Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Novel Framework for Learning Automata: A Statistical Hypothesis Testing Approach

CHONG DI[1], SHENGHONG LI[1], (Senior Member, IEEE), FANGQI LI[1] and KAIYUE QI [2]

[1]School of Cyber Security, School of Electronic Information and Electrical Engineering, Artificial Intelligence Institute, Shanghai Jiao Tong University, Shanghai 200240, China.

[2]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Kaiyue Qi (e-mail: tommy-qi@sjtu.edu.cn).

This research work is funded by the National Key Research and Development Project of China (2016YFB0801003).

ABSTRACT Learning automaton (LA), a powerful tool in reinforcement learning, is of crucial importance for its adaptivity in the stochastic environment and its applicability in various engineering fields. In particular, the LA adaptively explores the optimal action that maximizes the reward among all possible choices by interacting with the environment. However, the traditional frameworks for LA have several limitations in practical applications, e.g., the cost of parameter tuning and predicaments in massive-action environments, preventing them from being applied to time-sensitive and resources-restricted tasks. In this paper, we propose a novel LA framework based on the statistical hypothesis testing, where the actions are compared by statistical hypothesis iteratively and the suboptimal ones are dismissed, and the estimated optimal action is attained. Apart from the proposal, the theoretical analyses for the framework are given to reveal its ϵ-optimality. The proposed framework also features efficiency in massive-action environments and the parameter-free property. Comprehensive simulations are conducted in both benchmark and massive-action environments to demonstrate the superiority of the proposed framework over the ordinary schemes.

INDEX TERMS Learning automata, reinforcement learning, statistical inference, parameter-free

I. INTRODUCTION

Reinforcement learning (RL), one of the most fruitful branches in machine learning (ML), concerns with how agents ought to behave in a stochastic environment to maximize the cumulative rewards [1]- [4]. Learning automaton (LA) is a powerful tool in RL. It can adaptively explore the optimal action that maximizes the reward among all possible choices by interacting with the environment [5]. Early studies of LA date back to 1960s [6], but it has surged much-renewed interest owing to its modern application in a broad range of engineering contexts such as pattern recognition, signal processing, function optimization, and assignment problems (see, e.g. [7]- [16] and references therein).

As shown in Figure 1 [5], in the t-th interaction with the random environment, an LA selects an action $a(t)$ to interact with the environment and gets evaluative feedback $\beta(t)$, which is then utilized to update the internal state of the LA. The LA converges to the final state after a number of interactions, i.e., it learns the optimal action to interact with the given environment. LA can be divided into various genres according to their transition functions between internal states and constitutions of the action sets. From the transition functions, LA can be characterized as fixed structure stochastic automaton (FSSA) or variable structure stochastic automa-

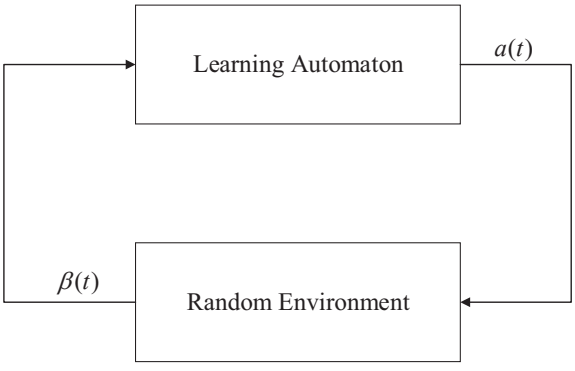


FIGURE 1. The interaction between an LA and the random environment.

ton (VSSA) [17]. The transition between states of FSSA is deterministic, while that of VSSA is stochastic. FSSA is the prototype of LA, and VSSA improves FSSA by being more flexible and having broader application scenarios. From the perspective of the action set, LA can be classified into continuous action set learning automata (CALA) and finite action set learning automata (FALA) [18]. The action set in CALA is an interval with uncountable elements, thus CALA is usually used for function optimization [19]. Meanwhile, FALA, whose action set is at most countable, is more extensively studied due to its numerous applications [20]. Besides, the random environment can be divided into P-model, Q-model or S-model depending on the type of its feedback [5]. The feedback is a binary value in $\{0, 1\}$ for P-model environments, a specific value in $\{\beta^1, \beta^2, \cdots, \beta^Q\}, (Q > 2)$ for Q-model environments, and an arbitrary value in $[0, 1]$ for S-model environments. As a typical case, the learning automata with a stochastic state transition function and a finite action set (VSFALA) in P-model environment have raised great attention from researchers, and plentiful theoretical achievements and learning schemes for VSFALA have emerged in the last decades [21].

One of the attractive properties of VSFALA is the ϵ-optimality which ensures that a VSFALA can converge to the optimal action with probability one as the number of interactions with the environment approaches to infinity [5]. In literature, the performance of various VSFALA schemes is evaluated by the convergence rate on the premise of a certain accuracy [22]. The accuracy is defined as the probability of a correct convergence, i.e., the probability for an LA to find the action with the highest reward probability. The convergence rate is the average times of iterations for an LA to converge correctly. The complexity of an LA scheme is usually measured by the convergence time which is defined as the consumed time on a device for an LA to converge correctly.

The framework for VSFALA was first studied in [17], where a popular learning process based on the action probability vector $\mathbf{P}$ was proposed. For an action set with r actions, $\mathbf{P}$ has r non-negative components $p_i, i = 1, 2, \cdots, r$ whose sum is one, where p_i represents the probability that an LA chooses the i-th action. The framework consists of three phases: (1) selecting an action according to the action probability vector; (2) interacting with the environment and getting the feedback; (3) updating the action probability vector. The VSFALA gets converged when the maximum action probability in $\mathbf{P}$ is greater than a predefined threshold $\mathcal{V}$, i.e., $\max_i \{p_i\} \geq \mathcal{V}$. Adhering to the framework, various schemes are proposed to speed up the convergence of VSFALA. Particularly, the family of estimator algorithms [22]- [27] that exploits historical information about the environment to guide the update is the most prevalent one. Notwithstanding the considerable improvements on the convergence rate of LA, almost all update strategies of $\mathbf{P}$ including the estimator algorithms suffer from two major limitations in practical applications:

- The parameter tuning cost. The majorities of ordinary schemes are based on the trilogy with multiple tunable parameters. So extra efforts are necessary to realize the trade-off between the accuracy and the convergence rate in a specific environment. Most traditional schemes are parameter-sensitive, and the cost of parameter tuning can be extremely expensive [28]. In practical applications, especially where interacting with the environment could be expensive, the enormous cost for parameter tuning is intimidating.
- The massive-action environments predicament. In massive-action environments, the inadequacy of the action probability vector based frameworks appears in two stages: the action selection and the action probability vector update. In the action selection stage, the time complexity of selecting an action by the action probability vector $\mathbf{P}$ is $\mathcal{O}(r)$, where r is the total number of actions. For massive-action environments where r is large, this stage can be extremely time-consuming. In the action probability vector update stage, the step size Δ for updating $\mathbf{P}$ is obtained by $\Delta = \frac{1}{rN}$, where $N \in \mathbb{N}^*$ is a resolution parameter, and it is evident that the step size declines inversely with the increase of the number of actions. Therefore in massive-action scenarios, the step size is small, which may slow down the convergence.

