm 1. In our analysis to follow, both ε and μ go to 0. This condition merely indicates that they go to 0 at the same rate.

We take the ordinary differential equation (ODE) approach—introduced by Ljung in [55] and developed later in [11]—in our convergence analysis. Thus, in lieu of working with the discrete iterates directly, we examine continuous-time piecewise constant interpolation of the iterates, which is defined as follows:

$$r^\mu(t) = r_n \quad \text{for} \quad t \in [n\mu, (n+1)\mu). \quad (14)$$

This will enable us to get a limit ODE, whose stationary points conclude the tracking result. These details are, however, relegated to a later section which provides the detailed analysis. Let $x^+ = \max\{0, x\}$. The following theorem is the main result of this paper and implies that the sequence $\{s_n\}$ generated by Algorithm 1 most frequently samples from the global optima set.

Theorem 4.1: Suppose Assumptions 1–4 hold. Let t_μ be any sequence of real numbers satisfying $t_\mu \to \infty$ as $\mu \to 0$. For any $\eta > 0$, there exists $\overline{\gamma}(\eta) > 0$ such that if $\gamma \le \overline{\gamma}(\eta)$ in (7), then $(r^\mu(\cdot + t_\mu) - \eta)^+ \to 0$ in probability as $\mu \to 0$. That is, for any $\delta > 0$

$$\lim_{\mu\to 0} \mathbb{P}\left(r^\mu(\cdot + t_\mu) - \eta \ge \delta\right) = 0.$$

Proof: See Section V for the detailed proof. ■

Interpretation of Theorem 4.1: The above theorem evidences the *tracking capability* of Algorithm 1 by looking at the worst case regret, and showing that it will be asymptotically less than some arbitrarily small value.[4] Clearly, ensuring a smaller worst case regret requires sampling the global optimizer more frequently. Put differently, the higher the difference between the objective function value of a feasible solution and the global optimum, the lower must be the empirical frequency of sampling

[3] This is in contrast to the decision theory literature, where the regret typically compares the average realized payoff for a course of action with the payoff that would have been obtained had a different course of action been chosen.

[4] This result is similar to the *Hannan consistency* notion [56] in repeated games, however, in a nonstationary setting.

that solution to maintain the regret below a certain small value. Here, $r^{\mu}(\cdot+t_{\mu})$ essentially looks at the asymptotic behavior of r_n. The requirement $t_{\mu}\to\infty$ as $\mu\to 0$ means that we look at r_n for a small μ and large n with $n\mu\to\infty$. It shows that, for a small μ, and with an arbitrarily high probability, r_n eventually spends nearly all of its time in the interval $[0,\eta+\delta)$ such that $\delta\to 0$ as $\mu\to 0$. Note that, for a small μ, r_n may escape from the interval $[0,\eta+\delta)$. However, if such an escape ever occurs, will be a "large deviations" phenomena—it will be a rare event. The order of the escape time (if it ever occurs) is often of the form $\exp(c/\mu)$ for some $c>0$. The probability of the escape in fact is exponentially small; see [11, Ch. 6.10] for details.

Remark 4.2: The choice of exploration factor γ essentially determines the size of the random disturbances $\boldsymbol{v}$ in (8), and affects both the convergence rate and efficiency of the proposed algorithm as illustrated in Fig. 2. Larger γ increases the exploration weight versus exploitation, hence, more time is spent on simulating nonpromising elements of the search space. In practice, one can initially start with $\gamma=\Delta\widetilde{D}$, where $\Delta\widetilde{D}$ is an upper bound on the greatest difference in objective function value between any two feasible solution based on the outputs of preliminary simulation experiments. To achieve an asymptotic regret of at most η in Theorem 4.1, one can then periodically solve for $\overline{\gamma}(\eta)$ in

$$\boldsymbol{b}^{\overline{\gamma}(\eta)}(\boldsymbol{\psi}_n)\cdot\boldsymbol{\psi}_n=\min_{i\in\mathcal{S}}\psi_{i,n}+\eta \tag{15}$$

and reset $0<\gamma<\overline{\gamma}(\eta)$ in the smooth best-response strategy (7).

Note that, to allow adaptivity to the time variations underlying the problem, Algorithm 1 selects nonoptimal states with some small probability. Thus, one would not expect the sequence of samples eventually spend all its time in the global optima set before it undergoes a jump. In fact, $\{s_n\}$ may visit each feasible solution infinitely often. Instead, the sampling strategy implemented by Algorithm 1 ensures the empirical frequency of sampling nonoptimal elements stays very low.

V. Analysis of the Smooth Best-Response Adaptive Search Algorithm

This section is devoted to both asymptotic and nonasymptotic analysis of the adaptive search algorithm. In this section, we work with the following two diagnostics.

(1) *Regret* r_n: It quantifies the ability of the adaptive search algorithm in tracking the random unpredictable jumps of the global optimum, and is defined in Section IV-A.

(2) *Empirical Sampling Distribution* $\mathbf{z}_n$: It is well-known that efficiency of a discrete stochastic optimization algorithm is defined as the percentage of times that the global optimum is sampled [8]. Accordingly, we define the vector

$$\mathbf{z}_n=\operatorname{col}\left(z_{1,n},\ldots,z_{S,n}\right)\in\mathbb{R}^S$$

where each element $z_{i,n}$ records the percentage of times that element i was sampled and simulated up to time n. It is iteratively updated via the stochastic approximation recursion

$$\mathbf{z}_{n+1}=\mathbf{z}_n+\mu\left[\mathbf{e}_{s_n}-\mathbf{z}_n\right],\quad 0<\mu<1 \tag{16}$$

where $\mathbf{e}_i\in\mathbb{R}^S$ denotes the unit vector whose ith component is equal to one. The small parameter μ introduces an exponential forgetting of the past sampling frequencies and allows us to track the evolution of underlying parameters. Besides quantifying efficiency, the maximizing element of the empirical sampling frequency will be used to locate the global optimizer in Section VI.

The rest of this section is organized into three subsections: Section V-A characterizes the limit system associated with discrete time iterates of the adaptive search scheme and the two diagnostics defined above. Next, Section V-B proves that such a limit system is globally asymptotically stable with probability one and characterizes its global attractors. The analysis up to this point considers μ small and n large, but μn remains bounded. Finally, to obtain the result presented in Theorem 4.1, we let μn go to infinity in Section V-C, and conclude asymptotic stability of the interpolated process associated with the regret.

A. Weak Convergence to Markovian Switching ODE

In this subsection, we use weak convergence methods to derive the limit dynamical system associated with the iterates $(r_n,\mathbf{z}_n)$. Before proceeding further, let us recall some definitions and notation:

Definition 5.1: Let Z_n and Z be $\mathbb{R}^n$-valued random vectors.

(1) Z_n converges to Z weakly, denoted by $Z_n\Rightarrow Z$, if for any bounded and continuous function $h(\cdot)$

$$\mathbb{E}h(Z_n)\to\mathbb{E}h(Z)\ \text{ as }\ n\to\infty.$$

(2) The sequence $\{Z(n)\}$ is tight if for each $\eta>0$, there exists a compact set K_η such that

$$\mathbb{P}(Z_n\in K_\eta)\geq 1-\eta\ \text{ for all }\ n.$$

Weak convergence is a generalization of convergence in distribution to a function space. The definitions of tightness and weak convergence extend to random elements in more general metric spaces. On a complete separable metric space, tightness is equivalent to relative compactness, which is known as Prohorov's Theorem [57]. By virtue of this theorem, we can extract convergent subsequences when tightness is verified. In what follows, we use a martingale problem formulation to establish the desired weak convergence. To this end, we first prove tightness. The limit process is then characterized using a certain operator related to the limit martingale problem. We refer the reader to [11, Ch. 7] for further details on weak convergence and related matters.

Define the vector of the true objective function values for all feasible solutions when $\theta_n=\theta$ is held fixed:

$$\overline{\mathbf{X}}(\theta):=\operatorname{col}\left(\overline{X}(1,\theta),\cdots,\overline{X}(S,\theta)\right) \tag{17}$$

where $\overline{X}(s,\theta)$ is defined in (4). Let further

$$\widehat{\boldsymbol{\psi}}_n:=\boldsymbol{\psi}_n-\overline{\mathbf{X}}(\theta_n) \tag{18}$$

denote the error in tracking the true objective function values via the simulation data at time n. Let further

$$\boldsymbol{Y}_n:=\begin{bmatrix}\widehat{\boldsymbol{\psi}}_n\\ r_n\\ \mathbf{z}_n\end{bmatrix}. \tag{19}$$

It can be easily verified that $\boldsymbol{Y}_n$ satisfies the recursion

$$\begin{aligned}\boldsymbol{Y}_{n+1}=\boldsymbol{Y}_n+\mu\left[\boldsymbol{A}\left(s_n,\widehat{\boldsymbol{\psi}}_n,X_n(s_n)\right)-\boldsymbol{Y}_n\right]&\\ +\mu\begin{bmatrix}\overline{\mathbf{X}}(\theta_n)-\overline{\mathbf{X}}(\theta_{n+1})\\ \mathbf{0}_{S+1}\end{bmatrix}&\end{aligned} \tag{20}$$

where

$$\boldsymbol{A}\left(s_n, \widehat{\boldsymbol{\psi}}_n, X_n(s_n)\right) = \begin{bmatrix} \mathbf{f}\left(s_n, \widehat{\boldsymbol{\psi}}_n + \overline{\mathbf{X}}(\theta_n), X_n(s_n)\right) - \overline{\mathbf{X}}(\theta_n) \\ X_n(s_n) - \overline{X}_{\min,n} \\ \mathbf{e}_{s_n} \end{bmatrix}. \tag{21}$$

Here, $\mathbf{f}(\cdot,\cdot,\cdot)$ is a vector-valued function, whose individual elements are defined in (6), and $\overline{\mathbf{X}}(\cdot)$ is defined in (17). As is widely used in the analysis of stochastic approximations, we consider the piecewise constant continuous time interpolated processes

$$\begin{aligned} \boldsymbol{Y}^{\mu}(t) &= \boldsymbol{Y}_n, \\ \theta^{\mu}(t) &= \theta_n, \end{aligned} \qquad \text{for } t \in [n\mu, (n+1)\mu). \tag{22}$$

we use $D([0,\infty) : G)$ to denote the space of functions that are defined in $[0,\infty)$, taking values in G, and are right continuous and have left limits (Càdlàg functions) with the Skorohod topology (see [11, p. 228]). The following theorem characterizes the limit process of the stochastic approximation iterates (20) as a Markovian switching ODE.

Theorem 5.1: Consider the recursion (20) and suppose Assumptions 1–4 hold. The interpolated process $(\boldsymbol{Y}^{\mu}(\cdot), \theta^{\mu}(\cdot))$ is tight in $D([0,\infty) : \mathbb{R}^{2S+1} \times \mathcal{M})$ and, as $\mu \to 0$, converges weakly to $(\boldsymbol{Y}(\cdot), \theta(\cdot))$ that is a solution of the Markovian switched ODE

$$\frac{d\boldsymbol{Y}}{dt} = \boldsymbol{F}(\boldsymbol{Y}, \theta(t)) - \boldsymbol{Y} \tag{23}$$

where

$$\boldsymbol{F}(\boldsymbol{Y}, \theta(t)) = \begin{bmatrix} \mathbf{0}_S \\ \boldsymbol{b}^{\gamma}\left(\widehat{\boldsymbol{\psi}} + \overline{\mathbf{X}}(\theta(t))\right) \cdot \overline{\mathbf{X}}(\theta(t)) - \overline{X}_{\min}(\theta(t)) \\ \boldsymbol{b}^{\gamma}\left(\widehat{\boldsymbol{\psi}} + \overline{\mathbf{X}}(\theta(t))\right) \end{bmatrix}. \tag{24}$$

Here, $\mathbf{0}_S$ is column zero vector of size S, $\overline{\mathbf{X}}(\cdot)$ is defined in (17), and

$$\overline{X}_{\min}(\theta) := \min_{\alpha \in \mathcal{S}} \overline{X}(\alpha, \theta) \tag{25}$$

where $\overline{X}(\alpha, \theta)$ is defined in Assumption 1. Further, $\theta(t)$ is a continuous time Markov chain with generator Q (see Assumption 3).

