The Sampling-and-Learning Framework: A Statistical View of Evolutionary Algorithms

Yang Yu
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
Email: yuy@lamda.nju.edu.cn

Hong Qian
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
Email: qianh@lamda.nju.edu.cn

Abstract—Evolutionary algorithms (EAs), a large class of general purpose optimization algorithms inspired from the natural phenomena, are widely used in various industrial optimizations and often show excellent performance. This paper presents an attempt towards revealing their general power from a statistical view of EAs. By summarizing a large range of EAs into the sampling-and-learning framework, we show that the framework directly admits a general analysis on the probable-absolute-approximate (PAA) query complexity. We particularly focus on the framework with the learning subroutine being restricted as a binary classification, which results in the sampling-and-classification (SAC) algorithms. With the help of the learning theory, we obtain a general upper bound on the PAA query complexity of SAC algorithms. We further compare SAC algorithms with the uniform search in different situations. Under the error-target independence condition, we show that SAC algorithms can achieve polynomial speedup to the uniform search, but not super-polynomial speedup. Under the one-side-error condition, we show that super-polynomial speedup can be achieved. This work only touches the surface of the framework. Its power under other conditions is still open.

I. INTRODUCTION

In many practical optimization problems, the objective functions are hidden or too complicated to be analyzed. Under this kind of circumstances, direct optimization algorithms are appealing, which follows the trial-and-error style with some heuristics. Evolutionary algorithms (EAs) [3] are a large family of such algorithms. The family includes genetic algorithms [17], evolutionary programming [26], evolutionary strategies [5], and also covers other nature-inspired heuristics including particle swarm optimization [25], ant colony optimization [11], estimation of distribution algorithms [29], etc.

Theoretical studies of EAs have been developed rapidly in the recent decades, particularly noticeable of the blooming of running time analysis [32], [2], [20]. With the development of several analysis techniques (e.g. [18], [44], [9], [37]), EAs have been theoretically investigated on problems from simple synthetic ones (e.g. [13]) to combinatorial problems (e.g. [35]) as well as NP-hard problems (e.g. [43]). During these analyses, effects of EAs components have been disclosed [42], including the crossover operators (e.g. [22], [31], [8], [33]), the population size (e.g. [21], [30], [40], [6]), etc. Measures of the performance also have developed to cover the approximation complexity (e.g. [19], [16], [43], [28]), the fixed-parameter complexity (e.g. [27], [33]), the complexity under fixed-budget computation [23], etc. While most of these analyses studied instances of EAs on problem cases, general performance analysis may even be more desired, as the application of EAs is nearly unlimited. The famous No-Free-Lunch Theorem [41] used a quite general framework of EAs and gave a general conclusion that any two EAs are with the same performance (at least on discrete domains) given no prior knowledge of the problem distribution, of which the general running time is exponential [44]. When the complexity of a problem class is bounded, a general convergence lower bound can be derived for a class of EAs [15]. For more general EAs, the Black-Box model can derive the best possible performance [12], [1], [50], [10]. We have learned that a general performance analysis relies on a general framework of EAs.

It has been noticed that various implementations of EAs share a common structure that consists of a cycle of sampling and model building [47]. In this work, we propose to study the sampling-and-learning (SAL) framework. EAs commonly employ some heuristic to reproduce solutions, which is captured by the sampling step of SAL; and they also distinguish the quality of the reproduced solutions to guide the next sampling (e.g., genetic algorithms remove a portion of the worst solutions), which is captured by the learning step of SAL. The SAL framework can simulate a wide range of EAs as well as other heuristic search methods, by specifying the sampling and the learning strategies.

We evaluate this framework by the probable-absolute-approximate (PAA) query complexity. PAA complexity counts the number of fitness evaluations before reaching to an approximate solution with a probability, which is close to the intuitive evaluation of EAs in practice. We show that the SAL framework immediately admits a general PAA upper bound. For a specific version of SAL that uses classification algorithms, named the SAC algorithms, we obtain a tighter upper bound. For a class of EAs [15]. For more general EAs, the Black-Box model can derive the best possible performance [12], [1], [50], [10]. We have learned that a general performance analysis relies on a general framework of EAs.

The rest of this paper is organized as follows: Section II introduces the SAL framework. In Section III, we compare the SAC algorithms, a specific version of the SAL framework, with the uniform search. Finally, Section IV concludes the paper.
II. The Sampling-and-Learning Framework

In this paper, we consider general minimization problems \( f \). We always denote \( X \) as the whole solution space which an algorithm will search among. In the analysis of this paper, we consider \( X \subseteq \mathbb{R}^n \) is a compact set (in the Euclidean space, the compact set is equivalent to the bounded and closed set) and \( f : X \rightarrow \mathbb{R} \) is a continuous function. Thus there must exist at least one solution \( x^* \in X \) such that \( f(x^*) = \min_{x \in X} f(x) \). We use \( D \) to denote sub-regions of \( X \) and define \( |D| = \int_D 1 dx \). For the sake of the intuitive evaluation of EAs in practice, we assume without loss of generality that \( |X| = 1 \) since \( X \) is a bounded and closed set. Denote \( D_\alpha = \{x \in X | f(x) \leq \alpha \} \) for any scaler \( \alpha, U_X \) as the uniform distribution over \( X \), \( T \) and \( D \) as the probability distributions. Besides, by \( \text{poly}(\cdot) \), we mean the set of all polynomials with the related variables, and by \( \text{superpoly}(\cdot) \), we mean the set of all functions that grow faster than any function in \( \text{poly}(\cdot) \) with the related variables.

**Definition 1** (Minimization Problem) A minimization problem consists of a continuous solution space \( X \) and a continuous function \( f : X \rightarrow \mathbb{R} \), where \( X \subseteq \mathbb{R}^n \) and \( X \) is a compact set. The goal is to find a solution \( x^* \in X \) such that \( f(x^*) \leq f(x) \) for all \( x \in X \).

Since \( X \) is a compact set and \( f \) is a continuous function, there must exist one solution \( x^* \in X \) such that \( f(x^*) = \max_{x \in X} f(x) \). Namely, \( f \) is bounded in \([f(x^*), f(x)]\). Therefore, in the rest of the paper, we assume without loss of generality that the value of \( f \) is bounded in \([0, 1]\), i.e., \( \forall x \in X : f(x) \in [0, 1] \). Given an arbitrary function \( g \) with bounded value range over the input domain, the bound can be implemented by a simple normalization \( f(x) = \frac{g(x) - g(x^*)}{\max_x g(x) - \min_x g(x^*)} \). Thus we assume in the rest of this paper that every minimization problem has its minimum value 0.

