# Analysis of Noisy Evolutionary Optimization When Sampling Fails

Chao Qian
University of Science and Technology of China
Hefei, China

Chao Bian
University of Science and Technology of China
Hefei, China

Yang Yu
Nanjing University
Nanjing, China

Ke Tang
Shenzhen Key Laboratory of Computational Intelligence, Southern University of Science and Technology
Shenzhen, China

Xin Yao
Shenzhen Key Laboratory of Computational Intelligence, Southern University of Science and Technology
Shenzhen, China

## ABSTRACT

In noisy evolutionary optimization, sampling is a common strategy to deal with noise, which evaluates the fitness of a solution multiple times (called *sample size*) independently and then uses the average to approximate the true fitness. Previous studies mainly focused on the empirical design of efficient sampling strategies, and the few theoretical analyses mainly proved the effectiveness of sampling with a fixed sample size in some situations. There are many fundamental theoretical issues to be addressed. In this paper, we first investigate the effect of sample size. By analyzing the (1+1)-EA on noisy LeadingOnes, we show that as the sample size increases, the running time can reduce from exponential to polynomial, but then return to exponential. This discloses that a proper sample size is crucial in practice. Then, we investigate what other strategies can work when sampling with any fixed sample size fails. By two illustrative examples, we prove that using parent populations can be better, and if using parent populations is also ineffective, adaptive sampling (i.e., sampling with an adaptive sample size) can work.

## CCS CONCEPTS

• **Theory of computation** → **Theory of randomized search heuristics**;

## KEYWORDS

Noisy optimization, evolutionary algorithms, sampling, running time analysis, computational complexity

**ACM Reference Format:**
Chao Qian, Chao Bian, Yang Yu, Ke Tang, and Xin Yao. 2018. Analysis of Noisy Evolutionary Optimization When Sampling Fails. In *GECCO '18: Genetic and Evolutionary Computation Conference, July 15–19, 2018, Kyoto, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3205455.3205643


*GECCO '18, July 15–19, 2018, Kyoto, Japan*

ACM ISBN 978-1-4503-5618-3/18/07…$15.00
https://doi.org/10.1145/3205455.3205643

## 1 INTRODUCTION

Evolutionary algorithms (EAs) are a type of general-purpose randomized optimization algorithms, inspired by natural evolution. They have been widely applied to solve real-world optimization problems, which are often subject to noise. Sampling is a popular strategy for dealing with noise: to estimate the fitness of a solution, it evaluates the fitness multiple ($m$) times (called *sample size*) independently and then uses the sample average to approximate the true fitness. Sampling reduces the variance of noise by a factor of $m$, but also increases the computation time for the fitness estimation of a solution by $m$ times. Previous studies mainly focused on the empirical design of efficient sampling methods, e.g., adaptive sampling [3, 4], which dynamically decides the sample size $m$ for each solution in each generation. The theoretical analysis on sampling was rarely touched.

Due to their sophisticated behaviors of mimicking natural phenomena, the theoretical analysis of EAs is difficult. Much effort thus has been devoted to understanding the behavior of EAs from a theoretical point of view [2, 14], but most of them focus on noise-free optimization. The presence of noise further increases the randomness of optimization, and thus also the difficulty of analysis.

For running time analysis (one essential theoretical aspect) in noisy evolutionary optimization, only a few results have been reported. The classic (1+1)-EA algorithm was first studied on the OneMax and LeadingOnes problems under various noise models [8, 12, 18]. The results showed that the (1+1)-EA is efficient only under low noise levels, e.g., for the (1+1)-EA solving OneMax in the presence of one-bit noise, the maximal noise level of allowing a polynomial running time is $\log n/n$, where the noise level is characterized by the noise probability $p \in [0, 1]$ and $n$ is the problem size. Later studies mainly proved the robustness of different strategies to noise, including using populations [5, 12, 17], sampling [18, 19] and threshold selection [20]. For example, the ($\mu$+1)-EA with $\mu = \Theta(\log n)$ can solve OneMax in polynomial time even if the probability of one-bit noise reaches 1. Note

that there was also a sequence of papers analyzing the running time of the compact genetic algorithm [11] and a simple ant colony optimization algorithm [6, 9, 10, 21] solving noisy problems, including OneMax as well as the combinatorial optimization problem single destination shortest paths.

The very few running time analyses involving sampling [18, 19] mainly showed the effectiveness of sampling with a large enough fixed sample size $m$. For example, for the (1+1)-EA solving OneMax under one-bit noise with $p = \omega(\log n/n)$, using sampling with $m = 4n^3$ can reduce the running time exponentially. In addition, Akimoto et al. [1] proved that using sampling with a large enough $m$ can make optimization under additive unbiased noise behave as optimization in a noise-free environment. However, there are still many fundamental theoretical issues that have not been addressed, e.g., how the sample size can affect the effectiveness of sampling, and what other strategies can work when sampling fails.

In this paper, we first theoretically investigate the effect of sample size. It may be believed that once the sample size $m$ reaches an effective value, the running time will always be polynomial as $m$ continues to increase. We give a counterexample, i.e., the (1+1)-EA solving LeadingOnes under one-bit noise with $p = 1$. Qian et al. [18] have shown that the running time will reduce from exponential to polynomial when $m = 4n^4 \log n/15$. We prove that the running time will return to exponential when $m \geq n^5$. Our analysis suggests that the selection of sample size should be careful in practice.

Then, we theoretically compare the two strategies of using parent populations and sampling on the robustness to noise. Previous studies have shown that both of them are effective for solving OneMax under one-bit noise [12, 18, 19], while using sampling is better for solving OneMax under additive Gaussian noise [19]. Here, we complement this comparison by constructing two specific noisy OneMax problems. For one of them, using parent populations is better, while for the other, using neither parent populations nor sampling is effective. For the latter case, we further prove that using adaptive sampling can reduce the running time exponentially, which provides some theoretical justification for the good empirical performance of adaptive sampling in practice [22, 26].

The rest of this paper is organized as follows. Section 2 introduces some preliminaries. Section 3 analyzes the effect of sample size. Sections 4 and 5 then show the effectiveness of using parent populations and adaptive sampling when sampling fails. Finally, Section 6 concludes the paper.

## 2 PRELIMINARIES

In this section, we first introduce the EAs studied in this paper, then introduce the sampling strategy, and finally present the analysis tools that we use throughout this paper.

### 2.1 Evolutionary Algorithms

The (1+1)-EA as described in Algorithm 1 maintains only one solution, and iteratively tries to produce one better solution. The ($\mu$+1)-EA as described in Algorithm 2 uses a parent population size $\mu$. In each iteration, it also generates one new solution $x'$, and then uses $x'$ to replace the worst solution in the population $P$ if $x'$ is not worse. When $\mu = 1$, the ($\mu$+1)-EA degenerates to the (1+1)-EA. Note that for the ($\mu$+1)-EA, a slightly different updating rule is also used [24]: $x'$ is simply added into $P$ and then the worst solution in $P \cup \{x'\}$ is deleted. Our results about the ($\mu$+1)-EA derived in the paper also apply to this setting.

In noisy optimization, only a noisy fitness value $f^{\mathrm{n}}(x)$ instead of the exact one $f(x)$ can be accessed. Note that in our analysis, we assume that the reevaluation strategy is used as in [6, 8, 12]. That is, besides evaluating the noisy fitness $f^{\mathrm{n}}(x')$ of the offspring solution, the noisy fitness values of parent solutions will be reevaluated in each iteration. The running time of EAs is usually defined as the number of fitness evaluations needed to find an optimal solution w.r.t. the true fitness function $f$ for the first time [1, 8, 12].

ALGORITHM 1 ((1+1)-EA). *Given a function $f$ over $\{0,1\}^n$ to be maximized, it consists of the following steps:*
1. *Let $x$ be a uniformly chosen solution.*
2. *Repeat until the termination condition is met*
3. *$x' :=$ flip each bit of $x$ independently with prob. $1/n$.*
4. *if $f(x') \geq f(x)$ then $x := x'$.*

ALGORITHM 2 (($\mu$+1)-EA). *Given a function $f$ over $\{0,1\}^n$ to be maximized, it consists of the following steps:*
1. *Let $P$ be a set of $\mu$ uniformly chosen solutions.*
2. *Repeat until the termination condition is met*
3. *$x :=$ uniformly selected from $P$ at random.*
4. *$x' :=$ flip each bit of $x$ independently with prob. $1/n$.*
5. *Let $z \in \arg\min_{z \in P} f(z)$; ties are broken randomly.*
6. *if $f(x') \geq f(z)$ then $P := (P \setminus \{z\}) \cup \{x'\}$.*

### 2.2 Sampling

Sampling as described in Definition 2.1 is a common strategy to deal with noise. It approximates the true fitness $f(x)$ using the average of a number of random evaluations. The number $m$ of random evaluations is called the *sample size*. Note that $m = 1$ is equivalent to that sampling is not used. Qian et al. [18, 19] have theoretically shown the robustness of sampling to noise. Particularly, they proved that by using sampling with some fixed sample size, the running time of the (1+1)-EA for solving OneMax and LeadingOnes under noise can reduce from exponential to polynomial.