Proof: The proof uses stochastic averaging theory based on [11]; see Appendix A for the detailed argument. ■

The limit system derived in the above theorem is a dynamical system modulated by a continuous-time Markov chain $\theta(t)$. At any given instance, the Markov chain dictates which regime the system belongs to. The system then follows the corresponding ODE until the modulating Markov chain jumps into a new state—that is, the limit system (23) is only piecewise deterministic.

B. Stability Analysis of the Markovian Switching ODE

We next proceed to analyze stability and characterize the set of global attractors of the limit system. In view of (23), there exists no explicit interconnection between the dynamics of $r(t)$ and $\mathbf{z}(t)$. To obtain the result presented in Theorem 4.1, we need to look only at the stability of

$$\frac{d}{dt}\begin{bmatrix} \widehat{\boldsymbol{\psi}} \\ r \end{bmatrix} = \begin{bmatrix} \mathbf{0}_S \\ \boldsymbol{b}^{\gamma}\left(\widehat{\boldsymbol{\psi}} + \overline{\mathbf{X}}(\theta(t))\right) \cdot \overline{\mathbf{X}}(\theta(t)) - \overline{X}_{\min}(\theta(t)) \end{bmatrix} - \begin{bmatrix} \widehat{\boldsymbol{\psi}} \\ r \end{bmatrix}. \tag{26}$$

The first component is asymptotically stable, and any trajectory $\widehat{\boldsymbol{\psi}}(t)$ decays exponentially fast to $\mathbf{0}_S$ as $t \to \infty$. This essentially establishes that realizing the simulation observations $X_n(s_n)$ provides sufficient information to construct an unbiased estimator of the true objective function values. Next, substituting the global attractor $\widehat{\boldsymbol{\psi}} = \mathbf{0}_S$, we analyze stability of

$$\frac{dr}{dt} = \boldsymbol{b}^{\gamma}\left(\mathbf{X}(\theta(t)) \cdot \overline{\mathbf{X}}(\theta(t)) - \overline{X}_{\min}(\theta(t)) - r. \tag{27}$$

Let

$$\mathbb{R}_{[0,\eta)} = \{r \in \mathbb{R}; 0 \le r < \eta\}. \tag{28}$$

We break down the stability analysis of (27) into two steps: First, we examine the stability of each deterministic subsystem, associated with each $\overline{\theta} \in \mathcal{M}$ when $\theta(t) = \overline{\theta}$ is held fixed. The set of global attractors is shown to comprise $\mathbb{R}_{[0,\eta)}$ for all $\overline{\theta} \in \mathcal{M}$. The slow switching condition then allows us to apply the method of multiple Lyapunov functions [58, Ch. 3] to analyze stability of the switched system.

Theorem 5.2: Consider the limit Markovian switched ODE given in (27). Let $r(0) = r_0$ and $\theta(0) = \theta_0$. For any $\eta > 0$, there exists $\overline{\gamma}(\eta)$ such that, if $\gamma < \overline{\gamma}(\eta)$ in (7), the following results hold:

1) If $\theta(t) = \theta$ is held fixed, the set $\mathbb{R}_{[0,\eta)}$ is globally asymptotically stable for each $\theta \in \mathcal{M}$, i.e.,

$$\lim_{t \to \infty} d\left(r(t), \mathbb{R}_{[0,\eta)}\right) = 0 \tag{29}$$

where $d(\cdot,\cdot)$ denotes the usual distance function.

2) For the Markovian switching ODE, the set $\mathbb{R}_{[0,\eta)}$ is globally asymptotically stable almost surely.

Proof: For detailed proof, see Appendix B. ■

The above theorem states that the set of global attractors of the switching ODE (27) is the same as that for all deterministic ODEs, obtained by fixing $\theta(t) = \theta \in \mathcal{M}$ in (27), and constitutes $\mathbb{R}_{[0,\eta)}$. This sets the stage for Section V-C which concludes the desired tracking result in Theorem 4.1.

C. Asymptotic Stability of the Interpolated Process

This section completes the proof of Theorem 4.1 by looking at the asymptotic stability of the interpolated process

$$\mathbf{y}^{\mu}(t) = \mathbf{y}_n := \begin{bmatrix} \widehat{\boldsymbol{\psi}}_n \\ r_n \end{bmatrix} \quad \text{for } t \in [n\mu, (n+1)\mu). \tag{30}$$

In Theorem 5.1, we considered μ small and n large, but μn remained bounded. This gives a limit switched ODE for the sequence of interest as $\mu \to 0$. Here, we study asymptotic stability and establish that the limit points of the switched ODE and the stochastic approximation algorithm coincide as $t \to \infty$. We thus consider the case where $\mu \to 0$ and $n \to \infty$, however, $\mu n \to \infty$ now. Nevertheless, instead of considering a two-stage limit by first letting $\mu \to 0$ and then $t \to \infty$, we study $\mathbf{y}^{\mu}(t + t_{\mu})$ and require $t_{\mu} \to \infty$ as $\mu \to 0$. The following corollary concerns asymptotic stability of the interpolated process.

Corollary 5.1: Let

$$\mathcal{Y}^{\eta} = \left\{\mathrm{col}(\boldsymbol{x}, r); \boldsymbol{x} = \mathbf{0}_S, r \in \mathbb{R}_{[0,\eta)}\right\}. \tag{31}$$

[4] It can be shown that the sequence $\{\boldsymbol{\psi}_n\}$ induces the same asymptotic behavior as the beliefs developed using the brute force scheme [8, Ch. 5.3] about objective function values.

Denote by $\{t_\mu\}$ any sequence of real numbers satisfying $t_\mu \to \infty$ as $\mu \to 0$. Assume $\{\mathbf{y}(n) : \mu > 0, n < \infty\}$ is tight or bounded in probability. Then, for each $\eta \geq 0$, there exists $\overline{\gamma}(\eta) \geq 0$ such that if $\gamma \leq \overline{\gamma}(\eta)$ in (7)

$$\mathbf{y}^\mu(\cdot + t_\mu) \to \mathcal{Y}^\eta, \quad \text{as } \mu \to 0. \tag{32}$$

Proof: We only give an outline of the proof, which essentially follows from Theorems 5.1 and 5.2. Define $\widehat{\mathbf{y}}^\mu(\cdot) = \mathbf{y}^\mu(\cdot + t_\mu)$. Then, it can be shown that $\widehat{\mathbf{y}}^\mu(\cdot)$ is tight. For any $T_1 < \infty$, take a weakly convergent subsequence of $\{\widehat{\mathbf{y}}^\mu(\cdot), \widehat{\mathbf{y}}^\mu(\cdot - T_1)\}$. Denote the limit by $(\widehat{\mathbf{y}}(\cdot), \widehat{\mathbf{y}}_{T_1}(\cdot))$. Note that $\widehat{\mathbf{y}}(0) = \widehat{\mathbf{y}}_{T_1}(T_1)$. The value of $\widehat{\mathbf{y}}_{T_1}(0)$ may be unknown, but the set of all possible values of $\widehat{\mathbf{y}}_{T_1}(0)$ (over all T_1 and convergent subsequences) belongs to a tight set. Using this and Theorems 5.1 and 5.2, for any $\varrho > 0$, there exists a $T_\varrho < \infty$ such that for all $T_1 > T_\varrho$, $d(\widehat{\mathbf{y}}_{T_1}(T_1), \mathcal{Y}^\eta) \geq 1 - \varrho$. This implies that $d(\widehat{\mathbf{y}}(0), \mathcal{Y}^\eta) \geq 1 - \varrho$, and the desired result follows. ■

VI. Case of Slow Random Time Variations

This section is devoted to the analysis of the adaptive search scheme when the random evolution of the parameters underlying the discrete stochastic optimization problem occurs on a timescale that is much slower as compared to the adaptation rate of the adaptive search algorithm with step-size μ. More precisely, we replace Assumption 4 with the following assumption.

Assumption 5: $0 < \varepsilon \ll \mu$ in the transition probability matrix P^ε in Assumption 3.

The above assumption introduces a different timescale to Algorithm 1, which leads to an asymptotic behavior that is fundamentally different as compared with the case $\varepsilon = O(\mu)$ that we analyzed in Section V. Under Assumption 5, $\theta(\cdot)$ is the slow component and $\boldsymbol{Y}(\cdot)$ is the fast transient in (23). It is well-known that, in such two timescale systems, the slow component is quasi-static—remains almost constant—while analyzing the behavior of the fast timescale. The weak convergence argument then shows that $(\boldsymbol{Y}^\mu(\cdot), \theta^\mu(\cdot))$ converges weakly to $(\boldsymbol{Y}(\cdot), \theta)$ as $\mu \to 0$ such that the limit $\boldsymbol{Y}(\cdot)$ is a solution to

$$\frac{d\boldsymbol{Y}}{dt} = \boldsymbol{F}(\boldsymbol{Y}, \theta) - \boldsymbol{Y} \tag{33}$$

where

$$\boldsymbol{F}(\boldsymbol{Y}, \theta) = \begin{bmatrix} \mathbf{0}_S \\ \boldsymbol{b}^\gamma\left(\widehat{\boldsymbol{\psi}} + \overline{\mathbf{X}}(\theta)\right) \cdot \overline{\mathbf{X}}(\theta) - \overline{X}_{\min}(\theta) \\ \boldsymbol{b}^\gamma\left(\widehat{\boldsymbol{\psi}} + \overline{\mathbf{X}}(\theta)\right) \end{bmatrix}. \tag{34}$$

Technical details are omitted for brevity; see [11, Ch. 8] for details.

Our task is then to investigate stability of this limit system. Recall (19). In view of (34), there exists no explicit interconnection between the dynamics of r and $\mathbf{z}$ in (33). Therefore, we start by looking at

$$\frac{d}{dt}\begin{bmatrix} \widehat{\boldsymbol{\psi}} \\ r \end{bmatrix} = \begin{bmatrix} \mathbf{0}_S \\ \boldsymbol{b}^\gamma\left(\widehat{\boldsymbol{\psi}} + \overline{\mathbf{X}}(\theta)\right) \cdot \overline{\mathbf{X}}(\theta) - \overline{X}_{\min}(\theta) \end{bmatrix} - \begin{bmatrix} \widehat{\boldsymbol{\psi}} \\ r \end{bmatrix}.$$

The first component is asymptotically stable, and any trajectory $\widehat{\boldsymbol{\psi}}(t)$ decays exponentially fast to $\mathbf{0}_S$ as $t \to \infty$. The first part in Theorem 5.2 also shows that, for the second component, the set $\mathbb{R}_{[0,\eta)}$ is globally asymptotically stable for all $\theta \in \mathcal{M}$.