In real-world applications, we expect EAs to achieve good enough solutions with a not quite small probability, which corresponds to approximation (e.g. \([43]\)) and probabilistic performance (e.g. \([43]\)). Combining the two, we study the probable-absolute-approximate (PAA) query complexity, which is the number of fitness evaluations that an algorithm takes before reaching an approximate quality, as defined in Definition 2. The PAA query complexity closely reflects our intuitive evaluation of EAs in practice.

**Definition 2** (Probable-Absolute-Approximate Query Complexity) Given a minimization problem \( f \), an algorithm \( A \), and any \( 0 < \delta < 1 \) as well as any approximation level \( \alpha^* > 0 \), then the probable-absolute-approximate (PAA) query complexity is the number of calls to \( f(\cdot) \) such that, with probability at least \( 1 - \delta \), \( A \) finds a solution \( x \) with \( f(x) \leq \alpha^* \).

A. The General Framework

Most EAs share a common trial-and-error structure with several important properties:

a) directly access the solution space, generate solutions, and evaluate the solutions;

b) the generation of new solutions depends only on a short history of past solutions;

c) both “global” and “local” heuristic operators are employed to generate new solutions.

We present a sampling-and-learning (SAL) framework in Algorithm 1 to capture these properties. The SAL framework starts from a random sampling in Step 1 like all EAs. Steps 2 and 13 record the best-so-far solutions throughout the search. SAL follows a cycle of learning and sampling stages. In Step 7, it learns a hypothesis \( h_t \) (i.e., a mapping from \( X \) to \( \mathbb{R} \)) via the learning algorithm \( \mathcal{L} \). Note that the learning algorithm allows to take the current data set \( T_t \), the last data set \( T_{t-1} \), and the last hypothesis \( h_{t-1} \) into account. Different EAs may make different use of them. Step 8 initializes the sample set for the next iteration. The sample set can be initialized as an empty set, or to preserve some good solutions from the previous iteration. In Steps 9 to 12, it samples from the distribution transformed from the hypothesis as well as from the whole solution space balanced by a probability. The distribution \( T_{h_t} \) implies the potential good regions learned by \( h_t \).

It should be noted that the SAL framework is not a concrete optimization algorithm but an abstract summary of a range of EAs, nor does the learning stage of the framework imply an accurate learning. We explain in the following how we could mimic several different EAs by the SAL framework. It is noticeable that the explanation is not a rigorous proof, but an intuitive illustration that the SAL framework can correspond to various implementations.

The genetic algorithms (GAs) \([17]\) deal with discrete solution spaces consisting of solutions represented as a vector of vocabulary. The element-wise mutation operator changes every element of a solution to a randomly selected word from the vocabulary with a probability. Converting this operation probability to the probability of generating a certain solution, let \( P_m(x^i|\cdot) \) be the probability of generating the solution.

**Algorithm 1** The sampling-and-learning (SAL) framework

**Input:**
- \( \alpha^* > 0 \): Approximation level
- \( T \in \mathbb{N}^+ \): Number of iterations
- \( m_0, \ldots, m_T \in \mathbb{N}^+ \): Number of samples
- \( \lambda \in [0, 1] \): Balancing parameters
- \( \mathcal{L} \): Learning algorithm
- \( T \): Distribution transformation of hypothesis

**Procedure:**
1. Collect \( S_0 = \{x_1, \ldots, x_{m_0}\} \) by i.i.d. sampling from the uniform distribution over \( X \)
2. \( \bar{x} = \arg\min_{x \in S_0} f(x) \).
3. Initialize the hypothesis \( h_0 \)
4. \( T_0 = \emptyset \)
5. for \( t = 1 \) to \( T \) do
6. Construct \( T_t = \{(x_1, y_1), \ldots, (x_{m_{t-1}}, y_{m_{t-1}})\} \), where \( x_i \in S_{t-1} \) and \( y_i = f(x_i) \)
7. \( h_t = \mathcal{L}(T_t, T_{t-1}, h_{t-1}, t) \), the learning step
8. Initialize \( S_t \) from \( T_t \)
9. for \( i = 1 \) to \( m_t \) do
10. Sample \( x_i \) from \( T_{h_t} \) with probability \( \lambda \)
11. \( S_t = S_t \cup \{x_i\} \)
12. end for
13. \( \bar{x} = \arg\min_{x \in S_t \cup \{\bar{x}\}} f(x) \)
14. end for
15. return \( \bar{x} \)
The SAL framework directly admits a general upper bound of the PAA query complexity, as stated in Theorem 1.

**Theorem 1** For any minimization problem $f$ and any approximation level $\alpha^* > 0$, with probability at least $1 - \delta$, a SAL algorithm will output a solution $x$ with $f(x) \leq \alpha^*$ using $m_2$ number of queried samples bounded from above by

$$O(m_0 + \max_{m \geq 1} \frac{1}{(1-\lambda)Pr_u + \lambda Pr_h} \ln \frac{1}{\delta})$$

where $Pr_u = \frac{R}{D_{\alpha^*} U(x) dx}$ is the success probability of uniform sampling,

$$\overline{Pr}_h = \frac{P}{m_1} \frac{T}{m_1} m_1 T$$

is the average success probability of sampling from the learnt hypothesis, $mPr_{hi}$ is the required sample size realizing $Pr_{hi}$, and $D_{\alpha^*} = \{x \in X | f(x) \leq \alpha^*\}$.

**Proof.** $m_0$ is the initial sample size. In every iteration, we need $mPr_{hi}$ samples to realize the probability $Pr_{hi}$, (generally the higher the probability the larger the sample size, but it depends on the concrete implement of the algorithm), thus $t = 1^mPr_{hi}$ number of samples is naturally required. We prove the rest of the bound.

Let’s consider the probability that after $T$ iterations, the SAL algorithm outputs a bad solution $x$ such that $f(x) > \alpha^*$. Since the $x$ is the best solution among all sampled examples, the probability is the intersection of events that every step of the sampling does not generate such a good solution.

1. For the sampling from uniform distribution over the whole solution space $X$, the probability of failure is $1 - Pr_u$.
2. For the sampling from the learnt hypothesis $h_1$ according to the distribution $Pr_{hi}$, the probability of failure is denoted as $1 - Pr_{hi}$.