Several parameter-free schemes have been proposed in recent years to address the problem of parameter tuning. The parameter-free concept, which is first presented in [29], indicates that a set of parameters can be universally applied to all environments without further tuning. The most representative parameter-free schemes are the parameter-free LA (PFLA) [28] and loss function-based LA (LFPLA) [30]. However, both schemes are supported by time-consuming and computing resources-consuming Monte-Carlo simulations [31], preventing the schemes from being applied to time-sensitive and resources-restricted tasks. Thus, the established parameter-free schemes are far from perfect. In the case of massive-action environments predicaments, Zhang et al. [32] present a fast LA (FLA) scheme that uses a programming trick to realize the efficient action selection by $\mathbf{P}$ and the immediate update of $\mathbf{P}$. However, the FLA scheme only reduces the time complexity of $\mathbf{P}$ based estimator algorithms, while the totality of iterations remains high. So the challenges in massive-action environments are still confronting VSFALA. In view of this, we propose a revolutionary LA framework that enjoys the desired parameter-free property as well as the high efficiency. The contributions of this paper are as follows.

1) We present a novel framework for LA without the action probability vector, it is parameter-free and is efficient in massive-action environments. The framework is based on a well-known method in statistical inference, the statistical hypothesis testing. To the best of our knowledge, this is the first usage of the statistical hypothesis testing

in the field of LA.
2) The three constitutes in the proposed framework show promising stability and efficiency, they are: the action selecting strategy, the action set adjusting strategy, and the convergence judgment.
3) The performance of the proposed framework is analyzed theoretically and rigorous proof of the ϵ-optimality is provided.
4) Comprehensive comparisons in benchmark environments and massive-action environments are given to validate the theoretical analyses and demonstrate that the proposed framework outperforms the action probability vector based ones.

The rest of this paper is organized as follows. Section II is dedicated to the problem formulation and the review of the action probability vector based framework. In Section III, we present the statistical hypothesis testing based framework and give the theoretical analyses, including the estimation of the convergence rate and the proof of ϵ-optimality. Section IV provides the experimentations that verified the advantages of the proposed framework. Finally, Section V concludes this paper.

II. PRELIMINARIES

A. STATEMENT OF THE PROBLEM

The VSFALA problem concerns an automaton with finite actions interacting with a P-model environment and is formulated as a triple $< A, B, D >$, where

- $A = \{a_1, a_2, \cdots, a_r\}$ is the finite set of actions. The action selected at the t-th iteration is denoted by $a(t) \in A$.
- $B = \{0, 1\}$ is the set of feedback to the automaton, where 0 and 1 denote a penalty and a reward respectively. The feedback at the t-th iteration is denoted by $\beta(t) \in B$.
- $D = \{d_1, d_2, \cdots, d_r\}$ is the set of reward probabilities, such

$$\Pr\{\beta(t) = 1 | a(t) = a_i\} = d_i. \tag{1}$$

When D is independent of time, the stochastic environment is stationary. Otherwise, it is a non-stationary environment with D a function of t. Since the performance of an LA in a stationary environment forms the basis of further generalization to non-stationary environments, the following discussions focus on the stationary environment. Without special notation, the environment mentioned in the following sections represents the P-model stationary environment.

By interacting with the environment, the LA aims to identify the optimal action a_m with the highest reward probability d_m, where

$$m = \arg \max_{i=1,2,\cdots,r} \{d_i\}. \tag{2}$$

B. ACTION PROBABILITY VECTOR BASED FRAMEWORK

The traditional framework for VSFALA is based on an action probability vector which is defined as $\mathbf{P}(t) = [p_1(t), p_2(t), \cdots, p_r(t)]$, where $p_i(t)$ denotes the probability of selecting the action a_i at the t-th iteration. The three-folded learning process of the $\mathbf{P}$ based frameworks is formally summarized as follows:

1) At the t-th iteration, an action $a(t)$ is selected according to the probability distribution of $\mathbf{P}(t)$, where

$$\Pr\{a(t) = a_i\} = p_i(t). \tag{3}$$

2) The environment receives the action and provides feedback $\beta(t)$, where

$$\begin{cases} \Pr\{\beta(t) = 1 | a(t) = a_i\} = d_i \\ \Pr\{\beta(t) = 0 | a(t) = a_i\} = 1 - d_i \end{cases}. \tag{4}$$

3) The action probability vector $\mathbf{P}(t)$ is updated:

$$\mathbf{P}(t+1) = \mathscr{T}\{\mathbf{P}(t), a(t), \beta(t)\}, \tag{5}$$

where $\mathscr{T}$ is the update strategy that characterizes different schemes.

The most frequently used update strategy is the reward-inaction strategy $\mathscr{T}_{\text{RI}}$ which is proposed in [33]. Suppose $a(t) = a_i$, $\mathscr{T}_{\text{RI}}$ is defined as below.

$$\text{IF } \beta(t) = 1 \text{ THEN } \begin{cases} p_j(t+1) = \max\{p_j(t) - \Delta, 0\}, \forall j \neq i \\ p_i(t+1) = 1 - \sum\limits_{j=1, j\neq m}^{r} p_j(t+1) \end{cases}. \tag{6}$$

Where $\Delta = \frac{1}{rN}$ is the step of modification. $\mathscr{T}_{\text{RI}}$ has been widely adopted in estimator algorithms such as the discretized pursuit scheme (DP_{RI}) [22], the stochastic estimator algorithm (SE_{RI}) [23], the discretized bayesian pursuit algorithm (DBPA) [24] and the optimal budget allocation based LA (LA_{OCBA}) [25]. Beyond the $\mathscr{T}_{\text{RI}}$ strategy, several variants have been proposed to improve the convergence rate, such as the generalized pursuit strategy $\mathscr{T}_{\text{GP}}$ in discretized generalized pursuit LA scheme (DGPA) [26] and the last-position elimination strategy $\mathscr{T}_{\text{LE}}$ in last-position elimination-based LA scheme (LELA) [27]. All schemes discussed above have been proved to be ϵ-optimal in stationary environments. For the detailed descriptions and the corresponding theoretical analyses, we suggest that [22]- [27] be referred.

III. THE PROPOSED FRAMEWORK

In this section, we propose the statistical hypothesis testing approach based framework (LA_{SHT}) for VSFALA. In the proposed framework, the actions that are judged to be inferior by the hypothesis testing would be eliminated from A adaptively. At length, the only action left is the optimal action chosen by LA_{SHT}. In this section, the necessary knowledge of the statistical hypothesis is reviewed firstly, followed by the proposal of LA_{SHT} and the theoretical analyses.

A. STATISTICAL HYPOTHESIS TESTING IN VSFALA

Suppose the action a_i is selected for the k-th time during the t-th iteration, the feedback follows (4):

$$\begin{cases} \Pr\{\beta_i(k) = 1\} = \Pr\{\beta(t) = 1 | a(t) = a_i\} = d_i \\ \Pr\{\beta_i(k) = 0\} = \Pr\{\beta(t) = 0 | a(t) = a_i\} = 1 - d_i \end{cases}, \tag{7}$$

where $\beta_i(k)$ denotes the feedback for the k-th time of choosing a_i. The reaction from the environment is a Bernoulli trial and the feedback $\beta_i(t)$ is a random variable following the Bernoulli distribution with parameter d_i. In a stationary environment, if an action a_i is picked for n times then the feedback $\beta_i(1), \beta_i(2), \cdots, \beta_i(n)$ are independent, identically distributed (i.i.d.) random variables. Therefore the sum of $\beta_i(k)$ follows a binomial distribution with parameters n and d_i:

$$\sum_{k=1}^{n} \beta_i(k) \sim B(n, d_i). \tag{8}$$

Theorem 1. *Suppose an action a_i has been selected for n times, the feedback sequence is $\{\beta_i(1), \beta_i(2), \cdots, \beta_i(n)\}$, $\Pr\{\beta_i(k) = 1\} = d_i$ and $\Pr\{\beta_i(k) = 0\} = 1 - d_i$ for $k = 1, 2, \cdots, n$. Then the following convergence in probability holds:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \beta_i(k) = d_i. \tag{9}$$