It remains to analyze stability of

$$\frac{d}{dt}\begin{bmatrix} \widehat{\boldsymbol{\psi}} \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_S \\ \boldsymbol{b}^\gamma\left(\widehat{\boldsymbol{\psi}} + \overline{\mathbf{X}}(\theta)\right) \end{bmatrix} - \begin{bmatrix} \widehat{\boldsymbol{\psi}} \\ \mathbf{z} \end{bmatrix}.$$

The first component is asymptotically stable. Substituting the global attractor $\widehat{\boldsymbol{\psi}} = \mathbf{0}_S$, we look at

$$\frac{d\mathbf{z}}{dt} = \boldsymbol{b}^\gamma\left(\overline{\mathbf{X}}(\theta)\right) - \mathbf{z}$$

which is globally asymptotically stable. Further, and any trajectory $\mathbf{z}(t)$ converges exponentially fast to $\boldsymbol{b}^\gamma(\overline{\mathbf{X}}(\theta))$ as $t \to \infty$.

Combining the above steps, one can obtain the same asymptotic stability result as that presented in Corollary 5.1. Further, using a similar argument, it can be shown that

$$\mathbf{z}^\mu(\cdot + t_\mu) \to \boldsymbol{b}^\gamma\left(\overline{\mathbf{X}}(\theta)\right) \quad \text{in probability as } \mu \to 0$$

where $\{t_\mu\}$ is any sequence of real numbers satisfying $t_\mu \to \infty$ as $\mu \to 0$. Therefore, for any $\delta > 0$

$$\lim_{\mu \to 0} \mathbb{P}\left(\left\|\mathbf{z}^\mu(\cdot + t_\mu) - \boldsymbol{b}^\gamma\left(\overline{\mathbf{X}}(\theta)\right)\right\| \geq \delta\right) = 0 \tag{35}$$

where $\|\cdot\|$ denotes the Euclidean norm. Feeding in the vector of true objective function values $\overline{\mathbf{X}}(\theta)$ for any θ, the smooth best-response strategy in Definition 3.1 outputs a randomized strategy with the highest probability assigned to the global optimum. This in turn implies that the maximizing element of the empirical sampling distribution $\mathbf{z}^\mu(\cdot + t_\mu)$ represents the global optimum. The following corollary summarizes this result.

Corollary 6.1: Denote the most frequently visited state by

$$s_n^{\max} = \arg\max_{i \in \mathcal{S}} \; z_{i,n}$$

where $z_{i,n}$ is the ith component of $\mathbf{z}_n$. Further, define the continuous time interpolated sequence

$$s^{\max,\mu}(t) = s_n^{\max} \quad \text{for} \quad t \in [n\mu, (n+1)\mu).$$

Then, under Assumptions 1, 3, and 5, $s^{\max,\mu}(\cdot + t_\mu)$ converges in probability to the global optima set $\mathcal{Q}(\theta)$. That is, for any $\delta > 0$

$$\lim_{\mu \to 0} \mathbb{P}\left(d\left(s^{\max,\mu}(\cdot + t_\mu), \mathcal{Q}(\theta)\right) \geq \delta\right) = 0 \tag{36}$$

where $d(\cdot, \cdot)$ denotes the usual distance function.

The above corollary simply asserts that, for a small μ and for any $\theta \in \mathcal{M}$, the most frequently sampled element, with a high probability, eventually spends nearly all its time in the global optima set.

VII. Numerical Examples

This section illustrates the performance of Algorithm 1, henceforth referred to as AS, using the examples in [26], [27]. The performance of the AS scheme will be compared against the following algorithms in the literature, which are proved to be globally convergent:

1) Random Search (RS) [26], [28]: RS is a modified hill-descending algorithm. Let $\boldsymbol{\pi}_n \in \mathbb{R}^S$ represent the empirical sampling frequency of the elements of the search space. The RS algorithm is summarized in Algorithm 2. For static discrete stochastic optimization problems, the constant step-size μ in (37) is replaced with the decreasing step-size $\mu_n = 1/n$. Each iteration of the RS algorithm requires one random number selection, $\mathcal{O}(S)$ arithmetic operations, one comparison and two independent simulation experiments.

Algorithm 2 Random search (RS) [26], [28]

```
Step 0. Initialization. Select $s_0 \in \mathcal{S}$. Set $\pi_{i,0} = 1$ if $i = s_0$, and 0 otherwise. Set $s_0^* = s_0$ and $n = 1$.
Step 1. Random search. Sample a candidate solution $s'_n$ uniformly from the set $\mathcal{M} - s_{n-1}$.
Step 2. Evaluation and Acceptance. Simulate to obtain $X_n(s_{n-1})$ and $X_n(s'_n)$. If $X_n(s'_n) < X_n(s_{n-1})$, let $s_n = s'_n$; otherwise, let $s_n = s_{n-1}$.
Step 3. Occupation Probability Update.
    $\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1} + \mu \left[ \mathbf{e}_{s_n} - \boldsymbol{\pi}_{n-1} \right], \quad 0 < \mu < 1 \qquad (37)$
    where $\mathbf{e}_i$ is a unit vector with the $i$th element being equal to one.
Step 4. Global Optimum Estimate.
    $s_n^* \in \arg\max_{i \in \mathcal{S}} \pi_{i,n}.$
Let $n \leftarrow n + 1$ and go to Step 1.
```

2) Upper Confidence Bound (UCB) [41], [43]: The UCB algorithms belong to the to the family of "follow the perturbed leader" algorithms. Let B denote an upper bound on the objective function and $\xi > 0$ be a constant. The UCB algorithm is summarized below in Algorithm 3. For a static discrete stochastic optimization problem, we set $\mu = 1$ in (39); otherwise, the discount factor μ has to be chosen in the interval (0, 1). Each iteration of the UCB algorithm requires $\mathcal{O}(S)$ arithmetic operations, one maximization and one simulation of the objective function.

Algorithm 3 Upper confidence bound (UCB) [41], [43]

```
Step 0. Initialization. For each $i \in \mathcal{S}$, simulate to obtain $X_0(i)$, and set $\widehat{X}_{i,0} = X_0(i)$, $m_{i,0} = 1$. Select $0 < \mu \leq 1$, and set $n = 1$.
Step 1. Sampling. Sample a candidate solution
    $s_n = \arg\max_{i \in \mathcal{S}} \left[ \widehat{X}_{i,n-1} + 2B \sqrt{\frac{\xi \ln(M_{n-1} + 1)}{m_{i,n-1}}} \right]$
    where $m_{i,n-1} = \sum_{\tau=1}^{n-1} \mu^{n-1-\tau} I(s_\tau = i)$
          $M_{n-1} = \sum_{i=1}^{S} m_{i,n-1} \qquad (38)$
    and $B$ is an upper bound on the objective function.
Step 2. Evaluation. Simulate to obtain $X_n(s_n)$.
Step 3. Update Belief.
    $\widehat{X}_{i,n} = \frac{1}{m_{i,n}} \sum_{\tau=1}^{n} \mu^{n-\tau} X_\tau(s_\tau) I(s_\tau = i). \qquad (39)$
Step 4. Global Optimum Estimate.
    $s_n^* \in \arg\max_{i \in \mathcal{S}} \widehat{X}_{i,n}.$
Let $n \leftarrow n + 1$ and go to Step 1.
```

Notice that the UCB algorithm is designed to find the global maximizer, whereas both the AS and RS schemes seek to find the global minimizer. Therefore, one needs to be careful to accordingly modify the simulation data (multiply by -1 in the case of UCB) when comparing these algorithms. Throughout this section, we use $\rho(x)$ as in Remark 3.1. In comparison to the RS and UCB algorithms, the proposed scheme requires $\mathcal{O}(S)$ arithmetic operations, one random number selection and one simulation of the objective function at each iteration.

In what follows, we start with a static discrete stochastic optimization example, and then proceed to the regime-switching setting.

A. Example 1: Static Discrete Stochastic Optimization

Consider the following example described in [26, Sec. 4]. Suppose that the demand d_n for a particular product has a Poisson distribution with parameter λ:

$$d_n \sim f(\alpha; \lambda) = \frac{\lambda^\alpha \exp(-\lambda)}{\alpha!}.$$

The objective is then to find the number that maximizes the demand probability, subject to the constraint that at most S units can be ordered. This problem can be formulated as a discrete deterministic optimization problem

$$\operatorname*{arg\,max}_{\alpha \in \{0,1,\ldots,S\}} \left[f(\alpha; \lambda) = \frac{\lambda^\alpha \exp(-\lambda)}{\alpha!} \right] \tag{40}$$

which can be solved analytically. Here, we aim to solve the following stochastic variant. Find

$$\operatorname*{arg\,min}_{\alpha \in \{0,1,\ldots,S\}} -\mathbb{E}\left\{ I(d_n = \alpha) \right\} \tag{41}$$

where $I(\cdot)$ denotes the indicator function. Clearly, the set of global optimizers of (40) and (41) coincide. This enables us to check the results obtained using the search schemes.

We consider the following two cases of the rate parameter λ in (40): i) $\lambda = 1$, which implies that the set of global optimizers is $\mathcal{Q} = \{0, 1\}$, and ii) $\lambda = 10$, in which case the set of global optimizers is $\mathcal{Q} = \{9, 10\}$. For each case, we further study the effect of the size of search space on the performance of algorithms by considering two instances: i) $S = 10$, and ii) $S = 100$. Since the problem is static in the sense that $\mathcal{Q}$ is fixed for each case, one can use the results of [59] to show that if the exploration factor γ in (7) decreases to zero sufficiently slowly, the sequence of samples $\{s_n\}$ converges almost surely to the global minimum. More precisely, we consider the following modifications to Algorithm 1:

(1) The constant step-size μ in (5) is replaced by decreasing step-size $\mu_n = 1/n$;

(2) The exploration factor γ in (7) is replaced by $1/n^\beta$, $0 < \beta < 1$.

By the above construction, $\{s_n\}$ will eventually become reducible with singleton communicating class $\mathcal{Q}$. That is, the sequence of samples $\{s_n\}$ eventually spends all its time in the global optimum. This is in contrast to the sequence of samples taken by Algorithm 1 in the regime-switching setting, which is discussed in the following example. The main reason for the above modifications is comparability to existing schemes, which use decreasing step-sizes, for the performance comparison purpose. In the context of stochastic recursive algorithms, it is unfair to compare the performance of a constant step-size algorithm with a diminishing step-size algorithm. Therefore, we take two approaches: In this example, the above modifications are made to devise a diminishing step-size variant of Algorithm 1. In Example 2, we make modifications on the existing algorithms to devise their corresponding constant step-size variants, and compare them with Algorithm 1 with no modifications.

TABLE I
EXAMPLE 1: PERCENTAGE OF INDEPENDENT RUNS OF ALGORITHMS THAT CONVERGED TO THE GLOBAL OPTIMUM SET IN n ITERATIONS. (a) $\lambda = 1$. (b) $\lambda = 10$

Iteration	$S = 10$			$S = 100$		
n	AS	RS	UCB	AS	RS	UCB
10	55	39	86	11	6	43
50	98	72	90	30	18	79
100	100	82	95	48	29	83
500	100	96	100	79	66	89
1000	100	100	100	93	80	91
5000	100	100	100	100	96	99
10000	100	100	100	100	100	100

(a)

Iteration	$S = 10$			$S = 100$		
n	AS	RS	UCB	AS	RS	UCB
10	29	14	15	7	3	2
100	45	30	41	16	9	13
500	54	43	58	28	21	25
1000	69	59	74	34	26	30
5000	86	75	86	60	44	44
10000	94	84	94	68	49	59
20000	100	88	100	81	61	74
50000	100	95	100	90	65	81

(b)

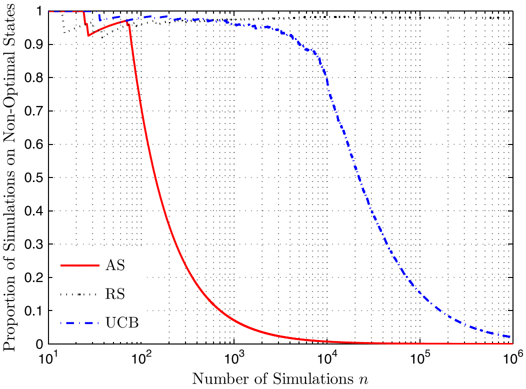

Fig. 1. Example 1. Proportion of simulation effort expended on states outside the global optima set ($\lambda = 1$, $S = 100$).