Since every sampling is independent, we can expand the probability of overall failures, i.e., for any solution $x$ belongs to the all sampled examples,

$$Pr(f(x) > \alpha^*) = (1 - Pr_{hi})^m_0.$$

**B. The Sampling-and-Classification Algorithms**

To further unfold the unknown term $Pr_{hi}$ in Theorem 1, we focus on a simplified version of the SAL framework that employs a classification algorithm in the learning stage. We call this type of algorithms as the sampling-and-classification (SAC) algorithms. In the learning stage of a SAC algorithm, as described in Algorithm 2, the learning algorithm first uses a threshold to transform the data set into a binary labeled data set, and then invokes the classification algorithm to learn from the binary data set. sign$[\cdot]$ is defined as sign$[v] = +1$ if $v \geq 0$ and $-1$ if $v < 0$. Note that SAC algorithms use the current
Algorithm 2 Learning sub-procedure for the sampling-and-classification (SAC) algorithms

Input: $T, T', h', t$; The input variables
$\alpha_1 > \ldots > \alpha_t$: Preset threshold parameters
$C$: Classification algorithm

Procedure:
1: Construct $B = \{(x_1, z_1), \ldots, (x_T, z_T)\}$ from that, for all $i$ and all $(x_i, y_i) \in T$, $z_i = \text{sign}[\alpha_i - y_i]$
2: $h = C(B)$
3: return $h$

data set $T$ in the learning algorithm, but not the last data set $T'$ and the last hypothesis $h'$. Putting Algorithm 2 into the framework of Algorithm 1, we always set $S_t = \emptyset$ for SAC, and $T_h$ will be some distribution over the positive area of $h$.

By these specifications, we can have a general PAA performance for SAC algorithms. According to Theorem 1, we need to estimate a lower bound of $P_{\text{Pr}_h}$, i.e., how likely the distribution $T_h$ will lead to a good solution. Recall $D_\alpha = \{x \in X | f(x) \leq \alpha\}$ for any scaler $0 < \alpha < 1$. Denote $D_h = \{x \in X | h(x) = +1\}$ for any hypothesis $h$, $U_{D_h}$ as the uniform distribution over $D_h$, and $D_{KL}$ as the Kullback-Leibler (KL) divergence. KL-Divergence measures how different one distribution departs from another one. For probability distributions $P$ and $Q$ of two continuous random variables, $D_{KL}(P||Q) = \int_{-\infty}^{\infty} \ln \left( \frac{p(x)}{q(x)} \right) p(x)dx$, where $p(x)$ and $q(x)$ are the probability densities of $P$ and $Q$. Let $\Delta$ denote the symmetric difference operator of two sets. We have a lower bound of the success probability as in Lemma 1.

Lemma 1 For any minimization problem $f$, any approximation level $\alpha^* > 0$, any hypothesis $h$, the probability that a solution sampled from an arbitrary distribution $T_h$ defined on $D_h$ will lead to a solution in $D_{\alpha^*}$ is lower bounded as

$$
\text{Pr}_h \geq \frac{|D_{\alpha^*} \cap D_h|}{|D_h|} - \frac{1}{2} \text{KL}(\text{T}_h||U_{D_h})
$$

Proof. Let $I[\cdot]$ denote the indicator function, namely, $I[\text{true}] = 1$ and $I[\text{false}] = 0$. The proof starts from the definition of the probability,

$$
\text{Pr}_h = \int_{D_h} T_h(x) \cdot I[x \in D_{\alpha^*}] dx
$$

$$
= \int_{D_h} \left( T_h(x) - U_{D_h}(x) + U_{D_h}(x) \right) \cdot I[x \in D_{\alpha^*}] dx
$$

$$
= \frac{|D_{\alpha^*} \cap D_h|}{|D_h|} + \int_{D_h} \left( T_h(x) - U_{D_h}(x) \right) \cdot I[x \in D_{\alpha^*}] dx
$$

$$
\geq \frac{|D_{\alpha^*} \cap D_h|}{|D_h|} + \sup_{x' \in D_h} \int_{D_h} T_h(x') - U_{D_h}(x') \cdot I[x \in D_{\alpha^*}] dx
$$

$$
\geq \frac{|D_{\alpha^*} \cap D_h|}{|D_h|} - \frac{1}{2} \text{KL}(\text{T}_h||U_{D_h}) \cdot I[x \in D_{\alpha^*}] dx
$$

$$
= \frac{|D_{\alpha^*} \cap D_h|}{|D_h|} - \frac{1}{2} \text{KL}(\text{T}_h||U_{D_h})
$$

where the last inequality is by Pinsker’s inequality.

We cannot pre-determine $D_h$, but we know that $h$ is derived by a binary classification algorithm from a data set which is labeled according to the threshold parameter $\alpha$. For the binary classification, we know that the generalization error, which is the expected misclassification rate, can be bounded above by the training error, which is the misclassification rate in the seen examples, as well as the generalization gap involving the complexity of the hypothesis space [24], as in Lemma 2. The VC($|H|$) is the VC-dimension measuring the complexity of $H$.

Lemma 2 (24) Let $H = \{h : X \to \{-1,+1\}\}$ be the hypothesis space containing a family of binary classification functions and VC($H$) = $d$, if there exist $m$ samples i.i.d. from $X$ according to some fixed unknown distribution $D$, then, $\forall h \in H$ and $\forall 0 < \eta < 1$, the following upper bound holds true with probability at least $1 - \eta$:

$$
\epsilon_D \leq \epsilon_D + 8m^{-1} \log(2emd^{-1}) + \log(4\eta^{-1})
$$

where $\epsilon_D$ is the expected error rate of $h$ over $D$ and $\epsilon_D$ is the error rate in the sampled examples from $D$, and when $\epsilon_D = 0$,

$$
\epsilon_D \leq 2m^{-1} \log(2emd^{-1}) + \log(2\eta^{-1})
$$

Again by Pinsker’s inequality, we know that the error $\epsilon_D$ under the distribution $D$ can be converted to the error $\epsilon_U$ under the uniform distribution, as

$$
\epsilon_U \leq \frac{\epsilon_D m^{\eta}}{1 - \frac{1}{2} \text{KL}(D||U)}
$$

where we only take the event that the generalization inequality holds with probability $1 - \eta$ into account. For simplicity, we denote the right-hand part as $\psi^{m,\eta}_{\epsilon_D,D,D_{KL}(D||U)}$ which decreases with $m$ and $\eta$, and increases with $\epsilon_D$, $d$, and $D_{KL}(D||U)$.

We can use this inequality to eliminate the $D_h$ in Lemma 1. In every iteration of SAC algorithms, there are $m_t$ samples collected, which make the error of $h_t$ bounded.