*Definition 2.1 (Sampling).* Sampling first evaluates the fitness of a solution $m$ times independently and obtains the noisy fitness values $f_1^{\mathrm{n}}(x), f_2^{\mathrm{n}}(x), \ldots, f_m^{\mathrm{n}}(x)$, and then outputs their average, i.e., $\hat{f}(x) = \sum_{i=1}^{m} f_i^{\mathrm{n}}(x)/m$.

Adaptive sampling dynamically decides the sample size for each solution in the optimization process, instead of using a fixed size. For example, one popular strategy [3, 4] is to first estimate the fitness of two solutions by a small number of samples, and then sequentially increase samples until the difference can be significantly discriminated. It has been found well useful in many applications [22, 26], while there has been no theoretical work supporting its effectiveness.

### 2.3 Analysis Tools

EAs often generate offspring solutions only based on the current population, thus, an EA can be modeled as a Markov chain $\{\xi_t\}_{t=0}^{+\infty}$ (e.g., in [13, 25]) by taking the EA's population space $\mathcal{X}$ as the chain's state space (i.e., $\xi_t \in \mathcal{X}$) and taking the set $\mathcal{X}^*$ of all optimal populations as the chain's target state space. Note that the population space $\mathcal{X}$ consists of all possible populations, and an optimal population contains at least one optimal solution.

Given a Markov chain $\{\xi_t\}_{t=0}^{+\infty}$ and $\xi_{\hat{t}}$, we define its *first hitting time* as $\tau = \min\{t \mid \xi_{\hat{t}+t} \in \mathcal{X}^*, t \geq 0\}$. The mathematical expectation of $\tau$, $\mathrm{E}(\tau \mid \xi_{\hat{t}}) = \sum_{i=0}^{+\infty} i \cdot \mathrm{P}(\tau = i \mid \xi_{\hat{t}})$, is called the *expected first hitting time* (EFHT). If $\xi_0$ is drawn from a distribution $\pi_0$, $\mathrm{E}(\tau \mid \xi_0 \sim \pi_0) = \sum_{\xi_0 \in \mathcal{X}} \pi_0(\xi_0)\mathrm{E}(\tau \mid \xi_0)$ is called the EFHT of the chain over the initial distribution $\pi_0$. Thus, the expected running time of the ($\mu$+1)-EA starting from $\xi_0 \sim \pi_0$ is $\mu + (\mu+1) \cdot \mathrm{E}(\tau \mid \xi_0 \sim \pi_0)$, where the term $\mu$ corresponds to evaluating the initial population, and the factor ($\mu$+1) corresponds to evaluating the offspring solution $x'$ and reevaluating the $\mu$ parent solutions in each iteration. For the (1+1)-EA, the expected running time is calculated by setting $\mu = 1$, i.e., $1 + 2 \cdot \mathrm{E}(\tau \mid \xi_0 \sim \pi_0)$. For the (1+1)-EA with sampling, it becomes $m + 2m \cdot \mathrm{E}(\tau \mid \xi_0 \sim \pi_0)$, since estimating the fitness of a solution needs $m$ independent evaluations. Note that we consider the expected running time of an EA starting from a uniform initial distribution in this paper.

In the following, we give two drift theorems that will be used to derive upper and lower bounds on the EFHT of Markov chains in the paper.

LEMMA 2.2 (MULTIPLICATIVE DRIFT [7]). *Given a Markov chain $\{\xi_t\}_{t=0}^{+\infty}$ and a distance function $V$ over $\mathcal{X}$, if for any $t \geq 0$ and any $\xi_t$ with $V(\xi_t) > 0$, there exists $c > 0$ such that $\mathrm{E}(V(\xi_t) - V(\xi_{t+1}) \mid \xi_t) \geq c \cdot V(\xi_t)$, then the EFHT satisfies that $\mathrm{E}(\tau \mid \xi_0) \leq \frac{1+\log(V(\xi_0)/V_{\min})}{c}$, where $V_{\min}$ denotes the minimum among all possible positive values of $V$.*

LEMMA 2.3 (NEGATIVE DRIFT [15, 16]). *Let $X_t$, $t \geq 0$, be real-valued random variables describing a stochastic process. Suppose there exists an interval $[a, b] \subseteq \mathbb{R}$, two constants $\delta, \epsilon > 0$ and, possibly depending on $l := b - a$, a function $r(l)$ satisfying $1 \leq r(l) = o(l/\log(l))$ such that for all $t \geq 0$:*

(1) $\mathrm{E}(X_t - X_{t+1} \mid a < X_t < b) \leq -\epsilon$,

(2) $\mathrm{P}(|X_{t+1} - X_t| \geq j \mid X_t > a) \leq r(l)/(1+\delta)^j$ *for* $j \in \mathbb{N}^+$.

*Then there is a constant $c > 0$ such that for $T := \min\{t \geq 0 : X_t \leq a \mid X_0 \geq b\}$ it holds $\mathrm{P}(T \leq 2^{cl/r(l)}) = 2^{-\Omega(l/r(l))}$.*

## 3 THE EFFECT OF SAMPLE SIZE

Previous studies [18, 19] have shown that for noisy evolutionary optimization, sampling with some fixed sample size $m$ can reduce the running time exponentially in some situations. For example, for the (1+1)-EA solving OneMax under one-bit noise with the noise probability $p = \omega(\log n/n)$, the expected running time is super-polynomial [8]; while by using sampling with $m = 4n^3$, the running time reduces to polynomial [18]. Then, a natural question is that whether the running time will always be polynomial by using any polynomially bounded sample size larger than the effective $m$. It may be believed that the answer is yes, since the sample size $m$ has been effective and using a larger sample size will make the fitness estimation more accurate. For example, for the (1+1)-EA solving OneMax under one-bit noise, it is easy to see from Lemma 6 in [18] that using a larger sample size than $4n^3$ will make the probability of accepting a true worse solution in the comparison continue to decrease and the running time will obviously stay polynomial. In this section, we give a counterexample by considering the (1+1)-EA solving the LeadingOnes problem under one-bit noise, which suggests that the selection of sample size should be careful in practice.

The LeadingOnes problem as presented in Definition 3.1 aims to maximize the number of consecutive 1-bits counting from the left of a solution. Its optimal solution is $11\ldots1$ (denoted as $1^n$). As presented in Definition 3.2, one-bit noise flips a random bit of a solution before evaluation with probability $p$. When $p = 1$, it was known [18] that the expected running time of the (1+1)-EA is exponential, while the running time will reduce to polynomial by using sampling with $m = 4n^4 \log n/15$. We prove in Theorem 3.3 that the running time of the (1+1)-EA will return to exponential if $m \geq n^5$.

*Definition 3.1 (LeadingOnes).* The LeadingOnes Problem of size $n$ is to find a binary string $x^*$ that maximises

$$f(x) = \sum_{i=1}^{n} \prod_{j=1}^{i} x_j.$$

*Definition 3.2 (One-bit Noise).* Given a parameter $p \in [0, 1]$, let $f^{\mathrm{n}}(x)$ and $f(x)$ denote the noisy and true fitness of a solution $x \in \{0,1\}^n$, respectively, then

$$f^{\mathrm{n}}(x) = \begin{cases} f(x) & \text{with prob. } 1-p, \\ f(x') & \text{with prob. } p, \end{cases}$$

where $x'$ is generated by flipping a randomly chosen bit of $x$.

From Lemma 9 in [18], we can find the reason why sampling is effective only with a moderate sample size in this case. Let $f(x)$ and $f^{\mathrm{n}}(x)$ denote the true and noisy fitness of a solution, respectively. In most cases, if $f(x) > f(y)$, the expected gap between $f^{\mathrm{n}}(x)$ and $f^{\mathrm{n}}(y)$ is positive, which implies that a larger sample size is better since it will decrease $\mathrm{P}(\hat{f}(x) \leq \hat{f}(y))$. However, when $x = 1^n$ and $y$ is close to the optimum $1^n$, the expectation of $f^{\mathrm{n}}(1^n) - f^{\mathrm{n}}(y)$ can be negative, which implies that a larger sample size is worse since it will increase $\mathrm{P}(\hat{f}(1^n) \leq \hat{f}(y))$. Thus, neither a small sample size nor a large sample size is effective. The sample size of $m = 4n^4 \log n/15$ just makes a good tradeoff, which can lead to a not too large probability of $\hat{f}(1^n) \leq \hat{f}(y)$ and a sufficiently small probability of $\hat{f}(x) \leq \hat{f}(y)$ for two solutions $x$ and $y$ with $f(x) > f(y)$ and $\mathrm{E}(f^{\mathrm{n}}(x) - f^{\mathrm{n}}(y)) > 0$.