In this example, we set $\beta = 0.2$ and $\gamma = 0.01$. We further set $B = 1$, and $\xi = 0.5$ in Algorithm 3. To compare the computational costs of the schemes and give a fair comparison, we use the number of simulation experiments performed by each algorithm when evaluating its performance. Close scrutiny of Table I leads to the following observations: In all three algorithms, the speed of convergence decreases when either S or λ (or both) increases. However, the effect of increasing λ is more substantial since the objective function values of the worst and best states are closer when $\lambda = 10$. Given equal number of simulation experiments, higher percentage of cases that a particular method has converged to the global optima indicates convergence at a faster rate. Table I confirms that the AS algorithm ensures faster convergence in each case.

To evaluate and compare efficiency of the algorithms, the sample path of the number of simulation experiments performed on nonoptimal feasible solutions, i.e., $1 - \sum_{i \in \mathcal{Q}} z_{i,n}$, is plotted in Fig. 1, when $\lambda = 1$ and $S = 100$. As can be seen, since the RS method randomizes among all (except the previously sampled) feasible solutions at each iteration, it performs approximately 98% of the simulations on nonoptimal elements. Further, the UCB algorithm switches to its exploitation phase after a longer period of exploration as compared to the AS algorithm. Fig. 1 thus confirms that the AS algorithm has the supreme balance between exploration of the search space and exploitation of the simulation data. Fig. 2 further illustrates the sensitivity of both the convergence speed and the efficiency of Algorithm 1 to the exploration parameter γ. As can be seen, larger γ increases both the number of simulation experiments that the algorithm requires to locate the global optimum, and the proportion of simulations performed on nonoptimal feasible solutions.

Finally, Fig. 3 considers Algorithm 1 in a static setting, however, with no modifications to the step-size (i.e., constant step-size μ) and illustrates the implication of the convergence result in Theorem 4.1, namely, lower asymptotic regrets are achievable by choosing smaller step-sizes μ. As can be seen,

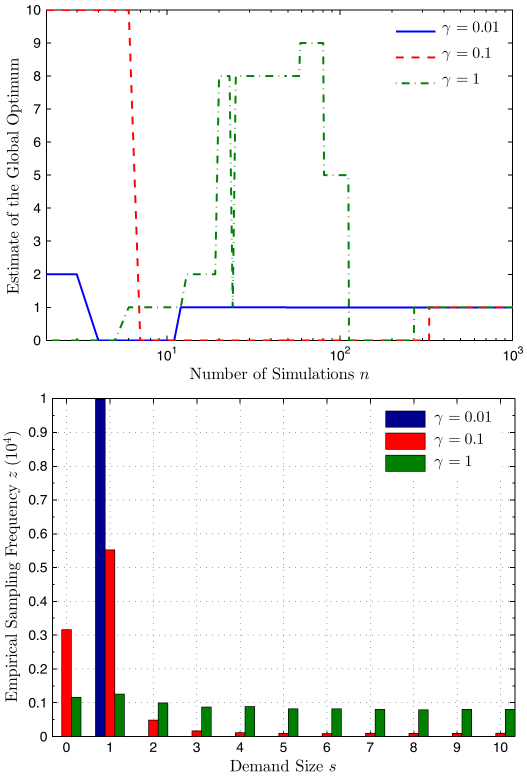

Fig. 2. Example 1. Sensitivity to γ when $\lambda = 1$ and $S = 10$. (Top) Convergence speed. (Bottom) The distribution of simulation experiments (efficiency) after $n = 10^4$ iterations.

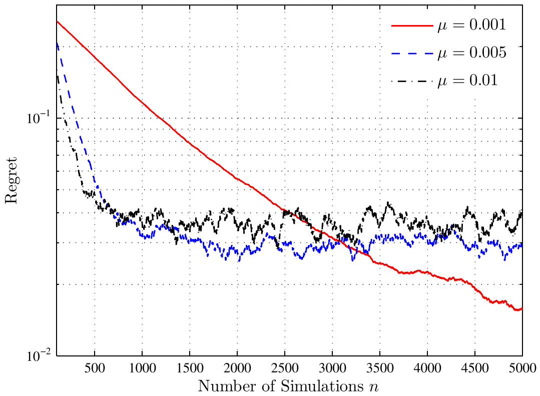

Fig. 3. Example 1. Smaller adaptation rate μ ensures lower asymptotic regret at the expense of higher transient regret due to lower learning rate ($\lambda = 1$, $S = 10$).

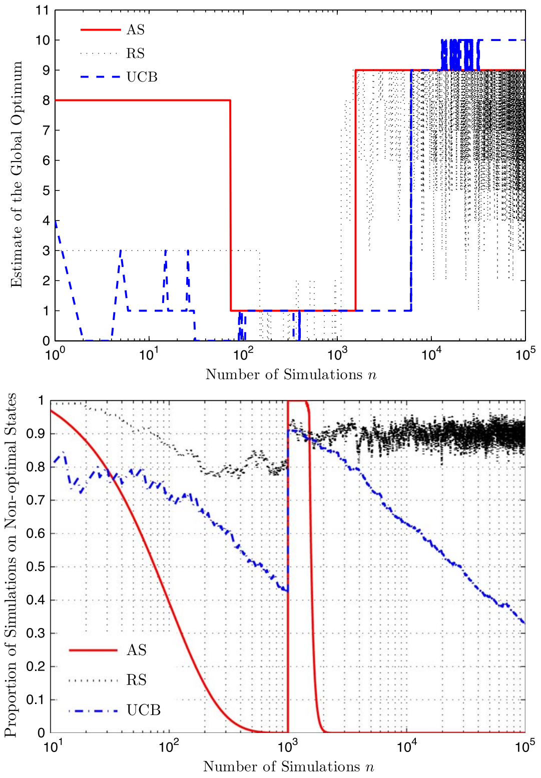

Fig. 4. Example 2. (Top) Sample path of the estimate of the global optimum. The global optima set is {0, 1}, for $0 < n < 10^3$, and {9, 10}, for $10^3 \leq n < 10^5$. (Bottom) Proportion of simulation experiments on nonoptimal feasible solutions when the global optima set evolves at $n = 10^3$.

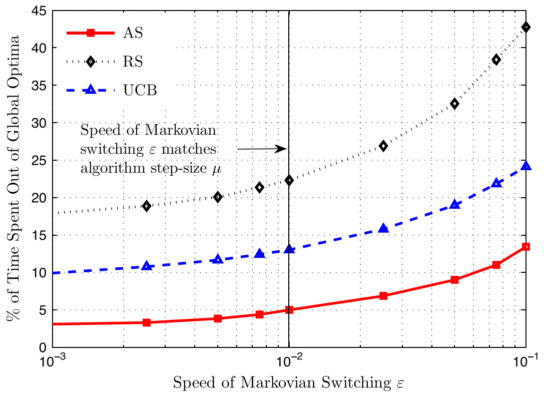

Fig. 5. Example 2. Proportion of time the global optimum estimate does not match the true global optimum versus the switching speed of the global optimizers set after $n = 10^6$ iterations.

lower asymptotic regrets come at the expense of higher regrets experienced in the transient state, due to slower learning rate μ. This further justifies the use of a decreasing step-size in the static settings; that is, the step-size is larger at the beginning to enable faster learning and improve transient behavior and becomes smaller over time to asymptotically achieve zero regret. In nonstationary environments, however, decreasing step-sizes will prevent the algorithm to follow the random unpredictable jumps of the global optimum, thereby justifying the use of constant step-sizes.

B. Example 2: Regime-Switching Discrete Stochastic Optimization

Consider the discrete stochastic optimization problem described in Example 1 with the exception that now $\lambda(\theta_n)$ jump changes between 1 and 10 according to a Markov chain $\{\theta_n\}$ with state space $\mathcal{M} = \{1, 2\}$, and transition probability matrix

$$P^{\varepsilon} = I + \varepsilon Q, \quad Q = \begin{bmatrix} -0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix}. \tag{42}$$

The regime-switching discrete stochastic optimization we aim to solve is then (41), where $S = 10$

$$d_n \sim f\left(\alpha; \lambda(\theta_n)\right) = \frac{\lambda^{\alpha}(\theta_n) \exp\left(-\lambda(\theta_n)\right)}{\alpha!} \tag{43}$$

and $\lambda(1) = 1$ and $\lambda(2) = 10$. The sets of global optimizers are then, $\mathcal{Q}(1) = \{0, 1\}$ and $\mathcal{Q}(2) = \{9, 10\}$, respectively. In the rest of this section, we assume $\gamma = 0.1$, and $\mu = \varepsilon = 0.01$.

Fig. 4 (top) shows tracking capability of the algorithms for a slow Markov chain $\{\theta_n\}$ that undergoes a jump from $\theta = 1$ to $\theta = 2$ at $n = 10^3$. As can be seen, contrary to the RS algorithm, both the AS and UCB methods properly track the changes; however, the AS algorithm responds faster to the jump. In the sample path shown, $s_0 = 8$. Due to the reinforcement learning nature of the AS algorithm, sampling a particular feasible solution increases the probability of sampling it again in the following period. This, together with the fact that the global optimizer is estimated by the most frequently sampled feasible solution, explains why the estimate of the global optimizer is 8 until 75 simulations have been performed. However, as n increases, more simulations are preformed on other feasible solutions due to the exploration component in the sampling strategy, hence, the estimates of objective function values become more accurate. Simultaneously, the sampling strategy is being updated using these more accurate estimates by overweighting the feasible solution with the minimum estimate value. This steadily increases the empirical frequency of sampling the true global optimizer, and enables the algorithm to detect it after 75 iterations. This effect is illustrated in Fig. 4 (bottom) by showing that the frequency of sampling nonoptimal feasible solutions declines as n increases. Evidently, the AS algorithm provides the supreme balance between exploration and exploitation, and properly responds to the random switching.

Finally, Fig. 5 compares efficiency of algorithms for different values of ε in (42), which determines the speed of random switchings. Each point on the graph is an average over 100 independent runs of 10^6 iterations of the algorithms. As expected, the estimate of the global optimum spends more time in the global optimum for all methods as the speed of time variations decreases. The estimate provided by the AS algorithm, however, differs from the true global optimum less frequently as compared with the RS and UCB methods.

VIII. Conclusion

This paper has considered discrete stochastic optimization via simulation where the global optimum is nonstationary and undergoes random unpredictable jumps over time. We proposed a class of sampling-based adaptive search algorithms based on a smooth best-response strategy, that is inspired by fictitious play learning rules in the game theory. The convergence analysis proved that, if the underlying time variations occur on the same timescale as the updates of the proposed adaptive search algorithm, it will properly track the time variations of the global optima set by showing that a regret measure can be made and kept arbitrarily small. Further, if the global optimum evolves on a slower timescale, the most frequently sampled solution tracks the global optima set. The proposed scheme thus exhausts most of its simulation efforts on the global optima set, which promotes its deployment as an online control mechanism to enable self-configuration of large-scale stochastic systems. It further allows correlation amongst the collected simulation observations, which is more realistic in practice. Numerical examples confirmed superior efficiency and convergence speed as compared with the existing random search and pure exploration methods.