Theorem 2 For any minimization problem $f$, any constant $0 < \eta < 1$, any approximation level $\alpha^* > 0$, the average success probability of sampling from the learnt hypothesis of any SAC algorithm is lower bounded as

$$
\text{Pr}_h \geq \frac{1 - \eta}{t=1} \sum_{i=1}^{m_t} \left[ D_{\alpha^*} \cdot \frac{1 - 2 \psi^{m_t,\eta}_{\epsilon_D,d,D_{KL}(D||U)}}{D_{\alpha^*} + \psi^{m_t,\eta}_{\epsilon_D,d,D_{KL}(D||U)}} \right] - |D_{\alpha^*}| \cdot \frac{1}{2} \text{KL}(\text{T}_h||U_{D_h})
$$

where $D_t = \lambda T_{h_t} + (1 - \lambda)U_X$ is the sampling distribution at iteration $t$, $\epsilon_D$ is the training error rate of $h_t$, $d$ is the VC-dimension of the learning algorithm.

Proof. By set operators,

$$
|D_{\alpha^*} \cap D_{h_t}| = |D_{\alpha^*} \cup D_{h_t} - D_{\alpha^*} \Delta D_{h_t}|
$$

$$
\geq |D_{\alpha^*} \cup D_{h_t} - D_{\alpha^*} \Delta D_{h_t} - D_{\alpha^*} \Delta D_{h_t}|
$$

$$
= |D_{\alpha^*} \cup D_{h_t} - D_{\alpha^*} \Delta D_{h_t} - \epsilon_U \cdot t
$$
= |D_{\alpha^*} \cup D_{h_t}| + |D_{\alpha^*}| - |D_{\alpha^*}| - \epsilon_{\Delta h_t, t},
where $\Delta$ is the symmetric difference operator of two sets and $\epsilon_{\Delta h_t, t}$ is the expected error rate of $h_t$ under $\mathcal{U}_X$. The first inequality is by the triangle inequality, and the last equation is by that $D_{\alpha^*}$ is contained in $D_{\alpha^*}$.

Since $|D_{h_t}| - |D_{\alpha^*}| \leq |D_{h_t} \Delta D_{\alpha^*}| = \epsilon_{\Delta h_t, t}$, we can bound $|D_{h_t}|$ as $|D_{\alpha^*}| + \epsilon_{\Delta h_t} \leq |D_{h_t}| \leq |D_{\alpha^*}| - \epsilon_{\Delta h_t, t}$.

Now, we can apply Lemma 1 and the success probability of sampling from $D_{h_t}$ is lower bounded as

$$
\Pr_{h_t} \geq \frac{|D_{\alpha^*} \cap D_{h_t}| - |D_{\alpha^*} \cap D_{h_t}|}{|D_{h_t}|} \cdot \frac{1}{2 \KL(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{h_t}})} \geq \frac{|D_{\alpha^*} - \epsilon_{\Delta h_t, t}|}{|D_{\alpha^*} + \epsilon_{\Delta h_t, t}} - \frac{|D_{\alpha^*}| - \epsilon_{\Delta h_t, t}|}{|D_{\alpha^*}| - \epsilon_{\Delta h_t, t}} \geq \frac{|D_{\alpha^*}| - \epsilon_{\Delta h_t, t}|}{|D_{\alpha^*}| - \epsilon_{\Delta h_t, t}} \cdot \frac{1}{2 \KL(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{h_t}})}.
$$

Substituting this lower bound and the probability $1 - \eta$ of the generalization bound into $\Pr_{h_t}$ obtains the theorem. ■

Combining Theorem 1 and Theorem 2 results an upper bound on the sampling complexity of SAC algorithms. Although the expression looks sophisticated, it can still reveal relative variables that generally effect the complexity. One could design various distributions for $\mathcal{T}_{h_t}$ to sample potential solutions, however, without any a priori knowledge, the uniform sampling is the best in terms of the worst case performance. Meanwhile, without any a priori knowledge, a small training error at each stage from a learning algorithm with a small VC-dimension can also improve the performance.

III. SAC ALGORITHMS V.S. UNIFORM SEARCH

When EAs are applied, we usually expect that they can achieve a better performance than some baselines. The uniform search can serve as a baseline, which searches the solution space always by sampling solutions uniformly at random. In other words, the uniform search is the SAL algorithm with $\lambda = 0$. In this section, we study the performance of SAC algorithms relative to the uniform search.

SAC algorithms will degenerate to uniform search if $\lambda = 0$. Thus, it is easy to know that the PAA query complexity of uniform search is

$$\Theta \left( \frac{1}{\Pr_{h_t}} \cdot \ln \frac{1}{\delta} \right).$$

Contrasting this with Theorem 1, we can find that how much a SAC algorithm improves from the uniform search depends on the average success probability $\Pr_{h_t}$ that relies on the learnt hypothesis. A SAC algorithm is not always better than the uniform search. Without any restriction, $\Pr_{h_t}$ can be zero and thus the SAC algorithm is worse. We are then interested in investigating the conditions under which SAC algorithms can accelerate from the uniform search.

A. A Polynomial Acceleration Condition

**Condition 1** (Error-Target Independence) In SAC algorithms, for any $t$ and any approximation level $\alpha^* > 0$, when sampling a solution $x$ from $\mathcal{U}_X$, the event $x \in D_{h_t} \Delta D_{\alpha^*}$ and the event $x \in D_{\alpha^*}$ are independent.

We call SAC algorithms that are under the error-target independence condition as SAC algorithms. The condition is defined using the independence of random variables. From the set perspective, it is equivalent with

$$|D_{\alpha^*} \cap (D_{\alpha^*} \Delta D_{h_t})| = |D_{\alpha^*}| \cdot |(D_{\alpha^*} \Delta D_{h_t})|.$$ 

Under the condition, we can bound from below the probability of sampling a good solution, as stated in Lemma 3.

**Lemma 3** For SAC algorithms, it holds for all $t$ that

$$|D_{\alpha^*} \cap (D_{h_t})| \geq \frac{|D_{\alpha^*}|(1 - \epsilon_{\Delta h_t, t})}{|D_{\alpha^*}| + \epsilon_{\Delta h_t, t}},$$

where $\epsilon_{\Delta h_t, t}$ is the expected error rate of $h_t$ under $\mathcal{U}_X$.