THEOREM 3.3. *For the (1+1)-EA solving LeadingOnes under one-bit noise with $p = 1$, the expected running time is exponential [18]; if using sampling with $m = 4n^4 \log n/15$, the expected running time is polynomial [18]; if using sampling with $m \geq n^5$, the expected running time is exponential.*

PROOF. We only need to prove the case $m \geq n^5$. Our main idea is to show that before reaching the optimal solution $1^n$, the algorithm will first find the solution $1^{n-1}0$ or $1^{n-2}01$ with a probability of at least $(1-\frac{1}{2^{n-2}})\cdot\frac{1}{n+1}$; while the probability of leaving $1^{n-1}0$ or $1^{n-2}01$ is exponentially small. Combining these two points, the theorem holds.

Let a Markov chain $\{\xi_t\}_{t=0}^{+\infty}$ model the analyzed evolutionary process. Let $\mathrm{LO}(x)$ denote the true number of leading 1-bits of a solution $x$. For any $t \geq 1$, let $C_t$ denote the event that at time $t$, the (1+1)-EA finds a solution with at least $n-2$ leading 1-bits for the first time, i.e., $\mathrm{LO}(\xi_t) \geq n-2$ and $\forall t' < t$ : $\mathrm{LO}(\xi_{t'}) < n-2$; let $A_t$ and $B_t$ denote the subsets of $C_t$, which require that $\xi_t \in \{1^{n-1}0, 1^{n-2}01\}$ and $\xi_t \in \{1^n, 1^{n-2}0^2\}$, respectively. Thus, before reaching the optimal solution $1^n$, the (1+1)-EA can find a solution in $\{1^{n-1}0, 1^{n-2}01\}$ with probability at least $\sum_{t=1}^{\infty}\mathrm{P}(A_t \mid C_t)\cdot\mathrm{P}(C_t)$.

We then show that $\mathrm{P}(A_t \mid C_t) \geq 1/(n+1)$. Assume that $\xi_{t-1} = x$, where $\mathrm{LO}(x) < n-2$. Let $\mathrm{P}_{\mathrm{mut}}(x,y)$ denote the probability that $x$ is mutated to $y$ by bit-wise mutation. Then,

$$\mathrm{P}(A_t \mid C_t) = \big(\mathrm{P}_{\mathrm{mut}}(x,1^{n-1}0)\cdot\mathrm{P}(\hat{f}(1^{n-1}0) \geq \hat{f}(x)) + \mathrm{P}_{\mathrm{mut}}(x,1^{n-2}01)\cdot\mathrm{P}(\hat{f}(1^{n-2}01) \geq \hat{f}(x))\big)/\mathrm{P}(C_t). \quad (1)$$

For $\mathrm{P}(\hat{f}(1^{n-1}0) \geq \hat{f}(x))$ and $\mathrm{P}(\hat{f}(1^{n-2}01) \geq \hat{f}(x))$, we apply Hoeffding's inequality to get a lower bound $1-e^{-n/2}$. By the definition of one-bit noise, we get, for any $0 \leq k \leq n-1$,

$$\mathrm{E}(f^{\mathrm{n}}(1^k01^{n-k-1})) = \sum_{j=1}^{k}\frac{1}{n}\cdot(j-1) + \frac{1}{n}\cdot n + \frac{n-k-1}{n}\cdot k.$$

Then, we have, for any $1 \leq k \leq n-1$,

$$\mathrm{E}(f^{\mathrm{n}}(1^k01^{n-k-1})) - \mathrm{E}(f^{\mathrm{n}}(1^{k-1}01^{n-k})) = (n-k-1)/n. \quad (2)$$

Thus, for any $k \leq n-3$, $\mathrm{E}(f^{\mathrm{n}}(1^{n-1}0)) - \mathrm{E}(f^{\mathrm{n}}(1^k01^{n-k-1})) \geq \mathrm{E}(f^{\mathrm{n}}(1^{n-2}01)) - \mathrm{E}(f^{\mathrm{n}}(1^{n-3}01^2)) = 1/n$. Since $\mathrm{LO}(x) \leq n-3$ and $\mathrm{E}(f^{\mathrm{n}}(x)) \leq \mathrm{E}(f^{\mathrm{n}}(1^{\mathrm{LO}(x)}01^{n-\mathrm{LO}(x)-1}))$, we have

$$\mathrm{E}(f^{\mathrm{n}}(1^{n-1}0)) - \mathrm{E}(f^{\mathrm{n}}(x)) \geq 1/n.$$

Let $r = \mathrm{E}(\hat{f}(x) - \hat{f}(1^{n-1}0))$. Since the $\hat{f}$ value by sampling is the average of $m$ independent evaluations, $r = \mathrm{E}(f^{\mathrm{n}}(x)) - \mathrm{E}(f^{\mathrm{n}}(1^{n-1}0)) \leq -1/n$. Then, we have

$$\mathrm{P}(\hat{f}(x) \geq \hat{f}(1^{n-1}0)) = \mathrm{P}(\hat{f}(x) - \hat{f}(1^{n-1}0) - r \geq -r) \leq \exp\left(-2m^2r^2/(m(2n)^2)\right) \leq e^{-n/2}, \quad (3)$$

where the first inequality is by Hoeffding's inequality and $-n \leq f^{\mathrm{n}}(x) - f^{\mathrm{n}}(1^{n-1}0) \leq n$, and the last is by $r \leq -1/n$ and $m \geq n^5$. It is easy to see from Eq. (2) that $\mathrm{E}(f^{\mathrm{n}}(1^{n-2}01)) = \mathrm{E}(f^{\mathrm{n}}(1^{n-1}0))$. Thus, we can similarly get

$$\mathrm{P}(\hat{f}(x) \geq \hat{f}(1^{n-2}01)) \leq e^{-n/2}. \quad (4)$$

By applying Eqs. (3) and (4) to Eq. (1), we get

$$\mathrm{P}(A_t|C_t) \geq (1-e^{-n/2})\frac{\mathrm{P}_{\mathrm{mut}}(x,1^{n-1}0)+\mathrm{P}_{\mathrm{mut}}(x,1^{n-2}01)}{\mathrm{P}(C_t)}.$$

Since $\mathrm{P}(B_t \mid C_t) \leq (\mathrm{P}_{\mathrm{mut}}(x,1^n) + \mathrm{P}_{\mathrm{mut}}(x,1^{n-2}0^2))/\mathrm{P}(C_t)$,

$$\frac{\mathrm{P}(A_t|C_t)}{\mathrm{P}(B_t|C_t)} \geq (1-e^{-n/2})\frac{\mathrm{P}_{\mathrm{mut}}(x,1^{n-1}0)+\mathrm{P}_{\mathrm{mut}}(x,1^{n-2}01)}{\mathrm{P}_{\mathrm{mut}}(x,1^n)+\mathrm{P}_{\mathrm{mut}}(x,1^{n-2}0^2)}.$$

If $x_{n-1} = x_n = 0$ or $x_{n-1} = x_n = 1$,

$$\frac{\mathrm{P}(A_t \mid C_t)}{\mathrm{P}(B_t \mid C_t)} \geq (1-e^{-n/2})\frac{\frac{1}{n}(1-\frac{1}{n})+\frac{1}{n}(1-\frac{1}{n})}{\frac{1}{n^2}+(1-\frac{1}{n})^2} \geq \frac{1}{n}.$$

If $x_{n-1}+x_n = 1$, we can similarly derive that $\frac{\mathrm{P}(A_t|C_t)}{\mathrm{P}(B_t|C_t)} \geq \frac{1}{n}$. Since $\mathrm{P}(A_t \mid C_t) + \mathrm{P}(B_t \mid C_t) = 1$, our claim that $\mathrm{P}(A_t \mid C_t) \geq 1/(n+1)$ holds.

Thus, the probability that the (1+1)-EA first finds a solution in $\{1^{n-1}0, 1^{n-2}01\}$ before reaching $1^n$ is at least

$$\sum_{t=1}^{\infty}\mathrm{P}(A_t \mid C_t)\cdot\mathrm{P}(C_t) \geq (1/(n+1))\cdot\sum_{t=1}^{\infty}\mathrm{P}(C_t) = (1/(n+1))\mathrm{P}(\mathrm{LO}(\xi_0) < n-2) = (1/(n+1))\left(1-1/2^{n-2}\right),$$

where the first equality is because the union of the events $C_t$ with $t \geq 1$ implies that the time of finding a solution with at least $n-2$ leading 1-bits is at least 1, which is equivalent to that the initial solution $\xi_0$ has less than $n-2$ leading 1-bits ; and the last equality is due to the uniform initial distribution.