Appendix A
Proof of Theorem 5.1

We first prove tightness of the interpolated process $\boldsymbol{Y}^{\mu}(\cdot)$. Consider the sequence $\{\boldsymbol{Y}_n\}$, defined in (19). Because of the boundedness of the sequences, for any $0 < T_1 < \infty$ and $n \leq T_1/\mu$, $\{\boldsymbol{Y}_n\}$ is bounded w.p.1. As a result, for any $\tilde{\kappa} > 0$

$$\sup_{n \leq \frac{T_1}{\mu}} \mathbb{E}\|\boldsymbol{Y}_n\|^{\tilde{\kappa}} < \infty$$

where in the above and hereafter $\|\cdot\|$ denotes the Euclidean norm. We will also use t/μ to denote the integer part of t/μ for each $t > 0$. Next, considering the interpolated process $\boldsymbol{Y}^{\mu}(\cdot)$ (defined in (22)) and the recursion (20), for any $t, u > 0$, $\delta > 0$, and $u < \delta$, it can be verified that

$$\boldsymbol{Y}^{\mu}(t+u) - \boldsymbol{Y}^{\mu}(t) = \mu \sum_{k=\frac{t}{\mu}}^{(t+u)/\mu-1} \left[\boldsymbol{A}\Big(s_k, \widehat{\boldsymbol{\psi}}_k, X_k(s_k)\Big) - \boldsymbol{Y}_k\right] + \sum_{k=\frac{t}{\mu}}^{(t+u)/\mu-1} \begin{bmatrix} \overline{\mathbf{X}}(\theta_k) - \overline{\mathbf{X}}(\theta_{k+1}) \\ \mathbf{0}_{S+1} \end{bmatrix} \tag{44}$$

where $\boldsymbol{A}(\cdot,\cdot,\cdot)$ is a vector function defined in (21). The boundedness of the sequences $\{X_n(s,\theta)\}$, $\{s_n\}$, and $\{\theta_n\}$ then implies that

$$\mathbb{E}_t^{\mu} \|\boldsymbol{Y}^{\mu}(t+u) - \boldsymbol{Y}^{\mu}(t)\|^2 = O(u)$$

where $\mathbb{E}_t^{\mu}$ denotes the conditional expectation given the σ-algebra generated by the μ-dependent past data up to time t. By virtue of the tightness criteria [60, Th. 3, p. 47] or [11, Ch. 7], it follows that

$$\lim_{\delta \to 0} \limsup_{\mu \to 0} \left[\mathbb{E}\left\{\sup_{0 \leq u \leq \delta} \mathbb{E}_t^{\mu} \|\boldsymbol{Y}^{\mu}(t+u) - \boldsymbol{Y}^{\mu}(t)\|^2\right\}\right] = 0. \tag{45}$$

Therefore, $\boldsymbol{Y}^{\mu}(\cdot)$ is tight in $D([0,\infty] : \mathbb{R}^{2S+1})$. In view of [28, Prop. 4.4], $\theta^{\mu}(\cdot) \Rightarrow \theta(\cdot)$ such that $\theta(\cdot)$ is a continuous time Markov chain with generator Q; see Assumption 3. As a result, the pair $(\boldsymbol{Y}^{\mu}(\cdot), \theta^{\mu}(\cdot))$ is tight in $D([0,\infty] : \mathbb{R}^{2S+1} \times \mathcal{M})$.

Using Prohorov's theorem [11], one can extract a weakly convergent subsequence. For notational simplicity, we still denote the subsequence by $\boldsymbol{Y}^{\mu}(\cdot)$ with limit $\boldsymbol{Y}(\cdot)$. By the Skorohod representation theorem [11] (with a slight abuse of notation), $\boldsymbol{Y}^{\mu}(\cdot) \to \boldsymbol{Y}(\cdot)$ w.p.1, and the convergence is uniform on any compact interval. We now proceed to characterize the limit $\boldsymbol{Y}(\cdot)$ using martingale averaging methods.

First, one can show the last term in (44) contributes nothing to the limit. To obtain the desired limit, it will then be proved that the limit $(\boldsymbol{Y}(\cdot), \theta(\cdot))$ is the solution of the martingale problem with operator $\mathcal{L}$ defined as follows. For all $i \in \mathcal{M}$,

$$\begin{aligned} \mathcal{L}g(\boldsymbol{x}, i) &= \nabla'_{\boldsymbol{x}} g(\boldsymbol{x}, i)\left[\boldsymbol{F}(\boldsymbol{x}, i) - \boldsymbol{x}\right] + Qg(\boldsymbol{x}, \cdot)(i) \\ Qg(\boldsymbol{x}, \cdot)(i) &= \sum_{j \in \mathcal{M}} q_{ij} g(\boldsymbol{x}, j) \end{aligned} \tag{46}$$

and, for each $i \in \mathcal{M}$, $g(\cdot, i) : \mathbb{R}^r \mapsto \mathbb{R}$ with $g(\cdot, i) \in C_0^1$ (C^1 function with compact support). Further, $\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, i)$ denotes the gradient of $g(\boldsymbol{x}, i)$ with respect to $\boldsymbol{x}$, and $\boldsymbol{F}(\cdot,\cdot)$ is defined in (24). Using an argument similar to [12, Lemma 7.18], one can show that the martingale problem associated with the operator $\mathcal{L}$ has a unique solution. Thus, it remains to prove that the limit $(\boldsymbol{Y}(\cdot), \theta(\cdot))$ is the solution of the martingale problem. To this end, it suffices to show that, for any positive arbitrary integer κ_0, and for any $t, u > 0$, $0 < t_\iota \leq t$ for all $\iota \leq \kappa_0$, and any bounded continuous function $h(\cdot, i)$, for all $i \in \mathcal{M}$

$$\mathbb{E}h\left(\boldsymbol{Y}(t_\iota), \theta(t_\iota) : \iota \leq \kappa_0\right) \times \left[g\left(\boldsymbol{Y}(t+u), \theta(t+u)\right) - g\left(\boldsymbol{Y}(t), \theta(t)\right) - \int_t^{t+u} \mathcal{L}g\left(\boldsymbol{Y}(v), \theta(v)dv\right)\right] = 0. \tag{47}$$

To verify (47), we work with $(\boldsymbol{Y}^{\mu}(\cdot), \theta^{\mu}(\cdot))$ and prove that the above equation holds as $\mu \to 0$.

By the weak convergence of $(\boldsymbol{Y}^{\mu}(\cdot), \theta^{\mu}(\cdot))$ to $(\boldsymbol{Y}(\cdot), \theta(\cdot))$ and the Skorohod representation, it can be seen that

$$\begin{aligned} &\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_\iota), \theta^{\mu}(t_\iota) : \iota \leq \kappa_0\right) \\ &\quad \times \left[g\left(\boldsymbol{Y}^{\mu}(t+u), \theta^{\mu}(t+u)\right) - g\left(\boldsymbol{Y}^{\mu}(t), \theta^{\mu}(t)\right)\right] \\ &\to \mathbb{E}h\left(\boldsymbol{Y}(t_\iota), \theta(t_\iota) : \iota \leq \kappa_0\right) \\ &\quad \times \left[g\left(\boldsymbol{Y}(t+u), \theta(t+u)\right) - g\left(\boldsymbol{Y}(t), \theta(t)\right)\right]. \end{aligned}$$

Now, choose a sequence of integers $\{n_\mu\}$ such that $n_\mu \to \infty$ as $\mu \to 0$, but $\delta_\mu = \mu n_\mu \to 0$, and partition $[t, t+u]$ into subintervals of length δ_μ. Then

$$\begin{aligned} &g\left(\boldsymbol{Y}^{\mu}(t+u), \theta^{\mu}(t+u)\right) - g\left(\boldsymbol{Y}^{\mu}(t), \theta^{\mu}(t)\right) \\ &\quad = \sum_{\ell:\ell\delta_\mu = t}^{t+u} \left[g\left(\boldsymbol{Y}_{\ell n_\mu + n_\mu}, \theta_{\ell n_\mu + n_\mu}\right) - g\left(\boldsymbol{Y}_{\ell n_\mu}, \theta_{\ell n_\mu + n_\mu}\right)\right] \\ &\quad\quad + \sum_{\ell:\ell\delta_\mu = t}^{t+u} \left[g\left(\boldsymbol{Y}_{\ell n_\mu}, \theta_{\ell n_\mu + n_\mu}\right) - g\left(\boldsymbol{Y}_{\ell n_\mu}, \theta_{\ell n_\mu}\right)\right] \end{aligned} \tag{48}$$

where $\sum_{\ell:\ell\delta_\mu = t}^{t+u}$ denotes the sum over ℓ in the range $t \leq \ell\delta_\mu < t + u$.

First, we consider the second term on the r.h.s. of (48); see (49), shown at the top of the next page. Concerning the first term on the r.h.s. of (48), see (50) shown at the top of the next page, where $\nabla_{\boldsymbol{x}} y$ denotes the gradient column vector with respect to vector $\boldsymbol{x}$, and $\nabla'_{\boldsymbol{x}} g$ represents its transpose. For notational simplicity, we write $\nabla_{\widehat{\boldsymbol{\psi}}} g(\boldsymbol{Y}_{\ell n_\mu}, \theta_{\ell n_\mu})$, $\nabla_r g(\boldsymbol{Y}_{\ell n_\mu}, \theta_{\ell n_\mu})$, and $\nabla_{\mathbf{z}} g(\boldsymbol{Y}_{\ell n_\mu}, \theta_{\ell n_\mu})$ as $\nabla_{\widehat{\boldsymbol{\psi}}} g$, $\nabla_r g$, and $\nabla_{\mathbf{z}} g$ respectively. The rest of the proof is divided into three steps, each concerning one of the terms in (50).