**Proof.** For the numerator,

$$|D_{\alpha^*} \cap (D_{h_t})| = |D_{\alpha^*}| - |D_{\alpha^*} \cap (D_{\alpha^*} \Delta D_{h_t})| = |D_{\alpha^*}| \cdot |(D_{\alpha^*} \Delta D_{h_t})| \geq |D_{\alpha^*}| \cdot (1 - \epsilon_{\Delta h_t, t}),$$

where the first equation is by $D_{\alpha^*} \subseteq D_{\alpha^*}$, and the second equality is by the error-target independence condition. For the denominator, we consider the worst case that all errors are out of $D_{h_t}$ and then $|D_{\alpha^*}| \leq |D_{\alpha^*}| + \epsilon_{\Delta h_t, t}$. ■

Similar to Theorem 2, we can bound from below the average success probability of sampling from the positive area of the learnt hypothesis.

$$\Pr_{h_t} \geq \frac{1 - \eta}{2 \KL(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{\alpha^*}})} \cdot t \cdot \frac{|D_{\alpha^*}|(1 - \epsilon_{\Delta h_t, t})}{|D_{\alpha^*}| + \epsilon_{\Delta h_t, t}} - |D_{\alpha^*}| \cdot \frac{1}{2 \KL(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{\alpha^*}})}.$$

We compare the uniform search with the SAC algorithms using uniform sampling within $D_{h_t}$, i.e., $\KL(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{h_t}}) = 0$, which is an optimistic situation. Then by Lemma 3

$$\Pr_{h_t} \geq \frac{1 - \eta}{2 \KL(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{\alpha^*}})} \cdot t \cdot \frac{|D_{\alpha^*}|(1 - \epsilon_{\Delta h_t, t})}{|D_{\alpha^*}| + \epsilon_{\Delta h_t, t}} - |D_{\alpha^*}| \cdot \frac{1}{2 \KL(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{\alpha^*}})}.$$

By plugging $\epsilon_{\Delta h_t, t} \leq \frac{\epsilon_{\Delta h_t}}{1 - \sqrt{\frac{\epsilon_{\Delta h_t}}{2 \KL(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{\alpha^*}})}}} = Q \cdot \epsilon_{D_t}$, where $\epsilon_{D_t}$ is the expected error rate of $h_t$ under the distribution $\mathcal{D}_t = \lambda \mathcal{U}_{D_{h_t}} + (1 - \lambda) \mathcal{U}_X$ and $Q = (1 - \frac{1}{2 \KL(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{\alpha^*}})})^{-1}$.

$$\Pr_{h_t} \geq \frac{1 - \eta}{2 \KL(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{\alpha^*}})} \cdot t \cdot \frac{|D_{\alpha^*}|(1 - Q \cdot \epsilon_{D_t})}{|D_{\alpha^*}| + Q \cdot \epsilon_{D_t}}. \quad (1)$$

Note from Lemma 2 that, the convergence rate of the error is $O(\frac{1}{m})$ ignoring other variables and logarithmic terms from Lemma 2. We assume that SAC algorithms using learning algorithms
with convergence rate $\Theta(\frac{1}{m})$. We then find that such SAC$_t$ algorithms cannot exponentially improve the uniform search in the worst case, as Proposition 1.

**Proposition 1** Using learning algorithms with convergence rate $\Theta(\frac{1}{m})$, $\forall f$, $\alpha^* > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$, if the query complexity of the uniform search is superpoly($\frac{1}{\alpha^*}, n, \frac{1}{\delta}$), the query complexity of SAC$_t$ algorithms is also superpoly($\frac{1}{\alpha^*}, n, \frac{1}{\delta}$) in the worst case.

**Proof.** The query complexity of the uniform search being superpoly($\frac{1}{\alpha^*}, n, \frac{1}{\delta}$) implies that

$$\frac{1}{Pr_a} = 1 = \frac{1}{|D_{\alpha^*}|} = \text{superpoly}(\frac{1}{\alpha^*}, n, \frac{1}{\delta}).$$

For the SAC$_t$ algorithms, if we ask the learning algorithm to produce a classifier with error rate $\frac{1}{\text{superpoly}(\frac{1}{\alpha^*}, n, \frac{1}{\delta})}$, it will require superpoly($\frac{1}{\alpha^*}, n, \frac{1}{\delta}$) number of samples in the worst case, so that the proposition holds. To avoid this, we can only expect the error rate to be $\frac{1}{\text{poly}(\frac{1}{\alpha^*}, n, \frac{1}{\delta})}$ in order to keep the query complexity at each iteration small.

Meanwhile, we can only have $T = \text{poly}(\frac{1}{\alpha^*}, n, \frac{1}{\delta})$ iterations otherwise we will have super-polynomial number of samples.

Following the optimistic case of Eq.[1], since $Q \geq 1$, we consider one more optimistic situation that $Q = 1$. Let $\eta = 0.5$. Even though, in the worst case that $|D_{\alpha^*}| = |D_{\alpha^*}| + Qe_{D_t}$, we can have that

$$Pr_a = \frac{1}{2} \frac{X}{t=1} \frac{m_1}{m_t} \frac{|D_{\alpha^*}|(1-e_{D_t})}{|D_{\alpha^*}| + e_{D_t}} = \frac{1}{\text{poly}(\frac{1}{\alpha^*}, n, \frac{1}{\delta})} \frac{\text{superpoly}(\frac{1}{\alpha^*}, n, \frac{1}{\delta})}{\text{poly}(\frac{1}{\alpha^*}, n, \frac{1}{\delta})},$$

where it is noted that as long as $e_{D_t} = \text{poly}(\frac{1}{\alpha^*}, n, \frac{1}{\delta})$ the value of $|D_{\alpha^*}|$ cannot affect the result. Then substituting $Pr_a$ into Theorem 1 obtains the total samples $m_t = \text{superpoly}(\frac{1}{\alpha^*}, n, \frac{1}{\delta})$ that proves the proposition.

The proposition implies that the SAC$_t$ algorithms can face the same barrier as that of the uniform search. Nevertheless, the SAC$_t$ algorithms can still improve the uniform search within a polynomial factor. We show this by case studies.

**On Sphere Function Class:**

Given the solution space $X_n = \{(x_1, \ldots, x_n) \mid \forall i = 1, \ldots, n : x_i \in [0, 1]\}$, the Sphere Function class is $F^*_n = \{f^*_n \mid f^*_n : X_n \rightarrow \mathbb{R}\}$ where

$$f^*_n(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - x^*_i)^2.$$

Obviously, $|X_n| = 1$, $f^*_n \in [0, 1]$ is convex, and the optimal value is 0. It is important to notice that the volume of a $n$-dimensional hyper-sphere with radius $r$ is $\frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)}r^n$, where $\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}dt$, so that $|D_{\alpha^*}| = \int_{\frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)}}^{\infty} \frac{n\alpha}{2} \frac{n\alpha}{2} dt$.