Proof. Because $\beta_i(k)$ is a random variable following the Bernoulli distribution with parameter d_i, we have the expectation $\mathrm{E}[\beta_i(k)]$ and the variance $\mathrm{Var}[\beta_i(k)]$:

$$\mathrm{E}[\beta_i(k)] = d_i, \quad \mathrm{Var}[\beta_i(k)] = d_i(1 - d_i). \tag{10}$$

Consider the mean of the feedback sequence $\bar{\beta}_i = \frac{1}{n} \sum_{k=1}^{n} \beta_i(k)$, then,

$$\mathrm{E}[\bar{\beta}_i] = d_i, \quad \mathrm{Var}[\bar{\beta}_i] = \frac{1}{n^2} \sum_{k=1}^{n} \mathrm{Var}[\beta_i(k)] = \frac{1}{n} d_i(1 - d_i). \tag{11}$$

According to Chebyshev's Inequality, $\forall \epsilon > 0$,

$$\Pr\{|\bar{\beta}_i - \mathrm{E}[\bar{\beta}_i]| \geq \epsilon\} \leq \frac{\mathrm{Var}[\bar{\beta}_i]}{\epsilon^2} = \frac{d_i(1 - d_i)}{n\epsilon^2}. \tag{12}$$

This implies

$$\lim_{n\to\infty} \Pr\{|\bar{\beta}_i - \mathrm{E}[\bar{\beta}_i]| \geq \epsilon\} = 0, \tag{13}$$

which completes the proof. □

Theorem 1 indicates that if action a_i is chosen for infinite times, then the mean of the feedback sequence is the reward probability.

Lemma 1. *Suppose $\{X_n\}$ is a sequence of i.i.d random variables, then $\forall z \in \mathbb{R}$,*

$$\lim_{n\to\infty} \Pr\left\{ \frac{1}{\sqrt{n\mathrm{Var}[X_n]}} \left(\sum_{k=1}^{n} X_k - n\mathrm{E}[X_n] \right) \leq z \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx. \tag{14}$$

Lemma 1 is the famous Lévy-Lindberg central limit theorem whose proof can be found in [34].

Corollary 1. *Suppose an action a_i has been selected for n times, the feedback sequence is $\{\beta_i(1), \beta_i(2), \cdots, \beta_i(n)\}$, $\Pr\{\beta_i(k) = 1\} = d_i$ and $\Pr\{\beta_i(k) = 0|\} = 1 - d_i$. Then,*

$$\lim_{n\to\infty} \bar{\beta}_i \sim \mathcal{N}\left(d_i, \frac{d_i(1 - d_i)}{n} \right), \tag{15}$$

where $\bar{\beta}_i$ is the mean of the random feedback sequence $\{\beta_i(1), \beta_i(2), \cdots, \beta_i(n)\}$.

Proof. Since $\{\beta_i(1), \beta_i(2), \cdots, \beta_i(n)\}$ are i.i.d. random variables, then $\mathrm{E}[\beta_i(k)] = d_i$, $\mathrm{Var}[\beta_i(k)] = d_i(1 - d_i)$ from the proof of Theorem 1, and by Theorem 1,

$$\lim_{n\to\infty} \Pr\left\{ \frac{\sum_{k=1}^{n} \beta_i(k) - nd_i}{\sqrt{nd_i(1 - d_i)}} \leq z \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx. \tag{16}$$

This implies

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} \beta_i(k) - nd_i}{\sqrt{nd_i(1 - d_i)}} = \lim_{n\to\infty} \frac{\frac{1}{n} \sum_{k=1}^{n} \beta_i(k) - d_i}{\sqrt{\frac{d_i(1-d_i)}{n}}} \sim \mathcal{N}(0, 1), \tag{17}$$

which is tantamount to

$$\lim_{n\to\infty} \bar{\beta}_i = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \beta_i(k) \sim \mathcal{N}\left(d_i, \frac{d_i(1 - d_i)}{n} \right). \tag{18}$$

□

Corollary 1 indicates that when action a_i has been selected for n times, $n \to \infty$, then the statistic $\bar{\beta}_i$ is subject to a Gaussian distribution $\mathcal{N}(d_i, \frac{d_i(1-d_i)}{n})$. It can be deduced that the reward probabilities of two actions can be compared as the means of two Gaussian random variables. Traditionally, Z-test [35] is capable of distinguishing the means of two Gaussian random variables given two individual sampling sequences. Roughly, in Z-test, a z-value is computed from two sequences, and the conclusion about whether two means are the same is drawn with significance level α.

Formally, consider the feedback sequence $\{\beta_i(1), \beta_i(2), \cdots, \beta_i(n_i)\}$ for action a_i and the feedback sequence $\{\beta_j(1), \beta_j(2), \cdots, \beta_j(n_j)\}$ for action a_j, where $n_i = n_j = n$, Corollary 1 implies

$$\lim_{n\to\infty} \bar{\beta}_i \sim \mathcal{N}(d_i, \frac{d_i(1 - d_i)}{n}), \tag{19}$$

$$\lim_{n\to\infty} \bar{\beta}_j \sim \mathcal{N}(d_j, \frac{d_j(1 - d_j)}{n}). \tag{20}$$

To distinguish and compare the reward probabilities for a_i and a_j, we state the null hypothesis (H_0) and the alternative hypothesis (H_1):

- H_0: $d_i = d_j$,
- H_1: $d_i \neq d_j$.

The z-value of the test statistics is computed by

$$z = \frac{\sqrt{n}|\bar{\beta}_i - \bar{\beta}_j|}{\sqrt{\bar{\beta}_i(1 - \bar{\beta}_i) + \bar{\beta}_j(1 - \bar{\beta}_j)}}. \tag{21}$$

For any given significance level α, the critical value $|z_{\alpha/2}|$ is obtained from the z-table [35], such

$$1-\alpha/2=\frac{1}{2\pi}\int_{-\infty}^{|z_{\alpha/2}|}e^{-\frac{z^2}{2}}dz. \tag{22}$$

At length, $|z|$ and $|z_{\alpha/2}|$ are compared. If $|z|>|z_{\alpha/2}|$, we reject the null hypothesis and claim that $d_i\neq d_j$ with significance level α . Statistically speaking, α is the probability that Z-test indicates $d_i\neq d_j$ when actually $d_i=d_j$.

However, Z-test is applicable only when the sample size, i.e.,n is large [35]. When n is small, another statistical test known as the Student's t-test is utilized [35]. The Student's t-test is more sophisticated than Z-test in selecting different critical values for different sample sizes. Since it has been shown that the Student's t-test and the Z-test are nearly indistinguishable when the sample size n is larger than 30 [35], the Student's t-test is only revoked when $n<30$ to reduce the complexity.

For the Student's t-test, the null and alternative hypotheses are still:

- H_0: $d_i=d_j$,
- H_1: $d_i\neq d_j$.

In the Student's t-test, the t-statistics is computed by

$$t=\frac{\sqrt{n}|\bar{\beta}_i-\bar{\beta}_j|}{\sqrt{S_i^2+S_j^2-2S_{ij}}}, \tag{23}$$

where

$$S_i^2=\frac{1}{n-1}\sum_{k=1}^{n}(\beta_i(k)-\bar{\beta}_i)^2, \tag{24}$$

$$S_j^2=\frac{1}{n-1}\sum_{k=1}^{n}(\beta_j(k)-\bar{\beta}_j)^2, \tag{25}$$

and

$$S_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(\beta_i(k)-\bar{\beta}_i)(\beta_j(k)-\bar{\beta}_j). \tag{26}$$

For any significance level α, define the degrees of freedom $\mathring{n}$ [35],

$$\mathring{n}=n-1. \tag{27}$$

Then, the critical value $|t_{\alpha/2}(\mathring{n})|$ is obtained from the t-table [35], such that

$$1-\alpha/2=\int_{-\infty}^{|t_{\alpha/2}(\mathring{n})|}\frac{\Gamma(\frac{\mathring{n}+1}{2})}{\Gamma(\frac{\mathring{n}}{2})}\frac{1}{\sqrt{\mathring{n}\pi}}\frac{1}{(1+\frac{t^2}{\mathring{n}})^{\frac{\mathring{n}+1}{2}}}dt. \tag{28}$$

Finally, $|t|$ and $|t_{\alpha/2}(\mathring{n})|$ are compared. If $|t|>|t_{\alpha/2}(\mathring{n})|$, the null hypothesis is rejected, and we can accept the alternative hypothesis with significance levelα.