Step 1: We start by looking at the first term on the r.h.s. of (50), and rewrite it as (51), shown at the top of the next page, where $X_k(\cdot, \theta)$ is defined by Assumption 1. Note that

$$\begin{aligned}
&\lim_{\mu\to0}\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_\iota),\theta^{\mu}(t_\iota):\iota\le\kappa_0\right)\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\left[g\left(\boldsymbol{Y}_{\ell n_\mu},\theta_{\ell n_\mu+\mu}\right)-g\left(\boldsymbol{Y}_{\ell n_\mu},\theta_{\ell n_\mu}\right)\right]\right]\\
&\quad=\lim_{\mu\to0}\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_\iota),\theta^{\mu}(t_\iota):\iota\le\kappa_0\right)\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\sum_{i_0=1}^{\Theta}\sum_{j_0=1}^{\Theta}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\left[g\left(\boldsymbol{Y}_{\ell n_\mu},j_0\right)\mathbb{P}(\theta_{k+1}=j_0|\theta_k=i_0)-g\left(\boldsymbol{Y}_{\ell n_\mu},i_0\right)\right]I(\theta_k=i_0)\right]\\
&\quad=\mathbb{E}h\left(\boldsymbol{Y}(t_\iota),\theta(t_\iota):\iota\le\kappa_0\right)\left[\int_t^{t+u}Qg\left(\boldsymbol{Y}(v),\theta(v)\right)dv\right]
\end{aligned}\tag{49}$$

$$\begin{aligned}
&\lim_{\mu\to0}\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_\iota),\theta^{\mu}(t_\iota):\iota\le\kappa_0\right)\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\left[g\left(\boldsymbol{Y}_{\ell n_\mu+n_\mu},\theta_{\ell n_\mu+n_\mu}\right)-g\left(\boldsymbol{Y}_{\ell n_\mu},\theta_{\ell n_\mu+n_\mu}\right)\right]\right]\\
&\quad=\lim_{\mu\to0}\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_\iota),\theta^{\mu}(t_\iota):\iota\le\kappa_0\right)\\
&\quad\quad\times\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\delta_\mu\nabla'_{\widehat{\psi}}g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\left[\mathbf{f}\left(s_k,\widehat{\boldsymbol{\psi}}_k+\overline{\mathbf{X}}(\theta_k),X_k(s_k)\right)-\overline{\mathbf{X}}(\theta_k)\right]-\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\widehat{\boldsymbol{\psi}}_k\right]\right.\\
&\quad\quad\quad\left.+\delta_\mu\nabla'_{r}g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\left[X_k(s_k)-\overline{X}_{\min,k}\right]-\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}r_k\right]+\delta_\mu\nabla'_{\mathbf{z}}g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\mathbf{e}_{s_k}-\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\mathbf{z}_k\right]\right]
\end{aligned}\tag{50}$$

$$\begin{aligned}
&\lim_{\mu\to0}\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_\iota),\theta^{\mu}(t_\iota):\iota\le\kappa_0\right)\\
&\quad\times\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\delta_\mu\nabla'_{\widehat{\psi}}g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\sum_{\check{\theta}=1}^{\Theta}\sum_{\theta=1}^{\Theta}\mathbb{E}_{\ell n_\mu}\left\{I\left(\theta_k=\theta|\theta_{\ell n_\mu}=\check{\theta}\right)\left[\mathbf{f}\left(s_k,\widehat{\boldsymbol{\psi}}_k+\overline{\mathbf{X}}(\theta),X_k(s_k,\theta)\right)-\overline{\mathbf{X}}(\theta)\right]\right\}\right]\right]
\end{aligned}\tag{51}$$

for large k with $\ell n_\mu\le k<\ell n_\mu+n_\mu$ and $k-\ell n_\mu\to\infty$, by [28, Prop. 4.4], for some $\widehat{k}_0>0$

$$\begin{aligned}
(I+\varepsilon Q)^{k-\ell n_\mu}&=Z((k-\ell n_\mu)\varepsilon)+O(\varepsilon+\exp\left(-\widehat{k}_0(k-\ell n_\mu)\right)\\
\frac{dZ(t)}{dt}&=Z(t)Q,\ Z(0)=I.
\end{aligned}$$

For $\ell n_\mu\le k\le\ell n_\mu+n_\mu$, $\varepsilon=O(\mu)$ yields that $(k-\ell n_\mu)\varepsilon\to0$ as $\mu\to0$. For such k, $Z((k-\ell n_\mu)\varepsilon)\to I$. Therefore, by the boundedness of $\overline{\mathbf{X}}(\theta)$ and $\mathbf{f}_k$, it follows that, as $\mu\to0$

$$\begin{aligned}
&\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\left\|\mathbb{E}_{\ell n_\mu}\left\{\mathbf{f}\left(s_k,\widehat{\boldsymbol{\psi}}_k+\overline{\mathbf{X}}(\theta),X_k(s_k,\theta)\right)-\overline{\mathbf{X}}(\theta)\right\}\right\|\\
&\quad\times\left|\mathbb{E}_{\ell n_\mu}\left\{I(\theta_k=\theta)|I\left(\theta_{\ell n_\mu}=\check{\theta}\right)\right\}-I\left(\theta_{\ell n_\mu}=\check{\theta}\right)\right|\to0.
\end{aligned}$$

Therefore, (51) reduces to

$$\begin{aligned}
&\lim_{\mu\to0}\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_\iota),\theta^{\mu}(t_\iota):\iota\le\kappa_0\right)\\
&\times\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\delta_\mu\nabla'_{\widehat{\psi}}g\left[\sum_{\check{\theta}=1}^{\Theta}I\left(\theta_{\ell n_\mu}=\check{\theta}\right)\left[-\overline{\mathbf{X}}(\check{\theta})\right.\right.\right.\\
&\quad\left.\left.\left.+\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\mathbb{E}_{\ell n_\mu}\left\{\mathbf{f}\left(s_k,\widehat{\boldsymbol{\psi}}_k+\overline{\mathbf{X}}(\check{\theta}),X_k(s_k,\check{\theta})\right)\right\}\right]\right]\right].
\end{aligned}\tag{52}$$

It is more convenient to work with the individual elements of $\mathbf{f}(\cdot,\cdot,\cdot)$. Substituting for the ith element from (6) into the last line of (52), we obtain

$$\begin{aligned}
&\frac{1}{n_\mu}\sum_{\check{\theta}=1}^{\Theta}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}I\left(\theta_{\ell n_\mu}=\check{\theta}\right)\mathbb{E}_{\ell n_\mu}\\
&\quad\times\left\{\frac{X_k(i,\check{\theta})}{b_i^{\gamma}\left(\widehat{\boldsymbol{\psi}}_k+\overline{\mathbf{X}}(\check{\theta})\right)}\cdot I(s_k=i)\right\}\\
&\quad\quad=\frac{1}{n_\mu}\sum_{\check{\theta}=1}^{\Theta}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\mathbb{E}_{\ell n_\mu}\left\{X_k(i,\check{\theta})\right\}I\left(\theta_{\ell n_\mu}=\check{\theta}\right)
\end{aligned}\tag{53}$$

since $\mathbb{E}_{\ell n_\mu}\{I(s_k=i)\}=b_i^{\gamma}(\boldsymbol{\psi}_{\ell n_\mu})$. Recall that $\boldsymbol{\psi}_{\ell n_\mu}=\widehat{\boldsymbol{\psi}}_{\ell n_\mu}+\overline{\mathbf{X}}(\theta_{\ell n_\mu})$. However, in the last line of (52), $\theta_{\ell n_\mu}=\check{\theta}$ is held fixed. Therefore, $\boldsymbol{\psi}_{\ell n_\mu}=\widehat{\boldsymbol{\psi}}_{\ell n_\mu}+\overline{\mathbf{X}}(\check{\theta})$. Note further that $\theta_{\ell n_\mu}=\theta^{\mu}(\mu\ell n_\mu)$. In light of Assumptions 1–3, by the weak convergence of $\theta^{\mu}(\cdot)$ to $\theta(\cdot)$, the Skorohod representation, and using $\mu\ell n_\mu\to v$, it can be shown by combining (52) and (53) that, as $\mu\to0$

$$\begin{aligned}
&\frac{1}{n_\mu}\sum_{\check{\theta}=1}^{\Theta}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\mathbb{E}_{\ell n_\mu}\left\{X_k(i,\check{\theta})\right\}I\left(\theta^{\mu}(\mu\ell n_\mu)=\check{\theta}\right)\\
&\quad\to\sum_{\check{\theta}=1}^{\Theta}\overline{X}(i,\check{\theta})I\left(\theta(v)=\check{\theta}\right)=\overline{F}\left(i,\theta(v)\right)\ \text{in probability.}
\end{aligned}\tag{54}$$

$$\begin{aligned}&\lim_{\mu\to 0}\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_{\iota}),\theta^{\mu}(t_{\iota}):\iota\le\kappa_0\right)\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\frac{\delta_\mu}{n_\mu}\nabla'_{r}g\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\left[\sum_{\breve{\theta}=1}^{\Theta}\left[\sum_{i=1}^{S}\mathbb{E}_{\ell n_\mu}X_k(i,\breve{\theta})I(s_k=i)-\overline{X}_{\min}(\breve{\theta})\right]I\left(\theta_{\ell n_\mu}=\breve{\theta}\right)\right]\right]\\ &\quad=\lim_{\mu\to 0}\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_{\iota}),\theta^{\mu}(t_{\iota}):\iota\le\kappa_0\right)\\ &\quad\times\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\frac{\delta_\mu}{n_\mu}\nabla'_{r}g\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\left[\sum_{\breve{\theta}=1}^{\Theta}\left[\sum_{i=1}^{S}b_i^{\gamma}\left(\boldsymbol{\psi}_{\ell n_\mu}\right)\mathbb{E}_{\ell n_\mu}X_k(i,\breve{\theta})-\overline{X}_{\min}(\breve{\theta})\right]I\left(\theta_{\ell n_\mu}=\breve{\theta}\right)\right]\right]\end{aligned}\tag{57}$$

A similar argument for the first term in (52) yields

$$\begin{aligned}&\lim_{\mu\to 0}\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_{\iota}),\theta^{\mu}(t_{\iota}):\iota\le\kappa_0\right)\\ &\times\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\delta_\mu\nabla'_{\widehat{\psi}}g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\left[\mathbf{f}\left(s_k,\widehat{\boldsymbol{\psi}}_k+\overline{\mathbf{X}}(\theta_k),X_k(s_k)\right)\right.\right.\right.\\ &\qquad\left.\left.\left.-\overline{\mathbf{X}}(\theta_k)\right]\right]\right]\to\mathbf{0}_S.\end{aligned}\tag{55}$$

Using the technique of [11, Ch. 8], it can be shown that

$$\begin{aligned}&\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_{\iota}),\theta^{\mu}(t_{\iota}):\iota\le\kappa_0\right)\\ &\quad\times\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\delta_\mu\nabla'_{\widehat{\psi}}g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\widehat{\boldsymbol{\psi}}_k\right]\right]\\ &\quad\to\mathbb{E}h\left(\boldsymbol{Y}(t_{\iota}),\theta(t_{\iota}):\iota\le\kappa_0\right)\\ &\qquad\times\left[\int_t^{t+u}\nabla'_{\widehat{\psi}}g\left(\boldsymbol{Y}(v),\theta(v)\right)\widehat{\boldsymbol{\psi}}(v)dv\right]\quad\text{as }\mu\to 0.\end{aligned}\tag{56}$$

Step 2: Next, we concentrate on the second term in (50). By virtue of the boundedness of $X_k(i,\theta)$ (see Assumption 1), and using a similar argument as in Step 1, we have (57), shown at the top of the page. Here, we used $\mathbb{E}_{\ell n_\mu}\{I(s_k=i)\}=b_i^{\gamma}(\boldsymbol{\psi}_{\ell n_\mu})$ as in Step 1 of the proof. Recall that $\boldsymbol{\psi}_k=\widehat{\boldsymbol{\psi}}_k+\overline{\mathbf{X}}(\theta_k)$ from (18). Note further that $\boldsymbol{\psi}_{\ell n_\mu}=\boldsymbol{\psi}^{\mu}(\mu\ell n_\mu)$ and $\theta_{\ell n_\mu}=\theta^{\mu}(\mu\ell n_\mu)$. By the weak convergence of $\theta^{\mu}(\cdot)$ to $\theta(\cdot)$, the Skorohod representation, and using $\mu\ell n_\mu\to v$ and Assumptions 1–3, it can then be shown

$$\begin{aligned}&\sum_{\breve{\theta}=1}^{\Theta}\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}b_i^{\gamma}\left(\boldsymbol{\psi}^{\mu}(\mu\ell n_\mu)\right)\mathbb{E}_{\ell n_\mu}\left\{X_k(i,\breve{\theta})\right\}\\ &\qquad\times I\left(\theta^{\mu}(\mu\ell n_\mu)=\breve{\theta}\right)\\ &\quad\to b_i^{\gamma}\left(\widehat{\boldsymbol{\psi}}(v)+\overline{\mathbf{X}}\left(\theta(v)\right)\right)\overline{X}(i,\theta(v))\ \text{in prob. as }\mu\to 0.\end{aligned}\tag{58}$$