We then show that after finding $1^{n-1}0$ or $1^{n-2}01$, the probability of the (1+1)-EA leaving this state in each iteration is exponentially small. From Eqs. (3) and (4), we know that for any $x$ with $\mathrm{LO}(x) < n-2$ and $y \in \{1^{n-1}0, 1^{n-2}01\}$, $\mathrm{P}(\hat{f}(x) \geq \hat{f}(y)) \leq e^{-n/2}$. For $x \in \{1^{n-2}0^2, 1^n\}$ and $y \in \{1^{n-1}0, 1^{n-2}01\}$, it is easy to verify that $\mathrm{E}(f^{\mathrm{n}}(y) - f^{\mathrm{n}}(x)) = \sum_{j=1}^{n-1}\frac{1}{n}(j-1) + \frac{1}{n}\cdot n - \sum_{j=1}^{n}\frac{1}{n}(j-1) = \frac{1}{n}$. Using the same analysis as Eq. (3), we get, for $x \in \{1^{n-2}0^2, 1^n\}$ and $y \in \{1^{n-1}0, 1^{n-2}01\}$, $\mathrm{P}(\hat{f}(x) \geq \hat{f}(y)) \leq e^{-n/2}$. Combining the above two cases, we get, for $x \notin \{1^{n-1}0, 1^{n-2}01\}$ and $y \in \{1^{n-1}0, 1^{n-2}01\}$, $\mathrm{P}(\hat{f}(x) \geq \hat{f}(y)) \leq e^{-n/2}$. Thus, our claim that the probability of leaving $\{1^{n-1}0, 1^{n-2}01\}$ is exponentially small holds. □

## 4 PARENT POPULATIONS CAN WORK ON SOME TASKS WHERE SAMPLING FAILS

Previous studies [12, 18, 19] have shown that both using parent populations and sampling can bring robustness to noise. For example, for the OneMax problem under one-bit noise with $p = \omega(\log n/n)$, the (1+1)-EA needs exponential time to find the optimum [8], while both using a parent population size $\mu \geq 12\log(15n)/p$ [12] and a sample size $m = 4n^3$ [18] can reduce the running time to polynomial. Then, a natural question is that whether there exist cases where only one of these two strategies is effective. This question has been partially solved. For the OneMax problem under additive Gaussian noise with large variances, it was shown that the ($\mu$+1)-EA with $\mu = \omega(1)$ needs super-polynomial time to find the optimum [11], while the (1+1)-EA using sampling can find the optimum in polynomial time [19]. In this section, we solve the other part of this question. That is, we prove that using parent populations can be better than using sampling.

Particularly, we compare the (1+1)-EA using sampling with the ($\mu$+1)-EA for solving OneMax under symmetric noise. The OneMax problem as presented in Definition 4.1 is to maximize the number of 1-bits, and the optimal solution is $1^n$. As presented in Definition 4.2, symmetric noise returns a false fitness $2n - f(x)$ with probability $1/2$. It is easy to see that under this noise model, the distribution of $f^{\mathrm{n}}(x)$ for any $x$ is

symmetric about $n$. We prove in Theorems 4.3 and 4.4 that the expected running time of the (1+1)-EA using sampling with any sample size $m$ is exponential, while the ($\mu$+1)-EA with $\mu = 3\log n$ can find the optimum in $O(n\log^3 n)$ time.

*Definition 4.1 (OneMax).* The OneMax Problem of size $n$ is to find a binary string $x^*$ that maximises

$$f(x) = \sum_{i=1}^{n} x_i.$$

*Definition 4.2 (Symmetric Noise).* Let $f^{\mathrm{n}}(x)$ and $f(x)$ denote the noisy and true fitness of a solution $x \in \{0,1\}^n$, respectively, then

$$f^{\mathrm{n}}(x) = \begin{cases} f(x) & \text{with prob. } 1/2, \\ 2n - f(x) & \text{with prob. } 1/2. \end{cases}$$

From the following analyses, we can find the reason why using sampling fails while using parent populations can work in this case. Under symmetric noise, the distribution of $f^{\mathrm{n}}(x)$ for any $x$ is symmetric about $n$. Thus, for any two solutions $x$ and $y$, the distribution of $f^{\mathrm{n}}(x) - f^{\mathrm{n}}(y)$ is symmetric about 0. By using sampling, the distribution of $\hat{f}(x) - \hat{f}(y)$ is still symmetric about 0, which implies that the offspring solution will always be accepted with probability at least $1/2$ in each iteration of the (1+1)-EA. Such a behavior is analogous to random walk, and thus the optimization is inefficient. The reason for the effectiveness of using parent populations is that the true best solution will be discarded only if it appears worse than all the other solutions in the population, the probability of which can be very small by using a logarithmic population size. Note that this finding is consistent with that in [12].

THEOREM 4.3. *For the (1+1)-EA solving OneMax under symmetric noise, if using sampling, the expected running time is exponential.*

PROOF. We use Lemma 2.3 to prove it. Let $X_t = |x|_0$ be the number of 0-bits of the solution $x$ after $t$ iterations of the (1+1)-EA. We consider the interval $[0, n/10]$, i.e., the parameters $a = 0$ (i.e., the optimum) and $b = n/10$ in Lemma 2.3.

Then, we analyze the drift $\mathrm{E}(X_t - X_{t+1} \mid X_t = i)$ for $1 \le i < n/10$. We divide the drift into two parts: positive $\mathrm{E}^+$ and negative $\mathrm{E}^-$. That is,

$$\mathrm{E}(X_t - X_{t+1} \mid X_t = i) = \mathrm{E}^+ - \mathrm{E}^-, \quad \text{where}$$
$$\mathrm{E}^+ = \textstyle\sum_{x':|x'|_0<i} \mathrm{P}_{\mathrm{mut}}(x,x') \cdot \mathrm{P}(\hat{f}(x') \ge \hat{f}(x)) \cdot (i - |x'|_0),$$
$$\mathrm{E}^- = \textstyle\sum_{x':|x'|_0>i} \mathrm{P}_{\mathrm{mut}}(x,x') \cdot \mathrm{P}(\hat{f}(x') \ge \hat{f}(x)) \cdot (|x'|_0 - i).$$

For the positive drift, we use a trivial upper bound 1 for $\mathrm{P}(\hat{f}(x') \ge \hat{f}(x))$. Then, we have

$$\mathrm{E}^+ \le \textstyle\sum_{x':|x'|_0<i} \mathrm{P}_{\mathrm{mut}}(x,x')(i - |x'|_0) \le i/n,$$

where the second inequality is directly from the proof of Theorem 5 in [18].

For the negative drift, we need to consider that the number of 0-bits is increased. We analyze the $n-i$ cases where only one 1-bit is flipped (i.e., $|x'|_0 = i+1$), which happens with probability $\frac{1}{n}(1-\frac{1}{n})^{n-1} \ge \frac{1}{en}$. Let $Y = f^{\mathrm{n}}(x) - f^{\mathrm{n}}(x')$. By the definition of symmetric noise, the value of $Y$ can be $-2i-1$, $-1$, $1$ and $2i+1$, each with probability $1/4$. It is easy to see that the distribution of $Y$ is symmetric about 0, i.e., $Y$ has the same distribution as $-Y$. Since $\hat{f}(x) - \hat{f}(x')$ is the average of $m$ independent random variables, which have the same distribution as $Y$, the distribution of $\hat{f}(x) - \hat{f}(x')$ is also symmetric about 0, and thus $\mathrm{P}(\hat{f}(x') \ge \hat{f}(x)) \ge 1/2$. Then,

$$\mathrm{E}^- \ge (n-i)/(en) \cdot (1/2) \cdot (i+1-i) = (n-i)/(2en).$$

By subtracting $\mathrm{E}^-$ from $\mathrm{E}^+$, we get

$$\mathrm{E}(X_t - X_{t+1} \mid X_t = i) \le i/n - (n-i)/(2en) \le -0.05,$$

where the last inequality is by $i < n/10$. That is, condition (1) of Lemma 2.3 holds with $\epsilon = 0.05$.

To make $|X_{t+1} - X_t| \ge j$, it is necessary to flip at least $j$ bits of $x$. Thus, we have

$$\mathrm{P}(|X_{t+1} - X_t| \ge j \mid X_t \ge 1) \le \binom{n}{j} \frac{1}{n^j} \le \frac{1}{j!} \le 2 \cdot \frac{1}{2^j},$$

which implies that condition (2) of Lemma 2.3 holds with $\delta = 1$ and $r(l) = 2$. Note that $l = b - a = n/10$. Thus, by Lemma 2.3, the expected running time is exponential. □

THEOREM 4.4. *For the ($\mu$+1)-EA solving OneMax under symmetric noise, if $\mu = 3\log n$, the expected running time is $O(n\log^3 n)$.*

PROOF. We use Lemma 2.2 to prove it. Note that the state of the corresponding Markov chain is currently a population, i.e., a set of $\mu$ solutions. We first construct a distance function $V$: for any population $P$, $V(P) = \min_{x\in P} |x|_0$, i.e., the minimum number of 0-bits of the solution in $P$. It is easy to see that $V(P) = 0$ iff $P \in \mathcal{X}^*$, i.e., $P$ contains the optimum $1^n$.