Note that the Student's t-test is the alternative of the Z-test in small sample size. Equation (21) and (23) are equivalent when the sample size approaches to infinity [35].

B. STATISTICAL HYPOTHESIS TESTING APPROACH BASED FRAMEWORK

A statistical hypothesis testing approach based framework LA_{SHT} is designed on the basis of the discussion above, The framework consists of four parts: the initialization, the action selecting strategy, the action set updating strategy, and the convergence judgment. The notations that are used throughout the framework are summarized in Table 1.

TABLE 1. Notations used in the proposal of LA_{SHT}

Symbol	Explanation
α	significance level
$\beta_i(k)$	feedback for the k-th choosing of action a_i
t	number of iteration, $t=1,2,\cdots,\mathcal{T}$
l	number of round, $l=1,2,\cdots,L$
$A(l)$	action set at the l-th round
$F_i(l)$	feedback sequence for action a_i at the l-th round
$\hat{d}_i(l)$	mean of the feedback sequence $F_i(l)$
$a_{m(l)}$	estimated optimal action after the l-th round

1) Initialization

In the beginning, the feedback sequence for each action is empty. However, as clarified in [1], the convergence is boosted by introducing prior knowledge by selecting every action for at least a certain number of times. In fact, for the majorities of LA schemes, there is an initialization process that chooses each action for some times to ensure the feedback sequence of each action is longer than some threshold. In our scheme, a useful trick known as the technique of optimistic initial values is adopted [1]. That is, the initial feedback sequence $F_i(0)$ for each action a_i is indiscriminately set to $\{0,1\}$.

2) Action Selecting Strategy

The time complexity of selecting an action by the probability distribution of the action probability vector is $\mathcal{O}(r)$ at each iteration, which is time-consuming in massive-action environments. In the proposed framework, an economical uniform selection strategy is adopted. So the sampling issue is no longer resorted to. The learning process is organized into multiple rounds, with each round consists of a series of iterations. The index of each round is denoted by $l, l=1,2,\cdots,L$, where L is the number of the total rounds until the convergence. The action set at the l-th round is denoted by $A(l)$, and its size by $|A(l)|$. Naturally, we have

$$|A(1)|=r. \tag{29}$$

At the l-th round, each action in the current action set $A(l)$ is unbiasedly selected once to interact with the environment. **Note that an iteration is defined as one interaction between an action with the environment. And the l-th round consists of $|A(l)|$ interactions with the environment, i.e., costs $|A(l)|$ iterations.**

3) Action Set Updating Strategy

The action set is adjusted after each round adaptively by eliminating the suboptimal actions from the action set. When the l-th round finishes, for each action $a_i \in A(l)$, the feedback sequence becomes $F_i(l) = \{\beta_i(1), \beta_i(2), \cdots, \beta_i(l+2)\}$. Similarly to the estimator algorithms, we define the estimated reward probability $\hat{d}_i(l)$ as the mean of the feedback sequence,

$$\hat{d}_i(l) = \frac{\sum_{k=1}^{|F_i(l)|} \beta_i(k)}{|F_i(l)|} = \frac{\sum_{k=1}^{l+2} \beta_i(k)}{l+2}. \tag{30}$$

The estimated optimal action $a_{m(l)}$ after the l-th round is the one in $A(l)$ with the highest estimated reward probability $\hat{d}_m(l)$, where

$$a_{m(l)} = \arg \max_{a_i \in A(l)} \hat{d}_i(l). \tag{31}$$

The type of chosen statistical test depends on $|F_i(l)| = l+2$ (sample size). Follow the discussions in Section III-A, the Student's t-test is adopted when $|F_i(l)| \leq 30$, while the Z-test is adopted in further rounds. For each non-optimal action $a_i \in A(l), i \neq m(l)$, the statistical test is conducted given the feedback sequences of action a_i and $a_{m(l)}$ and the significance level α. If the null hypothesis H_0 is rejected, then it is implied that the reward probabilities of action a_i and $a_{m(l)}$ are different with significance level α, so the action a_i is eliminated from the current action set. After all non-optimal actions are compared with $a_{m(l)}$, the reward probabilities of all the remaining actions in the action set are indistinguishable at a certain significance level, therefore all actions need to be explored further.

Remark 1. *The trade-off between exploration and exploitation is an inevitable dilemma in RL [1]. To maximize the cumulative rewards, an agent must examine the actions that it has tried and found to be superior in producing rewards, i.e.,* exploit *what it has already learned. Meanwhile, the agent has to try actions that have been chosen for few times in order to obtain sufficient information for every action that is possibly optimal, i.e., to* explore *all possibilities. Thus, it is necessary to balance exploration and exploitation as each action must be tried for plentiful times in a stochastic environment. In view of this, the proposed action selecting strategy is effective with the combination of the well-designed action set updating strategy, which automatically reaches a balance between exploration and exploitation.*

4) Convergence Judgment

In the light of the action set updating strategy, the cardinality of the action set $A(l)$ monotonically decreases. The proposed LA gets converged whenever there is only one action left in the action set. That is,

$$|A(L)| = 1. \tag{32}$$

The total number of interactions with the environment, i.e., the convergence rate is

$$\mathcal{T} = \sum_{t=1}^{L} |A(L)|. \tag{33}$$

The overall process of the proposed LA_{SHT} is summarized in Algorithm 1.

Algorithm 1 The Statistical Hypothesis Testing Approach Based Framework

Require: α: the significance level.

Initialize Round: $l = 1$.
Initialize Action set: $A(l) = A(1) = A$.
Initialize Feedback sequence: $\forall a_i \in A, F_i(0) = \{0, 1\}$.
repeat
  Actions in $A(l)$ interact with the environment one by one, and update the feedback sequences of each action:
  $$F_i(l) = \{F_i(l-1), \beta_i(l)\}, a_i \in A(l)$$
  Compute the estimated reward probabilities of each action and find out the estimated optimal action $a_{m(l)}$ by (30) and (31), respectively.
  for $a_i \in A(l), i \neq m(l)$ **do**
    if $(l+2) \leq 30$ **then**
      Given the feedback sequences $F_{m(l)}(l)$, $F_i(l)$ and the significance level α, implement the t-test.
    else
      Given the feedback sequences $F_{m(l)}(l)$, $F_i(l)$ and the significance level α, implement the Z-test.
    end if
    if The null hypothesis $H_0 : d_i = d_m$ is rejected **then**
      Eliminate action a_i from the current action set:
      $$A(l+1) = A(l) \setminus \{a_i\}.$$
    end if
  end for
  $l = l + 1$.
until $|A(l)| = 1$
Output: the optimal action a_m which is the only action left in action set $A(l)$.

C. THEORETICAL ANALYSIS

Since the Student's t-test is the counterpart of the Z-test when the sample size is small, and separate critical values for each sample size in t-test make the analysis intractable, so the analysis in this section is based upon the Z-test. Note that when $\alpha \to 0$, the elimination of action is a rare event in finite rounds, so the analysis does not lose generality.