Using a similar argument for the second term in (57), we conclude that, as $\mu\to 0$, we have (59), shown at the top of the next page. Finally, similar to (56), we have

$$\begin{aligned}&\mathbb{E}h(\boldsymbol{Y}^{\mu}(t_{\iota}),\theta^{\mu}(t_{\iota}):\iota\le\kappa_0)\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\delta_\mu\nabla'_{r}g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}r_k\right]\right]\\ &\quad\to\mathbb{E}h\left(\boldsymbol{Y}(t_{\iota}),\theta(t_{\iota}):\iota\le\kappa_0\right)\\ &\qquad\times\left[\int_t^{t+u}\nabla'_{r}g\left(\boldsymbol{Y}(v),\theta(v)\right)\ r(v)dv\right]\quad\text{as }\mu\to 0.\end{aligned}\tag{60}$$

Step 3: Next, we concentrate on the last term in (50). Using similar arguments as in Step 1 and 2, we have

$$\begin{aligned}&\lim_{\mu\to 0}\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_{\iota}),\theta^{\mu}(t_{\iota}):\iota\le\kappa_0\right)\\ &\qquad\times\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\delta_\mu\nabla'_{\mathbf{z}}g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\sum_{i=1}^{S}\mathbf{e}_i\cdot\mathbb{E}_{\ell n_\mu}\{I(s_k=i)\}\right]\right]\\ &\quad=\lim_{\mu\to 0}\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_{\iota}),\theta^{\mu}(t_{\iota}):\iota\le\kappa_0\right)\\ &\qquad\times\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\delta_\mu\nabla'_{\mathbf{z}}g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\sum_{\breve{\theta}=1}^{\Theta}I\left(\theta_{\ell n_\mu}=\breve{\theta}\right)\right.\right.\\ &\qquad\qquad\left.\left.\times\left[\sum_{i=1}^{S}b_i^{\gamma}\left(\widehat{\boldsymbol{\psi}}_{\ell n_\mu}+\overline{\mathbf{X}}(\breve{\theta})\right)\cdot\mathbf{e}_i\right]\right]\right].\end{aligned}\tag{61}$$

Noting that $\theta_{\ell n_\mu}=\theta^{\mu}(\mu\ell n_\mu)$, by the weak convergence of $\theta^{\mu}(\cdot)$ to $\theta(\cdot)$, the Skorohod representation, and using $\mu\ell n_\mu\to v$, it can then be shown

$$\begin{aligned}&\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\sum_{\breve{\theta}=1}^{\Theta}I\left(\theta^{\mu}(\mu\ell n_\mu)=\breve{\theta}\right)\\ &\quad\times\left[\sum_{i=1}^{S}b_i^{\gamma}\left(\widehat{\boldsymbol{\psi}}^{\mu}(\mu\ell n_\mu)+\overline{\mathbf{X}}(\breve{\theta})\right)\cdot\mathbf{e}_i\right]\\ &\qquad\to\boldsymbol{b}^{\gamma}\left(\widehat{\boldsymbol{\psi}}(v)+\overline{\mathbf{X}}\left(\theta(v)\right)\right)\ \text{in probability as }\mu\to 0.\end{aligned}\tag{62}$$

Therefore, as $\mu\to 0$, we have

$$\begin{aligned}&\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_{\iota}),\theta^{\mu}(t_{\iota}):\iota\le\kappa_0\right)\\ &\qquad\times\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\delta_\mu\nabla'_{\mathbf{z}}g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\sum_{i=1}^{S}\mathbf{e}_{s_k}\right]\right]\\ &\quad\to\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_{\iota}),\theta^{\mu}(t_{\iota}):\iota\le\kappa_0\right)\\ &\qquad\times\left[\int_t^{t+u}\nabla'_{\mathbf{z}}g\left(\boldsymbol{Y}(v),\theta(v)\right)\ \boldsymbol{b}^{\gamma}\left(\widehat{\boldsymbol{\psi}}(v)+\overline{\mathbf{X}}\left(\theta(v)\right)\right)dv\right].\end{aligned}\tag{63}$$

Finally, similar to (56), we have

$$\begin{aligned}&\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_{\iota}),\theta^{\mu}(t_{\iota}):\iota\le\kappa_0\right)\\ &\qquad\times\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\delta_\mu\nabla'_{\mathbf{z}}g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\mathbf{z}_k\right]\right]\\ &\quad\to\mathbb{E}h\left(\boldsymbol{Y}(t_{\iota}),\theta(t_{\iota}):\iota\le\kappa_0\right)\\ &\qquad\times\left[\int_t^{t+u}\nabla'_{\mathbf{z}}g\left(\boldsymbol{Y}(v),\theta(v)\right)\ \mathbf{z}(v)dv\right]\quad\text{as }\mu\to 0.\end{aligned}\tag{64}$$

Combining Steps 1 to 3 concludes the proof.

$$\begin{aligned}&\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_\iota),\theta^{\mu}(t_\iota):\iota\le\kappa_0\right)\times\left[\sum_{\ell:\ell\delta_\mu=t}^{t+u}\delta_\mu\nabla'_r g\left[\frac{1}{n_\mu}\sum_{k=\ell n_\mu}^{\ell n_\mu+n_\mu-1}\left[X_k(s_k)-\overline{X}_{\min,k}\right]\right]\right]\\&\quad\to\mathbb{E}h\left(\boldsymbol{Y}^{\mu}(t_\iota),\theta^{\mu}(t_\iota):\iota\le\kappa_0\right)\left[\int_t^{t+u}\nabla'_r g\left(\boldsymbol{Y}(v),\theta(v)\right)\left[\boldsymbol{b}^{\gamma}\left(\widehat{\boldsymbol{\psi}}(v)+\overline{\mathbf{X}}\left(\theta(v)\right)\right)\cdot\overline{\mathbf{X}}\left(\theta(v)\right)-\overline{X}_{\min}\left(\theta(v)\right)\right]dv\right]\end{aligned}\tag{59}$$

Appendix B
Proof of Theorem 5.2

We first prove that each subsystem—associated with each $\theta\in\mathcal{M}$ when $\theta(t)=\theta$ is held fixed—is globally asymptotically stable, and $\mathbb{R}_{[0,\eta)}$ is its global attracting set. Define the Lyapunov function

$$V_\theta(r)=r^2.\tag{65}$$

Taking the time derivative, and applying (27), we obtain

$$\frac{d}{dt}V_\theta(r)=2r\cdot\left[\boldsymbol{b}^{\gamma}\left(\overline{\mathbf{X}}(\theta)\right)\cdot\overline{\mathbf{X}}(\theta)-\overline{X}_{\min}(\theta)-r\right].$$

Since the objective function is bounded across the feasible set for each $\theta\in\mathcal{M}$, we have

$$\frac{d}{dt}V_\theta(r)\le 2r\cdot\left[C(\gamma,\theta)-r\right]\tag{66}$$

for some constant $C(\gamma,\theta)$. Recall the smooth best-response sampling strategy $\boldsymbol{b}^{\gamma}(\cdot)$ in Definition 3.1. The parameter γ simply determines the magnitude of perturbations applied to the objective function. It is then clear that $C(\gamma,\theta)$ is monotonically increasing in γ.

In view of (66), for each $\eta>0$, $\widehat{\gamma}$ can be chosen small enough such that, if $\gamma\le\widehat{\gamma}$ and $r\ge\eta$

$$\frac{d}{dt}V_\theta(r)\le -V_\theta(r).\tag{67}$$

Therefore, each subsystem is globally asymptotically stable and, for $\gamma\le\widehat{\gamma}$

$$\lim_{t\to\infty}d\left(r(t),\mathbb{R}_{[0,\eta)}\right)=0.$$

Finally, stability of the regime-switching ODE (23) is examined. We can use the above Lyapunov function to extend [15, Cor. 12] to prove global asymptotic stability w.p.1.

Theorem B. 1 [15, Cor. 12]: Consider the switched system

$$\begin{aligned}&\dot{\mathbf{w}}(t)=f\left(\mathbf{w}(t),\theta(t)\right)\\&\mathbf{w}(0)=\mathbf{w}_0,\ \theta(0)=\theta_0,\ \mathbf{w}(t)\in\mathbb{R}^r,\ \theta(t)\in\mathcal{M}\end{aligned}\tag{68}$$

where $\theta(t)$ is the state of a continuous time Markov chain with generator Q. Define

$$\overline{q}:=\max_{\theta\in\mathcal{M}}|q_{\theta\theta}|,\ \text{ and }\ \widetilde{q}:=\max_{\theta,\theta'\in\mathcal{M}}q_{\theta\theta'}.$$

Suppose there exist continuously differentiable functions $V_\theta:\mathbb{R}^n\to\mathbb{R}^+$, for each $\theta\in\mathcal{M}$, and class $\mathcal{K}_\infty$ functions $a_1,a_2:\mathbb{R}^+\to\mathbb{R}^+$ such that the following hold for some $\mathcal{H}\subset\mathbb{R}^r$:

1) $V_\theta(\mathbf{w})>0$ for all $\mathbf{w}\notin\mathcal{H}$;
2) $a_1(d(\mathbf{w},\mathcal{H}))\le V_\theta(\mathbf{w})\le a_2(d(\mathbf{w},\mathcal{H}))$, $\forall\mathbf{w}\in\mathbb{R}^r,\theta\in\mathcal{M}$;
3) $[\nabla V_\theta]'f(\mathbf{w},\theta)\le-\lambda V_\theta(\mathbf{w})$, $\forall\,\mathbf{w}\in\mathbb{R}^r$, $\forall\,\theta\in\mathcal{M}$;
4) $V_\theta(\mathbf{w})\le vV_{\theta'}(\mathbf{w})$, $\forall\,\mathbf{w}\in\mathbb{R}^r,\theta,\theta'\in\mathcal{M}$;
5) $(\lambda+\widetilde{q})/\overline{q}>v>1$.

Then, the regime-switching system (68) is globally asymptotically stable almost surely, and $\mathcal{H}$ constitutes its global attractor set.

The quadratic Lyapunov functions (65) satisfies hypothesis 2 in Theorem B.1; see (67). Further, since the Lyapunov functions are the same for all subsystems $\theta\in\mathcal{M}$, existence of $v>1$ in hypothesis 3 is automatically guaranteed. Hypothesis 4 simply ensures that the switching signal $\theta(t)$ is slow enough. Given that $\lambda=1$ in hypothesis 2, it remains to ensure that the generator Q of Markov chain $\theta(t)$ satisfies $1+\widetilde{q}>\overline{q}$. This is satisfied since, in view of (13)}, $|q_{\theta\tilde{\theta}}|\le 1$ for all $\theta,\tilde{\theta}\in\mathcal{M}$.

References

[1] V. I. Norkin, Y. M. Ermoliev, and A. Ruszczyński, "On optimal allocation of indivisibles under uncertainty," *Oper. Res.*, vol. 46, no. 3, pp. 381–395, 1998.

[2] J. R. Swisher, P. D. Hyden, S. H. Jacobson, and L. W. Schruben, "A survey of simulation optimization techniques and procedures," in *Proc. Winter Simulation Conf.*, 2000, vol. 1, pp. 119–128.

[3] K. Park and Y. Lee, "An on-line simulation approach to search efficient values of decision variables in stochastic systems," *Int. J. Adv. Manuf. Technol.*, vol. 25, no. 11/12, pp. 1232–1240, 2005.

[4] V. Krishnamurthy, X. Wang, and G. Yin, "Spreading code optimization and adaptation in CDMA via discrete stochastic approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1927–1949, Sep. 2004.