Then, we investigate $\mathrm{E}(V(\xi_t) - V(\xi_{t+1}) \mid \xi_t = P)$ for any $P$ with $V(P) > 0$ (i.e., $P \notin \mathcal{X}^*$). Assume that currently $V(P) = i$, where $1 \le i \le n$. We also divide the drift into two parts:

$$\mathrm{E}(V(\xi_t) - V(\xi_{t+1}) \mid \xi_t = P) = \mathrm{E}^+ - \mathrm{E}^-, \quad \text{where}$$
$$\mathrm{E}^+ = \textstyle\sum_{P':V(P')<i} \mathrm{P}(\xi_{t+1} = P' \mid \xi_t = P) \cdot (i - V(P')),$$
$$\mathrm{E}^- = \textstyle\sum_{P':V(P')>i} \mathrm{P}(\xi_{t+1} = P' \mid \xi_t = P) \cdot (V(P') - i).$$

For $\mathrm{E}^+$, we need to consider that the best solution in $P$ is improved. Let $x^* \in \arg\min_{x\in P} |x|_0$, then $|x^*|_0 = i$. In one iteration of the ($\mu$+1)-EA, a solution $x'$ with $|x'|_0 = i-1$ can be generated by selecting $x^*$ and flipping only one 0-bit in mutation, whose probability is $\frac{1}{\mu} \cdot \frac{i}{n}(1-\frac{1}{n})^{n-1} \ge \frac{i}{e\mu n}$. If $x'$ is not added into $P$, it must hold that $f^{\mathrm{n}}(x') < f^{\mathrm{n}}(x)$ for any $x \in P$, which happens with probability $1/2^\mu$ since $f^{\mathrm{n}}(x') < f^{\mathrm{n}}(x)$ iff $f^{\mathrm{n}}(x) = 2n - f(x)$. Thus, the probability that $x'$ is added into $P$ (which implies that $V(P') = i-1$) is $1 - 1/2^\mu$. We then get

$$\mathrm{E}^+ \ge \frac{i}{e\mu n} \cdot \left(1 - \frac{1}{2^\mu}\right) \cdot (i - (i-1)) = \frac{i}{e\mu n}\left(1 - \frac{1}{2^\mu}\right).$$

For $\mathrm{E}^-$, if there are at least two solutions $x, y$ in $P$ such that $|x|_0 = |y|_0 = i$, it obviously holds that $\mathrm{E}^- = 0$. Otherwise, $V(P') > V(P) = i$ implies that for the unique best solution $x^*$ in $P$ and any $x \in P \setminus \{x^*\}$, $f^{\mathrm{n}}(x^*) \le f^{\mathrm{n}}(x)$, which happens with probability $1/2^{\mu-1}$ since $f^{\mathrm{n}}(x^*) \le f^{\mathrm{n}}(x)$ iff $f^{\mathrm{n}}(x) = 2n - f(x)$. Thus, $\mathrm{P}(V(P') > i) \le 1/2^{\mu-1}$. Furthermore, $V(P')$

can increase by at most $n-i$. Thus, $\mathrm{E}^- \le (n-i)/2^{\mu-1}$. By subtracting $\mathrm{E}^-$ from $\mathrm{E}^+$, we get

$$\mathrm{E}(V(\xi_t)-V(\xi_{t+1}) \mid \xi_t) \ge \frac{i}{e\mu n} - \frac{i}{e\mu n 2^{\mu}} - \frac{n-i}{2^{\mu-1}} \ge \frac{i}{10n\log n},$$

where the last inequality holds with sufficiently large $n$. Note that $\mu = 3\log n$. Thus, by Lemma 2.2,

$$\mathrm{E}(\tau \mid \xi_0) \le 10n\log n(1+\log n) = O(n\log^2 n),$$

which implies that the expected running time is $O(n\log^3 n)$, since the algorithm needs to evaluate the offspring solution and reevaluate the $\mu$ parent solutions in each iteration. □

## 5 ADAPTIVE SAMPLING CAN WORK ON SOME TASKS WHERE BOTH SAMPLING AND PARENT POPULATIONS FAIL

In this section, we first theoretically investigate whether there exist cases where using neither parent populations nor sampling is effective. We give a positive answer by considering OneMax under segmented noise. Then, we prove that in such a situation, using adaptive sampling can be effective, which provides some theoretical justification for the good empirical performance of adaptive sampling in practice [22, 26].

As presented in Definition 5.1, the OneMax problem is divided into four segments. In one segment, the fitness is evaluated correctly, while in the other three segments, the fitness is disturbed by different noises. We prove in Theorem 5.2 that the expected running time of the (1+1)-EA using sampling with any sample size $m$ is exponential. From the proof, we can find the reason for the ineffectiveness of sampling. For two solutions $x$ and $x'$ with $|x'|_0 = |x|_0 + 1$ (i.e., $f(x) = f(x') + 1$), the expected gaps between $f^{\mathrm{n}}(x)$ and $f^{\mathrm{n}}(x')$ are positive and negative, respectively, in the segments of $\frac{n}{100} < |x|_0 \le \frac{n}{50}$ and $\frac{n}{200} < |x|_0 \le \frac{n}{100}$. Thus, in the former segment, a larger sample size is better since it will decrease $\mathrm{P}(\hat{f}(x) \le \hat{f}(x'))$, while in the latter segment, a larger sample size is worse since it will increase $\mathrm{P}(\hat{f}(x) \le \hat{f}(x'))$. Furthermore, there is no moderate sample size which can make a good tradeoff. Thus, sampling fails in this case.

*Definition 5.1 (OneMax under Segmented Noise).* For any $x \in \{0,1\}^n$, the noisy fitness value $f^{\mathrm{n}}(x)$ is calculated as:

(1) if $|x|_0 > \frac{n}{50}$, $f^{\mathrm{n}}(x) = n - |x|_0$;

(2) if $\frac{n}{100} < |x|_0 \le \frac{n}{50}$,

$$f^{\mathrm{n}}(x) = \begin{cases} n - |x|_0 & \text{with prob. } \frac{1}{2} + \frac{1}{n}, \\ 3n + |x|_0 & \text{with prob. } \frac{1}{2} - \frac{1}{n}; \end{cases}$$

(3) if $\frac{n}{200} < |x|_0 \le \frac{n}{100}$,

$$f^{\mathrm{n}}(x) = \begin{cases} 4n(n - |x|_0) & \text{with prob. } 1 - \frac{1}{n}, \\ (2n + |x|_0)^3 & \text{with prob. } \frac{1}{n}; \end{cases}$$

(4) if $|x|_0 \le \frac{n}{200}$,

$$f^{\mathrm{n}}(x) = \begin{cases} n^4(n - |x|_0) & \text{with prob. } \frac{1}{5}, \\ -n^4 - \delta & \text{with prob. } \frac{4}{5}, \end{cases}$$

where $\delta$ is randomly drawn from a continuous uniform distribution $\mathcal{U}[0,1]$, and $n/200 \in \mathbb{N}^+$.

THEOREM 5.2. *For the (1+1)-EA solving OneMax under segmented noise, if using sampling, the expected running time is exponential.*

PROOF. We divide the proof into two parts according to the range of $m$. Let $X_t = |x|_0$ denote the number of 0-bits of the solution $x$ after $t$ iterations of the (1+1)-EA. When $m \le \frac{n^4}{400}$, we use Lemma 2.3 to prove that starting from $X_0 \ge \frac{n}{50}$, the expected number of iterations until $X_t \le \frac{n}{100}$ is exponential. When $m > \frac{n^4}{400}$, we use Lemma 2.3 to prove that starting from $X_0 \ge \frac{n}{100}$, the expected number of iterations until $X_t \le \frac{n}{200}$ is exponential. Due to the uniform initial distribution, both $X_0 \ge \frac{n}{50}$ and $X_0 \ge \frac{n}{100}$ hold with a high probability. Thus, for any $m$, the expected running time until finding the optimum is exponential. For the proof of each part, condition (2) of Lemma 2.3 trivially holds, and we only need to show that $\mathrm{E}(X_t - X_{t+1} \mid X_t)$ is upper bounded by a negative constant.

[Part I: $m \le \frac{n^4}{400}$] We consider the interval $[\frac{n}{100}, \frac{n}{50}]$. As in the proof of Theorem 4.3, we compute the drift $\mathrm{E}(X_t - X_{t+1} \mid X_t = i)$ (where $\frac{n}{100} < i < \frac{n}{50}$) by $\mathrm{E}^+ - \mathrm{E}^-$. For $\mathrm{E}^-$, we consider the $n-i$ cases where only one 1-bit of $x$ is flipped in mutation. That is, $|x'|_0 = i+1$. We then show that the offspring solution $x'$ is accepted with probability at least $0.07$ (i.e., $\mathrm{P}(\hat{f}(x') \ge \hat{f}(x)) \ge 0.07$) by considering two subcases for $m$.