1) Analysis of the convergence rate $\mathcal{T}$

Consider two feedback sequences $F_i = \{\beta_i(1), \beta_i(2), \cdots\}$ and $F_j = \{\beta_j(1), \beta_j(2), \cdots\}$, where $\Pr\{\beta_i(k) = 1\} = d_i$,

$\Pr\{\beta_j(k)=1\}=d_j$ for $k=1,2,\cdots$ and $d_i>d_j$. Since $\beta_i(k)$ and $\beta_i(k)$ are i.i.d Bernoulli random variables, we have

$$\mathrm{E}[\beta_i(k)]=d_i,\quad \mathrm{E}[\beta_j(k)]=d_j. \tag{34}$$

By (21), if the null hypothesis H_0 is rejected with significance level α, the sample size n has to be no less than:

$$n\geq\left(\frac{z_{\alpha/2}\sqrt{d_i(1-d_i)+d_j(1-d_j)}}{d_i-d_j}\right)^2. \tag{35}$$

That is to say, for action a_i with reward probability d_i and action a_j with reward probability d_j, the prior estimation of the number of interactions with the environment to distinguish them with significance level α is

$$\hat{\mathcal{T}}_{ij}=2\cdot\left(\frac{z_{\alpha/2}\sqrt{d_i(1-d_i)+d_j(1-d_j)}}{d_i-d_j}\right)^2. \tag{36}$$

Suppose action a_m is the optimal action, the estimation of convergence rate $\hat{\mathcal{T}}$ within the proposed framework is

$$\hat{\mathcal{T}}=\sum_{i=1,i\neq m}^{r}\hat{\mathcal{T}}_{mi}+\max\{\hat{\mathcal{T}}_{mi}\}. \tag{37}$$

2) Proof of ϵ-Optimality

The ϵ-optimality is a crucial property of an LA, it indicates that, for any $\epsilon>0$, the LA converges to the optimal action in finite interactions with probability $1-\epsilon$. For the action probability vector based LA schemes, there is a conventional procedure to prove the ϵ-optimality [22]- [33]. For the proposed statistical hypothesis testing based framework, the ϵ-optimality has to be proved in a different manner.

Lemma 2. *For action set $A=\{a_1,a_2,\cdots,a_r\}$ and $D=\{d_1,d_2,\cdots,d_r\}$, given a significance level α, and an error ϵ_1, there exists a number of the round L_0, for all round $L>L_0$, LA$_{SHT}$ converges, i.e., $|A(L)|=1$ with probability $1-\epsilon_1$.*

Proof. For a given significance level α, $z_{\alpha/2}$ is obtained from (22). For action a_i and a_j, suppose $d_i\neq d_j$, according to (21), we have,

$$z=\sqrt{n}\cdot\Psi(\bar{\beta}_i,\bar{\beta}_j), \tag{38}$$

where $\bar{\beta}_i$ and $\bar{\beta}_j$ are the mean of feedback sequences $F_i(n)$ and $F_j(n)$, and

$$\Psi(\bar{\beta}_i,\bar{\beta}_j)=\frac{|\bar{\beta}_i-\bar{\beta}_j|}{\sqrt{\bar{\beta}_i(1-\bar{\beta}_i)+\bar{\beta}_j(1-\bar{\beta}_j)}}. \tag{39}$$

Since $\Psi(\bar{\beta}_i,\bar{\beta}_j)$ is continuous w.r.t both $\bar{\beta}_i$ and $\bar{\beta}_j$, by the definition of the convergence in probability [36] we can conclude that $\Psi(\bar{\beta}_i,\bar{\beta}_j)$ falls into the interval $(\Psi(d_i,d_j)-\delta,\Psi(d_i,d_j)+\delta)$ with probability $1-\epsilon_1$ when n is larger than $n(\epsilon_1,\delta)$. So when n is larger than $n(\epsilon_1,\delta)$ the z-value increases linearly $\sqrt{n}$ with probability $1-\epsilon_1$. Then, for any pair of actions a_i and a_j, let $\delta<|d_i-d_j|$, then for $n>n(\epsilon_1,\delta)$, $\Psi(\bar{\beta}_i,\bar{\beta}_j)\neq 0$ with probability $1-\epsilon_1$. And there always exists a n_0, $\forall n>\max\{n(\epsilon_1,\delta),n_0\}$ since z increases with n:

$$z=\sqrt{n}\cdot\Psi(\bar{\beta}_i,\bar{\beta}_j)>z_{\alpha/2}. \tag{40}$$

Without loss of generality, suppose a_i and a_j are the optimal action and suboptimal action respectively. Then for $L>\max\{n(\epsilon_1,\delta),n_0\}$, there is a distinction between them two and the suboptimal one is eliminated with significance level α with probability $1-\epsilon_1$. And all other inferior actions are naturally eliminated during the process. □

Lemma 3. *For each action in action set $A=\{a_1,a_2,\cdots,a_r\}$, when setting the significance level $\alpha\to 0$, the number of interactions with the environment of each action approaches to infinity.*

Proof. When setting $\alpha\to 0$, by (22), we have

$$\lim_{\alpha\to 0}\frac{1}{2\pi}\int_{-\infty}^{|z_{\alpha/2}|}e^{-\frac{t^2}{2}}dt\to 1. \tag{41}$$

This implies

$$\lim_{\alpha\to 0}|z_{\alpha/2}|\to\infty. \tag{42}$$

By (40), to reject the null hypothesis, the sample size should be infinity. That is, none action can be eliminated from the action set A before the total round approaches infinity when the significance level $\alpha\to 0$, which completes the proof. □

Theorem 2. *LA$_{SHT}$ is ϵ-optimal in every stationary random environment. That is, given any $\epsilon>0$, there exists a significance level $\alpha_0>0$, such that for all $\alpha<\alpha_0$:*

$$\Pr\{A(L)=\{a_m\}\}\geq 1-\epsilon, \tag{43}$$

where a_m is the optimal action that possesses the maximum reward probability.

Proof. To prove the ϵ-optimality, we first show that:

$$\lim_{\alpha\to 0}\Pr\{A(L)=\{a_m\}\}\to 1. \tag{44}$$

By setting $\epsilon_1\to 0$ in Lemma 2, the proposed LA with converge with probability one, we shall now show that LA converges to the optimal action with probability one, which can be proved by reductio ad absurdum.

Assume that LA converges to action $a_{m'}$ and $m'\neq m$. As action a_m is the optimal action, we have

$$d_{m'}<d_m. \tag{45}$$

Consider one of the convergence conditions, $\forall a_i\in A(L), i\neq m'$,

$$\hat{d}_{m'}>\hat{d}_i. \tag{46}$$

Lemma 3 implies that each action in A interacts with the environment infinity times when the significance level α approaches 0. According to Theorem 1, we have

$$\lim_{\alpha\to 0}\hat{d}_{m'}=d_m,\quad \lim_{\alpha\to 0}\hat{d}_m=d_m. \tag{47}$$

That is,

$$d_{m'}>d_m. \tag{48}$$

Obviously, (45) and (48) are contradictory. Thus, we reject the original assumption. Therefore it is implied that LA_{SHT} converges to the optimal action when the significance level α approaches 0.

To see how (44) implies the ϵ-optimality, we observe that (44) indicates that: given ϵ_2, $\forall \alpha < \epsilon$, $\exists \delta(\epsilon_2)$, such that $\Pr\{A(L) = \{a_m\}\} \geq 1 - \delta(\epsilon_2)$. To obtain ϵ-optimality, we only need to find a ϵ_2' such $\delta(\epsilon_2') < \epsilon$ for any possible ϵ.

When $\epsilon \geq 1$, this condition trivially holds. When $\epsilon \in (0, 1)$, it is sufficient to show that $\delta(\cdot)$ as the function of ϵ_2 does not have positive lower bound. This always holds for if $\delta(\cdot)$ is lower-bounded by $\delta' > 0$, then (44) would converge to $1 - \delta'$ rather than one. This completes the proof. □

IV. EXPERIMENTATIONS AND COMPARISONS

In this section, the proposed LA_{SHT} is evaluated and compared with other popular schemes in various stationary environments.