[5] I. Berenguer, X. Wang, and V. Krishnamurthy, "Adaptive MIMO antenna selection via discrete stochastic optimization," *IEEE Trans. Signal Process.*, vol. 53, no. 11, pp. 4315–4329, Nov. 2005.

[6] M. C. Fu, "Optimization via simulation: A review," *Ann. Oper. Res.*, vol. 53, no. 1, pp. 199–247, 1994.

[7] L. J. Hong and B. L. Nelson, "Discrete optimization via simulation using COMPASS," *Oper. Res.*, vol. 54, no. 1, pp. 115–129, 2006.

[8] G. C. Pflug, *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*. Norwell, MA, USA: Kluwer Academic, 1996.

[9] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*. Cambridge, MA, USA: MIT Press, 1998.

[10] A. Benveniste, M. Metivier, and P. Prioret, *Adaptive Algorithms and Stochastic Approximations*. New York, NY, USA: Springer-Verlag, 1990.

[11] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. New York, NY, USA: Springer-Verlag, 2003.

[12] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*. New York, NY, USA: Springer-Verlag, 1998.

[13] G. Yin and C. Zhu, *Hybrid Switching Diffusions: Properties and Applications*. New York, NY, USA: Springer-Verlag, 2009.

[14] D. Chatterjee and D. Liberzon, "Stability analysis of deterministic and stochastic switched systems via a comparison principle and multiple Lyapunov functions," *SIAM J. Control Optim.*, vol. 45, no. 1, pp. 174–206, 2007.

[15] D. Chatterjee and D. Liberzon, "On stability of randomly switched nonlinear systems," *IEEE Trans. Autom. Control*, vol. 52, no. 12, pp. 2390–2394, Dec. 2007.

[16] S. Andradóttir, "An overview of simulation optimization via random search," *Handbooks Oper. Res. Management Sci.*, vol. 13, pp. 617–631, 2006.

[17] R. Y. Rubinstein and A. Shapiro, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. Chichester, U.K.: Wiley, 1993.

[18] H. Chen and B. W. Schmeiser, "Stochastic root finding via retrospective approximation," *IIE Trans.*, vol. 33, no. 3, pp. 259–275, 2001.

[19] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM J. Optim.*, vol. 12, no. 2, pp. 479–502, 2002.

[20] T. Homem-De-Mello, "Variable-sample methods for stochastic optimization," *ACM Trans. Model. Comput. Simul.*, vol. 13, no. 2, pp. 108–133, 2003.

[21] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, Control*. Hoboken, NJ, USA: Wiley, 2003.

[22] W. J. Gutjahr and G. C. Pflug, "Simulated annealing for noisy cost functions," *J. Global Optim.*, vol. 8, no. 1, pp. 1–13, 1996.

[23] A. A. Prudius and S. Andradóttir, "Averaging frameworks for simulation optimization with applications to simulated annealing," *Naval Res. Logistics*, vol. 59, no. 6, pp. 411–429, 2012.

[24] F. Glover and M. Laguna, "Tabu search," in *Encyclopedia of Operations Research and Management Science*, S. I. Gass and C. M. Harris, Eds. New York, NY, USA: Springer, 1996, pp. 671–701.

[25] M. H. Alrefaei and S. Andradóttir, "Discrete stochastic optimization using variants of the stochastic ruler method," *Naval Res. Logistics*, vol. 52, no. 4, pp. 344–360, 2005.

[26] S. Andradóttir, "A global search method for discrete stochastic optimization," *SIAM J. Optim.*, vol. 6, no. 2, pp. 513–530, 1996.

[27] S. Andradóttir, "Accelerating the convergence of random search methods for discrete stochastic optimization," *ACM Trans. Model. Comput. Simul.*, vol. 9, no. 4, pp. 349–380, 1999.

[28] G. Yin, V. Krishnamurthy, and C. Ion, "Regime switching stochastic approximation algorithms with application to adaptive discrete stochastic optimization," *SIAM J. Optim.*, vol. 14, no. 4, pp. 1187–1215, 2004.

[29] S. Andradóttir and A. A. Prudius, "Balanced explorative and exploitative search with estimation for simulation optimization," *INFORMS J. Comput.*, vol. 21, no. 2, pp. 193–208, Spring 2009.

[30] V. I. Norkin, G. C. Pflug, and A. Ruszczyński, "A branch and bound method for stochastic global optimization," *Math. Prog.*, vol. 83, no. 1–3, pp. 425–450, 1998.

[31] L. Shi and S. Ólafsson, "Nested partitions method for global optimization," *Oper. Res.*, vol. 48, no. 3, pp. 390–407, 2000.

[32] M. Zlochin, M. Birattari, N. Meuleau, and M. Dorigo, "Model-based search for combinatorial optimization: A critical survey," *Ann. Oper. Res.*, vol. 131, no. 1–4, pp. 373–395, 2004.

[33] R. Rubinstein, "The cross-entropy method for combinatorial and continuous optimization," *Methodol. Comput. Appl. Prob.*, vol. 1, no. 2, pp. 127–190, 1999.

[34] R. Y. Rubinstein and D. P. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning*. New York, NY, USA: Springer Science & Business Media, 2013.

[35] M. Dorigo, G. Di Caro, and L. M. Gambardella, "Ant algorithms for discrete optimization," *Artif. l Life*, vol. 5, no. 2, pp. 137–172, 1999.

[36] M. Dorigo, M. Birattari, and T. Stützle, "Ant colony optimization," *IEEE Comput. Intell. Mag.*, vol. 1, no. 4, pp. 28–39, 2006.

[37] J. Hu, M. C. Fu, and S. I. Marcus, "A model reference adaptive search method for global optimization," *Oper. Res.*, vol. 55, no. 3, pp. 549–568, 2007.

[38] E. Zhou and J. Hu, "Gradient-based adaptive stochastic search for non-differentiable optimization," *IEEE Trans. Autom. Control*, vol. 59, no. 7, pp. 1818–1832, Jul. 2014.

[39] P. Larranaga and J. A. Lozano, Eds., *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Norwell, MA, USA: Kluwer Academic, 2002.

[40] M. Hauschild and M. Pelikan, "An introduction and survey of estimation of distribution algorithms," *Swarm Evol. Comput.*, vol. 1, no. 3, pp. 111–128, 2011.

[41] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multi-armed bandit problem," *Mach. Learn.*, vol. 47, no. 2/3, pp. 235–256, 2002.

[42] J.-Y. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits," in *Proc. 23th Conf. Learning Theory*, 2010, pp. 41–53.

[43] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Algorithmic Learning Theory*, ser. Lecture Notes in Computer Science, J. Kivinen, C. Szepesvári, E. Ukkonen, and T. Zeugmann, Eds. Berlin, Germany: Springer-Verlag, 2011, vol. 6925, pp. 174–188.

[44] A. Garcia, D. Reaume, and R. L. Smith, "Fictitious play for finding system optimal routings in dynamic traffic networks," *Transp. Res. B*, vol. 34, no. 2, pp. 147–156, 2000.

[45] T. J. Lambert Iii, M. A. Epelman, and R. L. Smith, "A fictitious play approach to large-scale optimization," *Oper. Res.*, vol. 53, no. 3, pp. 477–489, 2005.

[46] S. Yakowitz, P. L'ecuyer, and F. J. Vázquez-Abad, "Global stochastic optimization with low-dispersion point sets," *Oper. Res.*, vol. 48, no. 6, pp. 939–950, 2000.

[47] R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu, "Sample size selection in optimization methods for machine learning," *Math. Prog.*, vol. 134, no. 1, pp. 127–155, 2012.

[48] J. Linderoth, A. Shapiro, and S. Wright, "The empirical behavior of sampling methods for stochastic programming," *Ann. Oper. Res.*, vol. 142, no. 1, pp. 215–241, 2006.

[49] J. Hofbauer and W. H. Sandholm, "On the global convergence of stochastic fictitious play," *Econometrica*, vol. 70, no. 6, pp. 2265–2294, 2002.

[50] D. Fudenberg and D. K. Levine, "Conditional universal consistency," *Game Econ. Behav.*, vol. 29, no. 1/2, pp. 104–130, 1999.

[51] W. B. Powell and I. O. Ryzhov, *Optimal Learning*. Hoboken, NJ, USA: Wiley, 2012.

[52] D. Fudenberg and D. Levine, "Consistency and cautious fictitious play," *J. Econ. Dyn. Control*, vol. 19, no. 5–7, pp. 1065–1089, 1995.

[53] H. J. Kushner and J. Yang, "Analysis of adaptive step-size SA algorithms for parameter tracking," *IEEE Trans. Autom. Control*, vol. 40, no. 8, pp. 1403–1410, Aug. 1995.

[54] A. Blum and Y. Mansour, "From external to internal regret," *J. Mach. Learn. Res.*, vol. 8, no. 6, pp. 1307–1324, 2007.

[55] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Trans. Autom. Control*, vol. AC-22, no. 4, pp. 551–575, 1977.

[56] J. Hannan, "Approximation to bayes risk in repeated play," *Contributions to the Theory of Games*, vol. 3, pp. 97–139, 1957.

[57] P. Billingsley, *Convergence of Probability Measures*. New York, NY, USA: Wiley, 1968.

[58] D. Liberzon, *Switching in Systems and Control*. Boston, MA, USA: Birkhäuser, 2003.

[59] M. Benaïm and M. Faure, "Consistency of vanishingly smooth fictitious play," *Math. Oper. Res.*, vol. 38, no. 3, pp. 437–450, 2013.

[60] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes With Application to Stochastics Systems Theory*. Cambridge, MA, USA: MIT Press, 1984.

Omid Namvar Gharehshiran received the Ph.D. degree in statistical signal processing from the University of British Columbia, Vancouver, BC, Canada, in 2015. He is currently holding a NSERC Postdoctoral Fellowship at the Actuarial Science and Mathematical Finance group, Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada. His research interests span stochastic optimization and control, games and learning theory.

Vikram Krishnamurthy (S'90–M'91–SM'99–F'05) received the bachelor's degree from the University of Auckland, New Zealand, in 1988, and the Ph.D. in systems engineering from the Australian National University, Canberra, Australia, in 1992.

He is currently a Professor and holds the Canada Research Chair at the Department of Electrical Engineering, University of British Columbia, Vancouver, BC, Canada. His research interests include statistical signal processing, computational game theory, and stochastic control in social networks.

Dr. Krishnamurthy has served as Distinguished Lecturer for the IEEE Signal Processing Society and Editor-in-Chief of the IEEE Journal Selected Topics in Signal Processing. He received an honorary doctorate from KTH (Royal Institute of Technology), Sweden in 2013.

George Yin (F'02) received the B.S. degree in mathematics from the University of Delaware, Newark, DE, USA, in 1983 and the M.S. degree in electrical engineering and Ph.D. degree in applied mathematics from Brown University, Providence, RI, USA, in 1987.

He joined Wayne State University in 1987, and became a Professor in 1996. His research interests include stochastic systems and applications.

Dr. Yin has severed on the IFAC Technical Committee on Modeling, Identification and Signal Processing, and many conference program committees; he was Co-Chair of the SIAM Conference on Control and Its Application, 2011, and Co-Chair of a couple of AMS-IMS-SIAM Summer Research Conferences; he also chaired a number of SIAM prize selection committees. He is Chair of the SIAM Activity Group on Control and Systems Theory, and serves on the Board of Directors of the American Automatic Control Council. He is an Associate Editor of the *SIAM Journal on Control and Optimization*, and on the editorial board of a number of other journals. He was an Associate Editor of *Automatica* (2005–2011) and the IEEE Transactions on Automatic Control (1994–1998). He is a Fellow of SIAM.