The volume of a $n$-dimensional hyper-sphere with radius $r$ is $\frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)}r^n$, where $\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}dt$, so that $|D_{\alpha^*}| = \int_{\frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)}}^{\infty} \frac{n\alpha}{2} \frac{n\alpha}{2} dt$.

Proposition 2 For any function in $F^*_n$ and any approximation level $\alpha^* > 0$, SAC$_t$ algorithms can achieve the PAA query complexity, for any $n \geq 2$:

$$O\left(\frac{1}{\alpha^*}\right)^{\frac{n}{2}} \log \frac{1}{\alpha^*} \log \frac{1}{\alpha^*} \log \frac{1}{\alpha^*}$$

with probability at least $1 - \delta$.

**Proof.** We choose $\alpha_t = \frac{1}{\sqrt{t}}$ for all $t$, and use the number of iterations $T$ to approach $|D_{\alpha^*}| = |D_{\alpha^*}|$, for the approximation level $\alpha^*$. Solving this equation with the sphere volume results in $T = \log \frac{|C_{\alpha^*}|}{\pi^{\frac{n}{2}}/\Gamma\left(\frac{n}{2}+1\right)}$. We let the SAC$_t$ algorithm run $T = \log \frac{1}{\alpha^*}$ number of iterations. We assume $\log \frac{1}{\alpha^*}$ is an integer for simplicity, which does not affect the generality.
We show this by comparing SAC algorithms. SAC algorithms can handle problems with some local optima. Inherited from EAs, the Sphere Function class only consists of convex functions. As modeling EAs, SAC algorithms should be expected to be applied on the complex problems, while the Sphere Function class is non-convex and non-differentiable with some local optima, as depicted in Figure 1.

![Image of function landscape](image_url)

**Fig. 1.** The landscape of function $f_{spike}^0(x)$ in $[-1, 1]$.

It is easy to know $\min_{x \in \mathbb{X}_n} f(x) = 0$ and $\max_{x \in \mathbb{X}_n} f(x) \leq 1$ for any $f \in \mathcal{F}_{spike}$. For any $\alpha > 0$, we can bound the area $|D_\alpha| = [C_n \alpha^n, C_n(3\alpha)^n]$, where $C_n = \Theta\left(\frac{2\pi}{\sqrt{\pi n}}\right)$.

The Spike functions are non-convex and non-differentiable with some local optima. Using this SAC algorithm to minimize any member in the function class $\mathcal{F}_{spike}$, we obtain the PAA query complexity as in Proposition 3.

**Proposition 3** For any function in $\mathcal{F}_{spike}$ and any approximation level $\alpha^* > 0$, SAC1 algorithms can achieve the PAA query complexity

$$O\left(\frac{1}{\alpha^*}\right)^n \log \frac{1}{\delta} + n \log \frac{1}{\sqrt{\alpha^*}},$$

with probability at least $1 - \delta$.

**Proof.** For any function in $\mathcal{F}_{spike}$, we note that the function is convex in $D_\alpha$ when $\alpha$ is smaller than 0.05. We set $\alpha_1 = \frac{1}{20}$, so that when $t \geq 5$, the SAC1 algorithm with $\mathcal{L}_{spike}$ will deal with a convex function and thus the training error is zero. We use the number of iterations $T$ to achieve $|D_{\alpha T}| = |D_{\alpha_1}|$. Since $|D_{\alpha}| = [C_n \alpha^n, C_n(3\alpha)^n]$, we can obtain $T \geq \log \left(\frac{3(C_n)}{\sqrt{\alpha}}\right)$. We let the SAC1 algorithm run $T = \log \frac{1}{\sqrt{\alpha}}$ number of iterations and assume that $\log \frac{1}{\sqrt{\alpha}}$ is an integer.

In iteration $t \geq 5$, we want the error of the hypothesis $h_t$, $\epsilon_{D_t}$, to be $\frac{1}{20}$. Since the training error can be zero, we can solve the required sample size $m_t \leq m_T = O\left(\frac{n}{\sqrt{\alpha}} \log \frac{1}{\sqrt{\alpha}}\right)^2$. We thus obtain

$$\Pr_h \geq \frac{1}{2 \log \frac{1}{\sqrt{\alpha}}} \sum_{t=1}^{T} \frac{C_n(\alpha^*)^2}{2} \log \frac{1}{\sqrt{\alpha}} \frac{C_n(\alpha)}{2} + 2^{0.5}$$

We then follow Eq. [1]. We use uniform sampling within $D_h$, then $Q = \frac{1}{T}$. Letting the SAC1 algorithms use $m_T$ number of samples in every iteration, $\lambda = 0.5$ and $\eta = 0.5$, we have

$$m_t \leq m_T = O\left(\frac{n}{\sqrt{\alpha}} \log \frac{1}{\sqrt{\alpha}}\right)^2$$

using the inequality $\log x \leq \frac{1}{2} x - (\log 2c + 1)$ for any $x > 0$ and any $c > 0$. We thus obtain $\Pr_h \geq O\left(\frac{1}{\alpha^*}\right)^n \log \frac{1}{\delta} + n \log \frac{1}{\sqrt{\alpha^*}}$, which is $O\left(\frac{1}{\alpha^*}\right)^n \log \frac{1}{\delta} + n \log \frac{1}{\sqrt{\alpha^*}}$ using a constant $m_0$ and the max is upper bounded by plus.
long as the non-convexity is not quite severe, like the Spike algorithms, and thus the search process is interfered. But as

\[ \max \]

So we obtain the query complexity from Theorem 1, letting

\[ X \log \frac{1}{\sqrt{\alpha}} \ln \frac{1}{\delta} + n \log \frac{1}{\sqrt{\alpha}} \]

as the max is upper bounded by plus.

We observe from the proof that the non-convexity can result in non-zero training error for the learning algorithms in SAC₁ algorithms, and thus the search process is interfered. But as long as the non-convexity is not quite severe, like the Spike Functions, SAC₁ algorithms are not significantly affected, and can still be better than the uniform search by a factor near \( \frac{1}{\sqrt{\alpha}} / \log \frac{1}{\sqrt{\alpha}} \).