In the studies of LA, there are five benchmark environments, where the reward probabilities in each environment are:

- E_1: $D = \{0.65, 0.50, 0.45, 0.40, 0.35, 0.30, 0.25, 0.20, 0.15, 0.10\}$.
- E_2: $D = \{0.60, 0.50, 0.45, 0.40, 0.35, 0.30, 0.25, 0.20, 0.15, 0.10\}$.
- E_3: $D = \{0.55, 0.50, 0.45, 0.40, 0.35, 0.30, 0.25, 0.20, 0.15, 0.10\}$.
- E_4: $D = \{0.70, 0.50, 0.30, 0.20, 0.40, 0.50, 0.40, 0.30, 0.50, 0.20\}$.
- E_5: $D = \{0.10, 0.45, 0.84, 0.76, 0.20, 0.40, 0.60, 0.70, 0.50, 0.30\}$.

The complexity of an environment is reflected in two aspects: the number of actions r and the difference between the reward probabilities of the optimal action and suboptimal one denoted by δ. From this perspective, E_3 is the most complicated one among $E_1 - E_5$ with minimal $\delta = d_1 - d_2 = 0.05$. However, with a small r, these benchmarks are unrepresentative in most application scenarios. Therefore, we also consider a more complicated but practical massive-action environment, the monotonic environment [37]. The monotonic environment sets $d_1 = d_m$, $d_i = d_m(1 - \frac{i}{r})^\gamma$ for all $i = 2, 3, \cdots, r$. A series of monotonic environments were generated with different parameters as $E_6 - E_{10}$:

- $E_6 - E_{10}$: $d_m = 0.8, \gamma = 0.1, r = \{1000, 5000, 10000, 50000, 100000\}$.

The complexity of all environments tested is summarized in Table 2. Two sets of experiments are conducted for evaluation and comparison. Firstly, the influence of the significance level α on the performance of LA_{SHT} is presented. Secondly, the comparisons with schemes based on the action probability vector are given. All simulations are compiled by C++, using Intel(R) Core(R) CPU i7 @2.20GHz, 16.00GB RAM.

A. THE EFFECT OF THE SIGNIFICANCE LEVEL

As the only configurable parameter in LA_{SHT}, the significance level α is used to realize the trade-off between the accuracy and efficiency. Figure 2 presents the accuracy and the convergence rate of LA_{SHT} with different α in $E_1 - E_5$, performance on $E_6 - E_{10}$ are given in Section IV-B2. It can be concluded that a small value of α leads to high accuracy and low efficiency, while a large value of α leads to relatively low accuracy and high efficiency. The reason after this result is that when α is small, the LA_{SHT} is required to distinguish actions with higher confidence, which leads to higher accuracy. But to yield a distinction with higher confidence, a larger sampling size is necessary, so the efficiency declines. For α is large, the converse reasoning holds. Nevertheless, the performance of LA_{SHT} is steady when using the same significance level in different environments, which implies the parameter-free property of LA_{SHT}.

B. COMPARISONS WITH THE ACTION PROBABILITY VECTOR BASED SCHEMES

To show the superiority of the proposed framework, LA_{SHT} is compared with the most representative and the state-of-the-art schemes based on the action probability vector. The compared schemes are DP_{RI} [22] with the $\mathscr{T}_{\text{RI}}$ strategy, DGPA [26] with the $\mathscr{T}_{\text{GP}}$ strategy, and LELA [27] with the $\mathscr{T}_{\text{LE}}$ strategy.

1) Comparisons in Benchmark Environments

For the action probability vector based schemes, the parameters have to be tuned beforehand manually. Specifically, the resolution parameter N used to discretize the step-size when updating the action probability vector has to be determined for each environment. The tuning of N is a linear search problem, where millions of interactions with each environment are required. Table 3 shows the best parameters and the cost of parameter tuning for DP_{RI}, DGPA, and LELA in benchmark environments.

After tuning the parameters, the performance of LA_{SHT} is compared with well-tuned DP_{RI}, DGPA, and LELA. The simulations results are summarized in Table 4, from which we can construe the following conclusions:

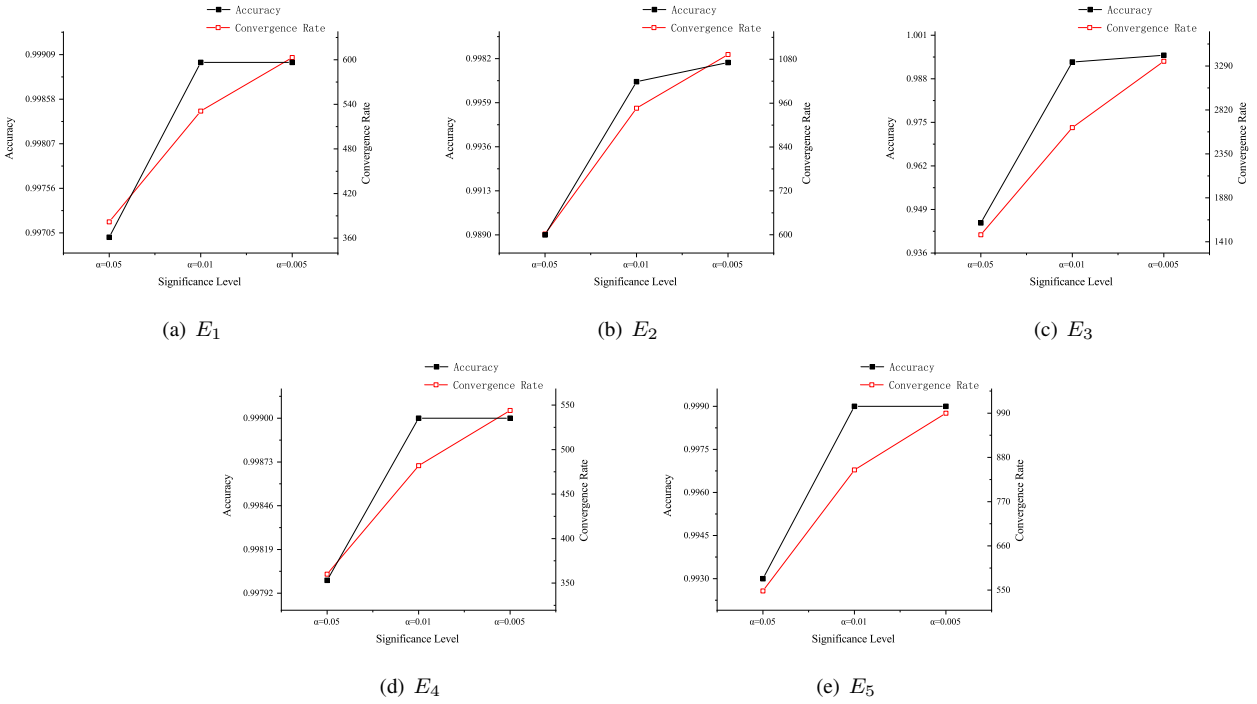
1) All accuracies are quite high, yielding the ϵ-optimality of LA. The difference between competitors and LA_{SHT} is reflected by the parameter setting. For DP_{RI}, DGPA, and LELA, the step-size for updating the action probability in each environment has been painstakingly tuned to achieve the fastest convergence rate with basic accuracy requirements. As for LA_{SHT}, an identical significance level can be used in any benchmarks, possessing natural adaptability in trade-offs between the accuracy and the convergence rate.
2) LA_{SHT} performs competitively in the convergence rate. Taking $\alpha = 0.01$ as a paradigm, compared with the action probability vector based schemes, LA_{SHT} converges with the fastest rate in every benchmark environment without losing accuracy. Meanwhile, define the improvements of the compared schemes relative to $\text{LA}_{\text{SHT}}(\alpha = 0.01)$ in benchmark environments by $\frac{\text{Con. Rate}_{\text{Compared Scheme}} - \text{Con. Rate}_{\text{LA}_{\text{SHT}}}}{\text{Con. Rate}_{\text{Compared Scheme}}}$. As shown in Figure 3, the improvements are significant.