(1) $m \ge 4$. For any $\frac{n}{100} < k \le \frac{n}{50}$, let $x^k$ denote a solution with $k$ number of 0-bits. According to Definition 5.1, we have

$$\begin{aligned} &\mathrm{E}(f^{\mathrm{n}}(x^k)) = (1/2 + 1/n)(n-k) \\ &\quad + (1/2 - 1/n)(3n+k) = 2n - 2 - 2k/n; \\ &\mathrm{Var}(f^{\mathrm{n}}(x^k)) = (1/2 + 1/n)(n-k)^2 \\ &\quad + (1/2 - 1/n)(3n+k)^2 - (2n - 2 - 2k/n)^2 \\ &\ge (1/2 - 1/n) \cdot (10n^2 + 2k^2 + 4kn) - 4n^2 \ge n^2. \end{aligned} \tag{5}$$

Let $Y = f^{\mathrm{n}}(x) - f^{\mathrm{n}}(x')$. Note that $|x|_0 = i \in (\frac{n}{100}, \frac{n}{50})$ and $|x'|_0 = i+1$. Then, we can easily get that $\mu := \mathrm{E}(Y) = 2/n$ and $\sigma^2 := \mathrm{Var}(Y) \ge 2n^2$. Let $Z = Y - \mu$. Then, we have $\mathrm{E}(Z) = 0$, $\mathrm{Var}(Z) = \sigma^2 \ge 2n^2$ and

$$\begin{aligned} \rho &:= \mathrm{E}(|Z|^3) \le 2\left(1/4 - 1/n^2\right) \cdot (2n + 2i + 1 + 2/n)^3 \\ &\quad + \left((1/2 - 1/n)^2 + (1/2 + 1/n)^2\right) \cdot (1 + 2/n)^3 \le 9n^3/2, \end{aligned}$$

where the last inequality holds with sufficiently large $n$. Note that $\hat{f}(x) - \hat{f}(x') - \mu$ is the average of $m$ independent random variables, which have the same distribution as $Z$. By Berry-Esseen inequality [23], we have

$$\mathrm{P}\left((\hat{f}(x) - \hat{f}(x') - \mu)\sqrt{m}/\sigma \le x\right) - \Phi(x) \ge -\rho/(2\sigma^3\sqrt{m}),$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. Thus, we have

$$\begin{aligned} &\mathrm{P}(\hat{f}(x) - \hat{f}(x') \le 0)) = \mathrm{P}(\hat{f}(x) - \hat{f}(x') - \mu \le -\mu) \\ &= \mathrm{P}\left((\hat{f}(x) - \hat{f}(x') - \mu)\sqrt{m}/\sigma \le -\mu\sqrt{m}/\sigma\right) \\ &\ge \Phi\left(-\mu\sqrt{m}/\sigma\right) - \rho/(2\sigma^3\sqrt{m}) \ge 0.07, \end{aligned}$$

where the second inequality is by $\mu = \frac{2}{n}$, $4 \le m \le \frac{n^4}{400}$, $\sigma \ge \sqrt{2}n$ and $\rho \le \frac{9}{2}n^3$.

(2) $m \leq 3$. It holds that $\mathrm{P}(\hat{f}(x') \geq \hat{f}(x)) \geq (\frac{1}{2} - \frac{1}{n})^3 \geq 0.1$, since it is sufficient that $f^{\mathrm{n}}(x')$ is always evaluated to $3n+i+1$ in $m$ independent evaluations.

Combining the above two cases, our claim that $\mathrm{P}(\hat{f}(x') \geq \hat{f}(x)) \geq 0.07$ holds. Note that $i < n/50$. Thus,

$$\mathrm{E}^- \geq ((n-i)/n)\,(1-1/n)^{n-1} \cdot 0.07 \geq 1.2/50.$$

For the positive drift, we can similarly get $\mathrm{E}^+ \leq \frac{i}{n} \leq \frac{1}{50}$ as in the proof of Theorem 4.3, since we optimistically assume that $x'$ is always accepted. Thus, the drift satisfies that

$$\mathrm{E}(X_t - X_{t+1} \mid X_t = i) = \mathrm{E}^+ - \mathrm{E}^- \leq -0.2/50.$$

[Part II: $m > \frac{n^4}{400}$] We consider the interval $[\frac{n}{200}, \frac{n}{100}]$, and compute the drift $\mathrm{E}(X_t - X_{t+1} \mid X_t = i)$ (where $\frac{n}{200} < i < \frac{n}{100}$) by $\mathrm{E}^+ - \mathrm{E}^-$. For the negative drift, we show that the probability of accepting the offspring solution $x'$ with $|x'|_0 = i+1$ is at least 0.9. For any $\frac{n}{200} < k < \frac{n}{100}$,

$$\mathrm{E}(f^{\mathrm{n}}(x^k) - f^{\mathrm{n}}(x^{k+1})) = (1-1/n) \cdot 4n - (1/n) \cdot (3(2n+k)^2 + 3(2n+k) + 1) \leq -8n;$$

and for any $\frac{n}{200} < k \leq \frac{n}{100}$,

$$\mathrm{Var}(f^{\mathrm{n}}(x^k)) = (2n+k)^6/n + (1-1/n)\,(4n(n-k))^2 - (\mathrm{E}(f^{\mathrm{n}}(x^k)))^2 \leq (1/n) \cdot 66n^6 + 16n^4 \leq 82n^5.$$

Then, $\mu := \mathrm{E}(\hat{f}(x) - \hat{f}(x')) \leq -8n$ and $\sigma^2 := \mathrm{Var}(\hat{f}(x) - \hat{f}(x')) \leq \frac{2}{m} \cdot 82n^5$. By Chebyshev's inequality and $m > \frac{n^4}{400}$, we have

$$\mathrm{P}(\hat{f}(x) \geq \hat{f}(x')) \leq \mathrm{P}(|\hat{f}(x) - \hat{f}(x') - \mu| \geq -\mu) \leq \sigma^2/\mu^2 \leq 0.1,$$

where the last inequality holds with sufficiently large $n$. Thus, $\mathrm{E}^- \geq \frac{n-i}{n}\left(1-\frac{1}{n}\right)^{n-1} \cdot 0.9 \geq \frac{99}{100e} \cdot 0.9 \geq 0.29$. For the positive drift, we still have $\mathrm{E}^+ \leq \frac{i}{n} \leq 0.01$. Thus, the drift satisfies that

$$\mathrm{E}(X_t - X_{t+1} \mid X_t = i) = \mathrm{E}^+ - \mathrm{E}^- \leq -0.28. \qquad \square$$

To prove the ineffectiveness of using parent populations, we derive a sufficient condition for the exponential running time of the ($\mu$+1)-EA required to solve OneMax under noise, which is inspired from Theorem 4 in [11]. We generalize their result from additive noise to arbitrary noise. The proof is provided in the appendix due to space limitations. As presented in Lemma 5.3, the condition intuitively means that when the solution is close to the optimum, the probability of discarding it from the population decreases linearly w.r.t. the population size $\mu$, which is, however, not small enough to make an efficient optimization. Note that for the case where using parent populations works in the last section, the probability of discarding the best solution from the population decreases exponentially w.r.t. $\mu$. By verifying this condition, we prove in Theorem 5.4 that the ($\mu$+1)-EA with $\mu \in poly(n)$ needs exponential time for solving OneMax under segmented noise. Let $poly(n)$ indicate any polynomial of $n$.

LEMMA 5.3. *For the ($\mu$+1)-EA (where $\mu \in poly(n)$) solving OneMax under noise, if for any $y$ with $|y|_1 > \frac{599n}{600}$ and any set of $\mu$ solutions $Q = \{x^1, x^2, \ldots, x^\mu\}$,*

$$\mathrm{P}(f^{\mathrm{n}}(y) < \min_{x^i \in Q} f^{\mathrm{n}}(x^i)) \geq 3/(5(\mu+1)), \qquad (6)$$

*then the expected running time is exponential.*

THEOREM 5.4. *For the ($\mu$+1)-EA (where $\mu \in poly(n)$) solving OneMax under segmented noise, the expected running time is exponential.*

PROOF. We use Lemma 5.3 to prove it. For any solution $y$ with $|y|_0 \leq n/200$ and $Q = \{x^1, \ldots, x^\mu\}$, let $A$ denote the event that $f^{\mathrm{n}}(y) < \min_{x^i \in Q} f^{\mathrm{n}}(x^i)$. We will show that $\mathrm{P}(A) \geq \frac{4}{5(\mu+1)}$, which implies that the condition Eq. (6) holds since $|y|_0 \leq n/200$ covers the required range of $|y|_1 > 599n/600$.

Let $B_l$ $(0 \leq l \leq \mu)$ denote the event that $l$ solutions in $Q$ are evaluated to have negative noisy fitness values. Note that for any $x$, $f^{\mathrm{n}}(x) < 0$ implies that $|x|_0 \leq n/200$, and $f^{\mathrm{n}}(x) = -n^4 - \delta$ where $\delta \sim \mathcal{U}[0,1]$. For any $0 \leq l \leq \mu$,

$$\mathrm{P}(A \mid B_l) \geq \mathrm{P}(f^{\mathrm{n}}(y) < 0 \mid B_l) \cdot \mathrm{P}(A \mid f^{\mathrm{n}}(y) < 0, B_l).$$

Under the conditions $f^{\mathrm{n}}(y) < 0$ and $B_l$, the noisy fitness values of $y$ and the corresponding $l$ solutions in $Q$ satisfy the same continuous distribution $-n^4 - \delta$ where $\delta \sim \mathcal{U}[0,1]$, thus

$$\mathrm{P}(A \mid f^{\mathrm{n}}(y) < 0, B_l) \geq 1/(l+1) \geq 1/(\mu+1).$$

Then, we get $\mathrm{P}(A \mid B_l) \geq \frac{4}{5} \cdot \frac{1}{\mu+1}$ and $\mathrm{P}(A) = \sum_{l=0}^{\mu} \mathrm{P}(A \mid B_l) \cdot \mathrm{P}(B_l) \geq \frac{4}{5(\mu+1)}$. By Lemma 5.3, the theorem holds. $\square$

We prove in Theorem 5.6 that the (1+1)-EA using adaptive sampling can solve OneMax under segmented noise in polynomial time. The employed adaptive sampling strategy is:

**[Adaptive Sampling]** For comparing two solutions $x$ and $y$, it first evaluates their noisy fitness once independently. If $3n \leq |f^{\mathrm{n}}(x) - f^{\mathrm{n}}(y)| < n^4$, this comparison result is directly used; otherwise, each solution will be evaluated $n^5$ times independently and the comparison will be based on the average value of these $n^5$ fitness evaluations.