B. A Super-Polynomial Acceleration Condition

We have shown in Proposition 4 that SAC₁ algorithms using common classification algorithms cannot super-polynomially improve from the uniform search in the worst case. An interesting question is therefore raised that when the super-polynomial improvement is possible.

Learned from the proof of Proposition 4, a straightforward way is to use a powerful classification algorithm with exponentially improved sample complexity, i.e., \( O(\ln \frac{1}{\delta}) \), so that only a polynomial number of samples is required to achieve a super-polynomially small error. Several active learning algorithms can do this in some circumstances (e.g. [7], [39]). Applying active learning algorithms needs a small modification of SAC₁. In iteration \( t \), instead of sampling from the uniform distribution in \( D_{b_t} \), the sampling is guided by the classifier. Nevertheless, the achieved error is still evaluated under the original (uniform) distribution. Using such learning algorithms denoted as \( L_{\alpha} \), we achieve Proposition 4 showing a super-polynomial acceleration from the uniform search on Sphere Functions.

**Proposition 4** For any function in \( F_{\alpha}^{n_{\text{sphere}}} \) and any approximation level \( \alpha > 0 \), SAC₁ algorithms using \( L_{\alpha} \) can achieve the PAA query complexity, for any \( n \geq 2 \),

\[ O \left( \frac{1}{\alpha} \ln \frac{1}{\alpha} + n \log \frac{1}{\alpha} \right) \]

with probability at least \( 1 - \delta \).

**Proof.** We choose \( \alpha_t = \frac{1}{\sqrt{T}} \) for all \( t \), and use the number of iterations \( T \) to approach \( |D_{\alpha_t}| = |D_{\alpha^*}| \), for the approximation level \( \alpha^* \). Solving this equation with the sphere volume results in \( T = \log \left( \frac{C_{\alpha^*}^{3/4}}{\sqrt{T}} \right) \). We let the SAC₁ algorithm run \( T = \log \frac{1}{\sqrt{\alpha}} \) number of iterations. We assume \( \log \frac{1}{\sqrt{\alpha}} \) is an integer for simplicity, which does not affect the generality.

In iteration \( t \), using \( L_{\alpha} \), we want the error of the hypothesis \( h_t, \epsilon_{\alpha_t} \), to be \( \frac{1}{\sqrt{T^n_{\alpha}}} \). Since the \( L_{\alpha} \) has the sample complexity \( O(\ln \frac{1}{\delta}) \), we ask for a hypothesis with zero training error, which requires the sample size \( m_t = O(tn) = O(n \log \frac{1}{\alpha_t}) \) with \( \eta \) being a constant.

We thus obtain

\[ P \sum_{t=1}^{T} m_t = O \left( n \log \frac{1}{\alpha_t} \right) \]

Following Eq.(1), we use uniform sampling within \( D_{b_t} \), then \( Q = \frac{1}{\sqrt{T}} \). Letting the SAC₁ algorithms use \( m_t \) number of samples in every iteration, \( \lambda = 0.5 \) and \( \eta = 0.5 \), we have

\[ P_{\alpha_t} \geq \frac{1}{2 \log \frac{1}{\alpha_t}} \sum_{t=1}^{T} \frac{1}{C_{\alpha_t}^{3/4}} \left( |D_{\alpha^*}| + Q \epsilon_{\alpha_t} \right) \]

\[ \geq \frac{C_{\alpha_t}^{3/4}}{2 \log \frac{1}{\alpha_t}} \sum_{t=1}^{T} \frac{1}{C_{\alpha_t}^{3/4}} \left( 1 - 2 \left( \frac{1}{\sqrt{\alpha}} \right)^{3/4} \right) \]

\[ \geq \frac{C_{\alpha_t}^{3/4}}{2 \log \frac{1}{\alpha_t}} \left( \frac{1}{C_{\alpha_t}^{3/4}} \right) \left( 1 - 2 \left( \frac{1}{\sqrt{\alpha}} \right)^{3/4} \right) \]

\[ \geq \frac{C_{\alpha_t}^{3/4}}{2 \log \frac{1}{\alpha_t}} \left( \frac{1}{C_{\alpha_t}^{3/4}} \right) \left( 1 - 2 \left( \frac{1}{\sqrt{\alpha}} \right)^{3/4} \right) \]

\[ \geq \frac{C_{\alpha_t}^{3/4}}{2 \log \frac{1}{\alpha_t}} \left( \frac{1}{C_{\alpha_t}^{3/4}} \right) \left( 1 - 2 \left( \frac{1}{\sqrt{\alpha}} \right)^{3/4} \right) \]

\[ \geq \frac{C_{\alpha_t}^{3/4}}{2 \log \frac{1}{\alpha_t}} \left( \frac{1}{C_{\alpha_t}^{3/4}} \right) \left( 1 - 2 \left( \frac{1}{\sqrt{\alpha}} \right)^{3/4} \right) \]

So we obtain the query complexity from Theorem 1, letting \( m_0 \) be a constant,

\[ O \left( \frac{1}{\alpha} \ln \frac{1}{\alpha} + n \log \frac{1}{\alpha} \right) \]

which is \( O \left( \frac{1}{\alpha} \ln \frac{1}{\delta} + n \log \frac{1}{\alpha} \right) \).

Meanwhile, we are more interested in exploring conditions under which the super-polynomial improvement is possible without requiring such powerful learning algorithms. For this purpose, we find the one-side-error condition.

**Condition 2** (One-Side-Error) In SAC₁ algorithms, for any \( t \) and any \( x \in X \), if \( x \in D_{b_t} \Delta D_{\alpha_t} \), it must hold that \( x \in D_{\alpha_t} \).

The condition implies that \( h_t \) can only make false-negative errors, i.e., wrongly classifies positive samples (inside \( D_{\alpha_t} \)) as negative, but no false-positive errors. One practical way to approach this condition is through the cost-sensitive classifiers (\[14, 46\]) with a very large mis-classification cost for negative samples. We call SAC₁ algorithms that are further under this condition as SAC₂ algorithms.

**Lemma 5** For SAC₂ algorithms, it holds for all \( t \) that

\[ |D_{b_t}| \leq |D_{\alpha_t}| \]

**Proof.** Note that for training \( h_t \) we label the samples from \( D_{\alpha_t} \) as positive and label the rest as negative. Since \( h_t \) only makes false-negative errors, i.e., every error is in \( D_{\alpha_t} \), we have \( D_{b_t} \subseteq D_{\alpha_t} \), which implies the lemma.