2) Comparisons in Massive-Action Environments

After showing that LA_{SHT} performs competitively in benchmark environments, its performance in massive-action en-

TABLE 2. The number of actions r and the minimal difference δ of the reward probabilities between the optimal action and suboptimal actions in environments $E_1 - E_{10}$.

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
r	10	10	10	10	10	1000	5000	10000	50000	100000
δ	0.15	0.10	0.05	0.20	0.08	0.40	0.34	0.31	0.27	0.25



(a) E_1 (b) E_2 (c) E_3

(d) E_4 (e) E_5

FIGURE 2. Accuracy and convergence rate of LA_{SHT} with various significance levels α in benchmark environments (250,000 experiments were performed for each parameter setting in each environment).

TABLE 3. The best parameters and the tuning cost of three representative schemes in benchmark environments $E_1 - E_5$.

Env.	DP_{RI}		DGPA		LELA	
	Parameter	Cost	Parameter	Cost	Parameter	Cost
E_1	N=298	4.85E+09	N=33	4.36E+08	N=10	8.49E+07
E_2	N=653	2.45E+10	N=65	1.64E+09	N=21	2.88E+08
E_3	N=2356	3.40E+11	N=204	1.59E+10	N=59	3.30E+09
E_4	N=216	2.54E+09	N=28	3.17E+08	N=13	7.91E+07
E_5	N=881	3.12E+10	N=55	1.19E+09	N=28	3.70E+08

vironments was then investigated in $E_6 - E_{10}$. Note that due to the unbearable cost of parameter tuning in massive action environments, we set the resolution parameter $N = 1$ that yields the highest convergence rate for both DGPA and LELA. And DP_{RI} is neglected for its poor performance, i.e., the low accuracy when setting $N = 1$. Since the number of actions r is the dominating variable in massive-action environments $E_6 - E_{10}$ rather than δ, it would be considered as the independent variable. Figure 4 and Figure 5 illustrate the convergence rate and the convergence time of schemes versus the number of actions, respectively. As the curves of LA_{SHT} with different α are indistinguishable, Table 5 lists the detailed simulation results.

The superiority of LA_{SHT} is evident. On the one hand, the convergence rate of schemes based on the action probability vector increases rapidly with the increase in the number of actions, while that of LA_{SHT} grows much slower. In the most extreme case E_{10}, where $r = 100,000$, the required number of interactions with the environments of LA_{SHT} ($\alpha = 0.01$) is less than a quarter of DGPA and is less than one-eighth of LELA. On the other hand, a comparison between convergence time indicates that the action probability vector based schemes can be particularly inefficient in massive-action environments. As shown in Figure 5, LA_{SHT} converges in much less time than DGPA or LELA. For example, in E_{10} where $r = 100,000$, the convergence time of LA_{SHT}

TABLE 4. Comparisons of the accuracy and the convergence rate among various schemes in benchmark environments (250,000 experiments were performed for each scheme setting in each environment).

Action Probability Vector based Schemes						
	DP_{RI}		DGPA		LELA	
	Accuracy	Convergence Rate	Accuracy	Convergence Rate	Accuracy	Convergence Rate
E_1	0.994	1075	0.997	923	0.998	693
E_2	0.995	2576	0.996	1718	0.998	1362
E_3	0.992	9544	0.995	5098	0.993	3779
E_4	0.995	786	0.997	774	0.999	745
E_5	0.993	2335	0.996	1443	0.997	1188
Statistical Hypothesis Testing Approach Based Framework						
	$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.005$	
	Accuracy	Convergence Rate	Accuracy	Convergence Rate	Accuracy	Convergence Rate
E_1	0.997	382	0.999	531	0.999	603
E_2	0.989	602	0.997	946	0.998	1093
E_3	0.945	1486	0.993	2631	0.995	3340
E_4	0.998	360	0.999	482	0.999	544
E_5	0.993	548	0.999	849	0.999	990

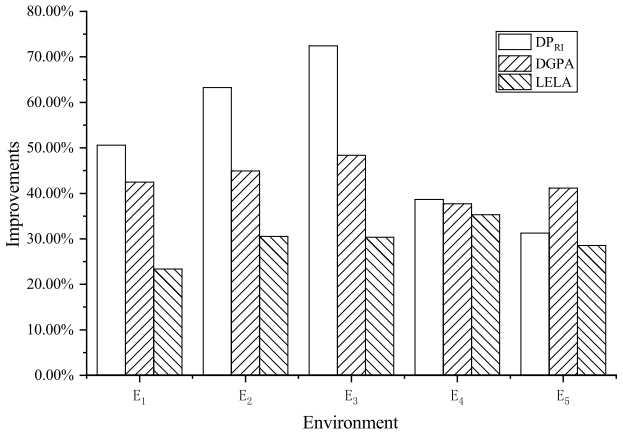


FIGURE 3. The improvements of the compared schemes relative to LA_{SHT} in benchmark environments.

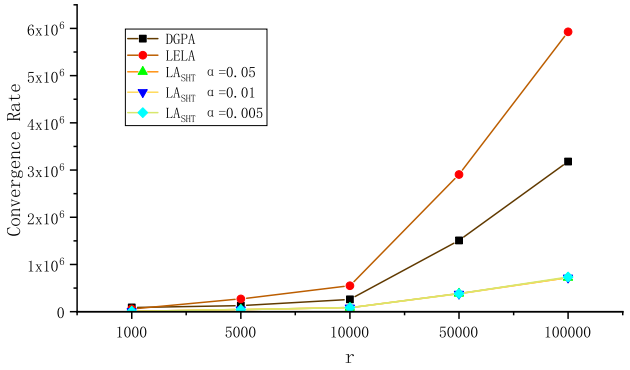


FIGURE 4. The convergence rate versus the number of actions r in massive-action environments $E_6 - E_{10}$ (250,000 experiments were performed for LA_{SHT} in each environment, 100 experiments were performed for both DGPA and LELA).

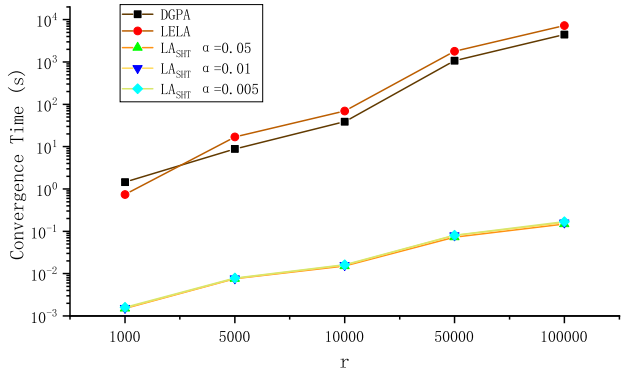


FIGURE 5. The convergence time versus the number of actions r in massive-action environments $E_6 - E_{10}$ (250,000 experiments were performed for LA_{SHT} in each environment, 100 experiments were performed for both DGPA and LELA).

($\alpha = 0.01$) is $0.147s$, while DGPA requires $4436.61s$ and LELA requires $7298.58s$, respectively. Thus it is obvious that significant improvements have been realized.