Intuitively, when the noisy fitness gap of two solutions is too small or too large, we need to increase the sample size to make a more confident comparison.

To prove Theorem 5.6, we apply the upper bound on the number of iterations of the (1+1)-EA solving noisy OneMax in [12]. Let $x^j$ denote any solution with $j$ 0-bits. Lemma 5.5 intuitively means that if the probability of recognizing the true better solution in the comparison is large, the running time can be upper bounded. From the proof, we can find why adaptive sampling is effective in this case. In the 2nd segment (or the 4th segment) of the problem, $\mathrm{E}(f^{\mathrm{n}}(x) - f^{\mathrm{n}}(y))$ is positive for two solutions $x$ and $y$ with $f(x) > f(y)$, while in the 3rd segment, it is negative. Thus, a large sample size is better in the 2nd and 4th segments, while a small one is better in the 3rd segment. According to the range of the noisy fitness gap of two solutions in each segment, the adaptive sampling strategy happens to allocate $n^5$ evaluations for comparing two solutions in the 2nd segment (or the 4th segment), while allocate only one evaluation in the 3rd segment; thus it works.

LEMMA 5.5. *[12] Suppose there is a positive constant $c \leq 1/15$ and some $2 < l \leq n/2$ such that*

$$\forall 0 < i \leq j : \mathrm{P}(\hat{f}(x^j) < \hat{f}(x^{i-1})) \geq 1 - l/n;$$
$$\forall l < i \leq j : \mathrm{P}(\hat{f}(x^j) < \hat{f}(x^{i-1})) \geq 1 - ci/n,$$

*then the (1+1)-EA optimizes noisy OneMax in expectation in $O(n\log n)+n2^{O(l)}$ iterations.*

THEOREM 5.6. *For the (1+1)-EA solving OneMax under segmented noise, if using adaptive sampling, the expected running time is polynomial.*

PROOF. We use Lemma 5.5 to prove it. By considering four cases for $i$, we analyze $\mathrm{P}(\hat{f}(x^j)\geq\hat{f}(x^{i-1}))$, where $0<i\leq j$.
(1) $i>\frac{n}{50}$. It holds that $\forall j\geq i$, $\mathrm{P}(\hat{f}(x^j)\geq\hat{f}(x^{i-1}))=0$, since $f^{\mathrm{n}}(x^j)$ is evaluated exactly and $f^{\mathrm{n}}(x^{i-1})$ must be larger.
(2) $\frac{n}{100}+1<i\leq\frac{n}{50}$. If $j>\frac{n}{50}$, we easily verify that $\mathrm{P}(\hat{f}(x^j)\geq\hat{f}(x^{i-1}))=0$. If $j\leq\frac{n}{50}$, $|f^{\mathrm{n}}(x^j)-f^{\mathrm{n}}(x^{i-1})|<3n$, thus both $x^j$ and $x^{i-1}$ will be evaluated $n^5$ times according to the adaptive sampling strategy. Let $Y=f^{\mathrm{n}}(x^{i-1})-f^{\mathrm{n}}(x^j)$. Based on Eq. (5), we easily get $\mu:=\mathrm{E}(Y)\geq\frac{2}{n}$, and for any $\frac{n}{100}<k\leq\frac{n}{50}$, $\mathrm{Var}(f^{\mathrm{n}}(x^k))\leq(\frac{1}{2}+\frac{1}{n})\cdot n^2+(\frac{1}{2}-\frac{1}{n})\cdot 10n^2\leq 6n^2$, thus $\sigma^2:=\mathrm{Var}(Y)\leq 12n^2$. By Chebyshev's inequality and $m=n^5$, we get $\mathrm{P}(\hat{f}(x^j)\geq\hat{f}(x^{i-1}))\leq\mathrm{P}(|\hat{f}(x^{i-1})-\hat{f}(x^j)-\mu|\geq\mu)\leq\frac{\sigma^2}{m\mu^2}\leq\frac{3}{n}$.
(3) $\frac{n}{200}+1<i\leq\frac{n}{100}+1$. If $j\geq\frac{n}{100}+1$, it holds that $\mathrm{P}(\hat{f}(x^j)\geq\hat{f}(x^{i-1}))=0$, since the noisy fitness in the 3rd segment of Definition 5.1 is always larger than that in the 2nd segment. If $j\leq\frac{n}{100}$, $3n\leq|f^{\mathrm{n}}(x^j)-f^{\mathrm{n}}(x^{i-1})|<n^4$, thus both $x^j$ and $x^{i-1}$ are just evaluated once. Then, we get $\mathrm{P}(\hat{f}(x^j)\geq\hat{f}(x^{i-1}))=1/n$, since $\hat{f}(x^j)\geq\hat{f}(x^{i-1})$ iff $\hat{f}(x^j)=(2n+j)^3$. Note that $\hat{f}$ is just $f^{\mathrm{n}}$ here, since it only performs one evaluation.
(4) $0<i\leq\frac{n}{200}+1$. If $j>\frac{n}{200}$, $0\leq f^{\mathrm{n}}(x^j)\leq n^4$. Note that $f^{\mathrm{n}}(x^{i-1})=n^4(n-i+1)$ or $f^{\mathrm{n}}(x^{i-1})\leq -n^4$. Thus, $|f^{\mathrm{n}}(x^j)-f^{\mathrm{n}}(x^{i-1})|\geq n^4$. If $j\leq\frac{n}{200}$, we can easily derive that $|f^{\mathrm{n}}(x^j)-f^{\mathrm{n}}(x^{i-1})|<n$ or $\geq n^4$. Thus, for any $j\geq i$, both $x^j$ and $x^{i-1}$ will be evaluated $n^5$ times. Let $Y=f^{\mathrm{n}}(x^{i-1})-f^{\mathrm{n}}(x^j)$. It is easy to verify $\mu:=\mathrm{E}(Y)\geq n^4/5$ and $\sigma^2:=\mathrm{Var}(Y)\leq 2n^{10}$. By Chebyshev's inequality, $\mathrm{P}(\hat{f}(x^j)\geq\hat{f}(x^{i-1}))\leq\frac{\sigma^2}{m\mu^2}\leq\frac{50}{n^3}$.

Thus, we have shown $\forall 0<i\leq j:\mathrm{P}(\hat{f}(x^j)\geq\hat{f}(x^{i-1}))\leq\log n/(15n)$ for sufficiently large $n$. Let $l=\log n$ and $c=1/15$. The condition of Lemma 5.5 is satisfied and the expected number of iterations is thus $O(n\log n)+n2^{O(\log n)}$, i.e. polynomial. Since in each iteration, a solution is evaluated by at most $n^5$ times, the expected running time is polynomial. □

## 6 CONCLUSION

In this paper, we analyze the effectiveness of sampling in noisy evolutionary optimization via running time analysis. Our analysis on noisy LeadingOnes shows that as the sample size increases, the running time of the (1+1)-EA first reduces from exponential to polynomial, but then returns to exponential. This discloses the importance of selecting a proper sample size. We also construct two artificial noisy problems to show that when sampling with any fixed sample size fails, using parent populations and adaptive sampling can work. Real noisy problems will be studied in the future.

## ACKNOWLEDGMENTS

We would like to thank Per Kristian Lehre for helpful discussions and thank the anonymous reviewers for their valuable comments. This work was supported by the Ministry of Science and Technology of China (2017YFC0804003), the NSFC (61603367, 61672478), the YESS (2016QNRC001), the Science and Technology Innovation Committee Foundation of Shenzhen (ZDSYS201703031748284), and the Royal Society Newton Advanced Fellowship (NA150123).