**Lemma 6** shows the one-side-error condition controls the size \( |D_{b_t}| \) to be bounded by \( |D_{\alpha_t}| \). Thus we can refine Lemma 5 as Lemma 6.

**Lemma 6** For SAC₂ algorithms, it holds for all \( t \) that

\[ \frac{|D_{\alpha_t} \cap D_{b_t}|}{|D_{b_t}|} \geq \frac{|D_{\alpha_t} \cap D_{b_t}|}{|D_{\alpha_t}|} \]

where \( \epsilon_{\alpha_t} \) is the expected error rate of \( h_t \) under \( U_X \).

**Proof.** Since the SAC₁ algorithm is also a SAC₂ algorithm, incorporating Lemma 5 into Lemma 3 proves the lemma.

We assume that \( L_{\alpha} \) is a learning algorithm that not only behaves like \( L_{\alpha} \) but also results a hypothesis
Proof. By Lemma 6, of iterations. We assume that we only ask for a random guess classification (i.e., error rate 1/2) in the proof of Proposition 5.

**Proposition 5** For any function in $F^n_{\text{sphere}}$ and any approximation level $\alpha^* > 0$, SAC algorithms can achieve the PA query complexity

$$O \log \frac{1}{\alpha^*} (\ln \frac{1}{\delta} + n),$$

with probability at least $1 - \delta$.

Proof. By Lemma 6

$$\left| D_{\alpha^*} \cap D_{t_i} \right| \geq D_{\alpha^*} \left( 1 - Q \epsilon_{D_i} \right),$$

where $\epsilon_{D_i}$ is the error of $h_t$ under its original distribution $D_t$, and $Q$ is the resulting factor of changing the distribution.

Let $\alpha_t = \frac{1}{\alpha}$ for all $t$, and use the number of iterations $T$ to achieve $\left| D_{\alpha^*} \right| = \left| D_{\alpha^*} \right|$, for the approximation level $\alpha^*$. Solving this equation with the sphere volume results in $T = \log \frac{C_n(\alpha^*)}{\alpha^*}$. Let the SAC algorithm run $T = \log \frac{1}{\alpha^*}$ number of iterations. We assume $\log \frac{1}{\alpha^*}$ is an integer for simplicity, which does not affect the generality.

In iteration $t$, using $L_{\text{sphere}}^+$, we want the error of the hypothesis $h_t, \epsilon_{D_t}$, to be a constant $\frac{1}{2}$. Since $L_{\text{sphere}}^+$ produces a hypothesis with zero training error, to achieve $\epsilon_{D_t} \leq \frac{1}{2}$ requires the number of samples in $O(n)$. We thus obtain

$$T = \sum_{t=1}^{n} m_t = O \ n \log \frac{1}{\alpha^*}.$$

We then follow Eq. (1). We use uniform sampling within $D_{h_t}$, then $Q = \frac{1}{\alpha^*}$. Letting the SAC algorithm use $m_T$ number of samples in every iteration, $\lambda = \frac{1}{3}$ and $\eta = 0.5$, we have

$$P_{h_t} \geq \frac{1}{2 \log \frac{1}{\alpha^*}} \sum_{t=1}^{T} \log \frac{1}{\alpha^*} \left| D_{\alpha^*} \left( 1 - Q \epsilon_{D_i} \right) \right| $$

$$\geq \frac{1}{2 \log \frac{1}{\alpha^*}} \sum_{t=1}^{T} \log \frac{1}{\alpha^*} \left| D_{\alpha^*} \right|$$

$$= \frac{C_n(\alpha^*)^{1/2}}{8 \log \alpha^*} \sum_{t=1}^{T} \log \frac{1}{\alpha^*} \left| D_{\alpha^*} \left( 1 - Q \epsilon_{D_i} \right) \right| $$

$$\geq \frac{C_n(\alpha^*)^{1/2}}{8 \log \alpha^*} \left( \alpha^* \right)^{1/2} - 1 = \Omega \log \frac{1}{\alpha^*}.$$

So we obtain from Theorem 1 the query complexity of the SAC algorithm, letting $m_0$ be a constant,

$$O \max \log \frac{1}{\alpha^*} \ln \frac{1}{\delta} \ n \log \frac{1}{\alpha^*},$$

which is $O \ln \frac{1}{\alpha^*} (\ln \frac{1}{\delta} + n)$.

Proposition 5 shows a super-polynomial improvement from the complexity of the uniform search. It is interesting to note that we only ask for a random guess classification (i.e., error rate 1/2) in the proof of Proposition 5.

IV. DISCUSSIONS AND CONCLUSIONS

This paper describes the sampling-and-learning (SAL) framework which is an abstract summary of a range of EAs. The SAL framework allows us to investigate the general performance of EAs from a statistical view. We show that the SAL framework directly admits a general upper bound on the PA query complexity, which is the number of fitness evaluations before an approximate solution is found with a probability.

Focusing on SAC algorithms, which are SAL algorithms using classification learning algorithms, we give a specific performance upper bound, and compare with uniform random search. We find two conditions that drastically effect the performance of SAC algorithms. Under the error-target independence condition, which assumes that the error of the learned classifier in each iteration is independent with the target approximation area, the SAC algorithms can obtain a polynomial improvement over the uniform search, but not a super-polynomial improvement. We demonstrate the improvement using the Sphere Function class consisting of convex functions as well as the Spike Function class consisting of non-convex functions. Further incorporating the one-side-error condition, which assumes that the classification only makes false-negative errors, the SAC algorithms can obtain a super-polynomial improvement over the uniform search.

On the one hand, our results show that the property of classification error in SAC algorithms greatly impacts the performance, which was never touched in previous studies, as far as we know. We expect the work could guide the design of novel search algorithms. On the other hand, how to satisfy the conditions is a non-trivial practical issue.

In the case study on the Sphere Function class, we find that a learning error rate no more than the random guess is sufficient to achieve a super-polynomial improvement under the conditions. This implies that an accurate learning algorithm may not be necessary for a good SAC algorithm. It is interesting that a recent work [4] also noticed that a learnable concept is not necessary for the trial-and-error search with a computation oracle.

In this paper, the SAC algorithms are analyzed in continuous domains, while the main body of theoretical studies of evolutionary algorithms focuses on the discrete domains. Thus understanding the performance of SAC algorithms in discrete domains is our future work. Moreover, in the SAC algorithms analyzed in this paper, the learning algorithm does not utilize the last hypothesis or the last data set. It would be interesting to investigate whether considering them will bring any significant difference.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation of China (61375061) and the Jiangsu Science Foundation (BK2012303).

REFERENCES