V. CONCLUSION

In this paper, we propose a novel LA framework based on the statistical hypothesis testing, which yields both the parameter-free property and the high efficiency. The proposed LA_{SHT} consists of four parts: the initialization, the action selecting strategy, the action set updating strategy, and the convergence judgment. The statistical hypothesis testing is utilized to eliminate inferior actions during the action set updating. And the balance between exploration and exploitation is attained through our realization. Theoretical analyses including the estimation of the convergence rate and the proof of the ϵ-optimality are also provided. Experimental results demonstrate that LA_{SHT} is parameter-free and ϵ-optimal. Comprehensive simulations in various environments verify

TABLE 5. Comparisons of the convergence rate and convergence time among LA_{SHT} with different significance levels in massive-action environments (250,000 experiments were performed for LA_{SHT} in each environment).

	$\alpha = 0.05$			$\alpha = 0.01$			$\alpha = 0.005$		
	Accuracy.	Con. Rate	Con. Time (s)	Accuracy	Con. Rate	Con. Time (s)	Accuracy	Con. Rate	Con. Time (s)
E_6	0.999	9403	0.00150	0.999	9560	0.00154	0.999	9603	0.00159
E_7	0.999	43286	0.00763	0.999	43362	0.00767	0.999	44207	0.00768
E_8	0.999	85505	0.01518	0.999	85602	0.15750	0.999	85622	0.01611
E_9	0.999	377899	0.07279	0.999	378534	0.08000	0.999	380720	0.08098
E_{10}	0.999	720787	0.14765	0.999	721773	0.16010	0.999	735587	0.16916

the superiority of LA_{SHT} to the action probability vector based schemes, especially in massive-action environments. Our further work includes extending our framework into non-stationary random environments and Q-, S-models.

REFERENCES

[1] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.

[2] Liu, Teng, et al. "Parallel reinforcement learning: a framework and case study." IEEE/CAA Journal of Automatica Sinica 5.4 (2018): 827-835.

[3] Xue, Lei, et al. "An adaptive strategy via reinforcement learning for the prisoner U+ 02BC s dilemma game." IEEE/CAA Journal of Automatica Sinica 5.1 (2018): 301-310.

[4] Cao, ZhengCai, et al. "Scheduling semiconductor testing facility by using cuckoo search algorithm with reinforcement learning and surrogate modeling." IEEE Transactions on Automation Science and Engineering 99 (2018): 1-13.

[5] Narendra, Kumpati S., and Mandayam AL Thathachar. Learning automata: an introduction. Courier Corporation, 2012.

[6] Tsetlin, Michael L. "On behaviour of finite automata in random medium." Avtom I Telemekhanika 22 (1961): 1345-1354.

[7] Guo, Haonan, et al. "Learning Automata Based Competition Scheme to Train Deep Neural Networks." IEEE Transactions on Emerging Topics in Computational Intelligence (2018).

[8] Khomami, Mohammad Mehdi Daliri, et al. "Minimum positive influence dominating set and its application in influence maximization: a learning automata approach." Applied Intelligence 48.3 (2018): 570-593.

[9] Di, Chong, et al. "Learning Automata based Access Class Barring Scheme for Massive Random Access in Machine-to-Machine Communications." IEEE Internet of Things Journal (2018).

[10] Cuevas, Erik, et al. "Fast algorithm for multiple-circle detection on images using learning automata." IET Image Processing 6.8 (2012): 1124-1135.

[11] Thathachar, M. A. L., and P. S. Sastry. "Learning Automata for Pattern Classification." Networks of Learning Automata. Springer, Boston, MA, (2004): 139-176.

[12] Zeng, Xianyi, and Zeyi Liu. "A learning automata based algorithm for optimization of continuous complex functions." Information Sciences 174.3-4 (2005): 165-175.

[13] Thapa, Rajan, et al. "A learning automaton-based scheme for scheduling domestic shiftable loads in smart grids." IEEE Access 6 (2018): 5348-5361.

[14] Zhu, Junpeng, et al. "Learning automata-based methodology for optimal allocation of renewable distributed generation considering network reconfiguration." IEEE Access 5 (2017): 14275-14288.

[15] Ai, Zhengyang, et al. "A Smart Collaborative Charging Algorithm for Mobile Power Distribution in 5G Networks." IEEE Access (2018).

[16] Yazidi, Anis, and Hugo L. Hammer. "Solving stochastic nonlinear resource allocation problems using continuous learning automata." Applied Intelligence (2018): 1-20.

[17] Varshavskii, V. I., and I. P. Vorontsova. "On the behavior of stochastic automata with a variable structure." Avtomatika i Telemekhanika 24.3 (1963): 353-360.

[18] Santharam, G., P. S. Sastry, and M. A. L. Thathachar. "Continuous action set learning automata for stochastic optimization." Journal of the Franklin Institute 331.5 (1994): 607-628.

[19] Guo, Ying, Shenghong Li, and Bo Fan. "Adaptive continuous action-set learning automata scheme." Electronics Letters 54.4 (2018): 242-244.

[20] Najim, Kaddour, and Alexander S. Poznyak. Learning automata: theory and applications. Elsevier, 2014.

[21] Rezvanian A, Saghiri AM, Vahidipour SM, Esnaashari M, Meybodi MR. Recent advances in learning automata. Springer, 2018.

[22] Oommen, B. John, and Joseph K. LanctÃt't. "Discretized pursuit learning automata." IEEE Transactions on systems, man, and cybernetics 20.4 (1990): 931-938.

[23] Papadimitriou, Georgios I., Maria Sklira, and Andreas S. Pomportsis. "A new class of/spl epsi/-optimal learning automata." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 34.1 (2004): 246-254.

[24] Zhang, Xuan, Ole-Christoffer Granmo, and B. John Oommen. "On incorporating the paradigms of discretization and Bayesian estimation to create a new family of pursuit learning automata." Applied intelligence 39.4 (2013): 782-792.

[25] Zhang, Junqi, et al. "Incorporation of optimal computing budget allocation for ordinal optimization into learning automata." IEEE Transactions on Automation Science and Engineering 13.2 (2016): 1008-1017.

[26] Agache, Mariana, and B. John Oommen. "Generalized pursuit learning schemes: New families of continuous and discretized learning automata." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 32.6 (2002): 738-749.

[27] Zhang, Junqi, Cheng Wang, and MengChu Zhou. "Last-position elimination-based learning automata." IEEE Trans. Cybernetics 44.12 (2014): 2484-2492.

[28] Ge, Hao. "A Parameter-Free Learning Automaton Scheme." arXiv preprint arXiv:1711.10111 (2017).

[29] Ge, Hao, et al. "A parameter-free gradient Bayesian two-action learning automaton scheme." Proceedings of the 2015 International Conference on Communications, Signal Processing, and Systems. Springer, Berlin, Heidelberg, (2016): 963-970.

[30] Guo, Ying, Hao Ge, and Shenghong Li. "A loss function based parameter-less learning automaton scheme." Neurocomputing 260 (2017): 331-340.

[31] Fishman, George. Monte Carlo: concepts, algorithms, and applications. Springer Science & Business Media, 2013.

[32] Zhang, JunQi, Cheng Wang, and MengChu Zhou. "Fast and epsilon-optimal discretized pursuit learning automata." IEEE transactions on cybernetics 45.10 (2015): 2089-2099.

[33] Thathachar, M. A. L. "Discretized reward-inaction learning automata." J. Cybernetics and Information Science 2 (1979): 24-29.

[34] Barany, Imre, and Van Vu. "Central limit theorems for Gaussian polytopes." The Annals of Probability 35.4 (2007): 1593-1621.

[35] Casella, George, and Roger L. Berger. Statistical inference. Vol. 2. Pacific Grove, CA: Duxbury, 2002.

[36] Cover, Thomas M., and Joy A. Thomas. Elements of information theory. John Wiley & Sons, 2012.

[37] Jamieson, Kevin, and Robert Nowak. "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting." Information Sciences and Systems (CISS), 2014 48th Annual Conference on. IEEE, (2014): 1-6.