## REFERENCES

[1] Y. Akimoto, S. Astete-Morales, and O. Teytaud. 2015. Analysis of runtime of optimization algorithms for noisy functions over discrete codomains. *Theoretical Computer Science* 605 (2015), 42–50.
[2] A. Auger and B. Doerr. 2011. *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. World Scientific, Singapore.
[3] J. Branke and C. Schmidt. 2004. Sequential sampling in noisy environments. In *Proceedings of PPSN'04*. Birmingham, UK, 202–211.
[4] E. Cantú-Paz. 2004. Adaptive sampling for noisy problems. In *Proceedings of GECCO'04*. Seattle, WA, 947–958.
[5] D.-C. Dang and P. K. Lehre. 2015. Efficient optimisation of noisy fitness functions with population-based evolutionary algorithms. In *Proceedings of FOGA'15*. Aberystwyth, UK, 62–68.
[6] B. Doerr, A. Hota, and T. Kötzing. 2012. Ants easily solve stochastic shortest path problems. In *Proceedings of GECCO'12*. Philadelphia, PA, 17–24.
[7] B. Doerr, D. Johannsen, and C. Winzen. 2012. Multiplicative drift analysis. *Algorithmica* 64, 4 (2012), 673–697.
[8] S. Droste. 2004. Analysis of the (1+1) EA for a noisy OneMax. In *Proceedings of GECCO'04*. Seattle, WA, 1088–1099.
[9] M. Feldmann and T. Kötzing. 2013. Optimizing expected path lengths with ant colony optimization using fitness proportional update. In *Proceedings of FOGA'13*. Adelaide, Australia, 65–74.
[10] T. Friedrich, T. Kötzing, M. Krejca, and A. Sutton. 2016. Robustness of ant colony optimization to noise. *Evolutionary Computation* 24, 2 (2016), 237–254.
[11] T. Friedrich, T. Kötzing, M. Krejca, and A. Sutton. 2017. The compact genetic algorithm is efficient under extreme Gaussian noise. *IEEE Transactions on Evolutionary Computation* 21, 3 (2017), 477–490.
[12] C. Gießen and T. Kötzing. 2016. Robustness of populations in stochastic environments. *Algorithmica* 75, 3 (2016), 462–489.
[13] J. He and X. Yao. 2001. Drift analysis and average time complexity of evolutionary algorithms. *Artificial Intelligence* 127, 1 (2001), 57–85.
[14] F. Neumann and C. Witt. 2010. *Bioinspired Computation in Combinatorial Optimization: Algorithms and Their Computational Complexity*. Springer-Verlag, Berlin, Germany.
[15] P. Oliveto and C. Witt. 2011. Simplified drift analysis for proving lower bounds in evolutionary computation. *Algorithmica* 59, 3 (2011), 369–386.
[16] P. Oliveto and C. Witt. 2012. Erratum: Simplified drift analysis for proving lower bounds in evolutionary computation. *arXiv:1211.7184* (2012).
[17] A. Prügel-Bennett, J. Rowe, and J. Shapiro. 2015. Run-time analysis of population-based evolutionary algorithm in noisy environments. In *Proceedings of FOGA'15*. Aberystwyth, UK, 69–75.
[18] C. Qian, C. Bian, W. Jiang, and K. Tang. 2017. Running time analysis of the (1+1)-EA for OneMax and LeadingOnes under bit-wise noise. *arXiv:1711.00956* (2017).
[19] C. Qian, Y. Yu, K. Tang, Y. Jin, X. Yao, and Z.-H. Zhou. 2018. On the effectiveness of sampling for evolutionary optimization in noisy environments. *Evolutionary Computation* 26, 2 (2018), 60–90.
[20] C. Qian, Y. Yu, and Z.-H. Zhou. 2018. Analyzing evolutionary optimization in noisy environments. *Evolutionary Computation* 26, 1 (2018), 1–41.
[21] D. Sudholt and C. Thyssen. 2012. A simple ant colony optimizer for stochastic shortest path problems. *Algorithmica* 64, 4 (2012), 643–672.
[22] A. Syberfeldt, A. Ng, R. John, and P. Moore. 2010. Evolutionary optimisation of noisy multi-objective problems using confidence-based dynamic resampling. *European Journal of Operational Research* 204, 3 (2010), 533–544.
[23] I. S. Tyurin. 2010. An improvement of upper estimates of the constants in the Lyapunov theorem. *Russian Mathematical Surveys* 65, 3 (2010), 201–202.
[24] Carsten Witt. 2006. Runtime analysis of the ($\mu$+1) EA on simple pseudo-Boolean functions. *Evolutionary Computation* 14, 1 (2006), 65–86.
[25] Y. Yu, C. Qian, and Z.-H. Zhou. 2015. Switch analysis for running time analysis of evolutionary algorithms. *IEEE Transactions on Evolutionary Computation* 19, 6 (2015), 777–792.
[26] Z. Zhang and T. Xin. 2007. Immune algorithm with adaptive sampling in noisy environments and its application to stochastic optimization problems. *IEEE Computational Intelligence Magazine* 2, 4 (2007), 29–40.

## 7 APPENDIX

**Proof of Lemma 5.3.** Let $X_i^t$ denote the number of solutions with $i$ 1-bits in $\xi_t$ (i.e., the population after $t$ iterations). Let $a=\frac{599n}{600}$ and $b=20$. We first use an inductive proof to show that

$$\forall t\geq 0, i>a:\ \mathrm{E}(X_i^t)\leq \mu b^{a-i}. \tag{7}$$

For $t=0$, due to the uniform initial distribution, we easily have $\mathrm{E}(X_i^0)=\mu\cdot\binom{n}{i}/2^n)$. Note that for $k\geq\frac{2n}{3}$, $\binom{n}{k+1}/\binom{n}{k}=\frac{n-k}{k+1}\leq\frac{n/3}{2n/3+1}\leq\frac{1}{2}$. Thus, $\binom{n}{i}/2^n\leq\binom{n}{\frac{3n}{4}}/\binom{n}{\frac{2n}{3}}\leq(\frac{1}{2})^{n/12}\leq b^{a-n}$, which implies that $\forall i>a, \mathrm{E}(X_i^0)\leq\mu b^{a-i}$. We then assume that $\forall 0\leq k\leq t, i>a:\ \mathrm{E}(X_i^k)\leq\mu b^{a-i}$, and analyze $\mathrm{E}(X_i^{t+1})$ for any $i>a$. Let $\boldsymbol{X^t}=(X_0^t,X_1^t,...,X_n^t)$, $\boldsymbol{l}=(l_0,l_1,...,l_n)$, $|\boldsymbol{l}|_1=\sum_{i=0}^n l_i$ and $p=\frac{3}{5(\mu+1)}$. Let $x'$ denote the offspring solution generated in the $(t+1)$-th iteration, and let $x^i$ denote a solution with $i$ 1-bits. Then, we have

$$\begin{aligned}
&\mathrm{E}(X_i^{t+1}-X_i^t)=\mathrm{E}(\mathrm{E}(X_i^{t+1}-X_i^t\mid\boldsymbol{X^t}))\\
&=\textstyle\sum_{|\boldsymbol{l}|_1=\mu}\mathrm{P}(\boldsymbol{X^t}=\boldsymbol{l})\cdot\\
&\quad\big(\mathrm{P}(|x'|_1=i,x'\text{ and any }x^i\text{ in }\xi_t\text{ are not deleted}\mid\boldsymbol{X^t}=\boldsymbol{l})\\
&\quad-\mathrm{P}(|x'|_1\neq i,\text{one }x^i\text{ in }\xi_t\text{ is deleted}\mid\boldsymbol{X^t}=\boldsymbol{l})\big)\\
&\leq\textstyle\sum_{|\boldsymbol{l}|_1=\mu}\mathrm{P}(\boldsymbol{X^t}=\boldsymbol{l})\cdot\big(\mathrm{P}(|x'|_1=i\mid\boldsymbol{X^t}=\boldsymbol{l})\cdot(1-(l_i+1)p)\\
&\quad-(1-\mathrm{P}(|x'|_1=i\mid\boldsymbol{X^t}=\boldsymbol{l}))\cdot l_ip\big)\\
&=\textstyle\sum_{|\boldsymbol{l}|_1=\mu}\mathrm{P}(\boldsymbol{X^t}=\boldsymbol{l})\cdot\big(\mathrm{P}(|x'|_1=i\mid\boldsymbol{X^t}=\boldsymbol{l})\cdot(1-p)-l_ip\big)\\
&=\textstyle\sum_{|\boldsymbol{l}|_1=\mu}\mathrm{P}(\boldsymbol{X^t}=\boldsymbol{l})\cdot\big(\sum_{j=0}^n\frac{l_j}{\mu}\cdot\mathrm{P}_{mut}(x^j,x^i)\cdot(1-p)-l_ip\big)\\
&=(1-p)\textstyle\sum_{j=0}^n\mathrm{P}_{mut}(x^j,x^i)\cdot\sum_{|\boldsymbol{l}|_1=\mu}\mathrm{P}(\boldsymbol{X^t}=\boldsymbol{l})\frac{l_j}{\mu}-\sum_{|\boldsymbol{l}|_1=\mu}\mathrm{P}(\boldsymbol{X^t}=\boldsymbol{l})l_ip\\
&=(1-p)\textstyle\sum_{j=0}^n\mathrm{P}_{mut}(x^j,x^i)\cdot\sum_{l_j=0}^{\mu}\mathrm{P}(X_j^t=l_j)\frac{l_j}{\mu}-\sum_{l_i=0}^{\mu}\mathrm{P}(X_i^t=l_i)l_ip\\
&=\frac{1-p}{\mu}\cdot\sum_{j=0}^n\mathrm{P}_{mut}(x^j,x^i)\cdot\mathrm{E}(X_j^t)-p\cdot\mathrm{E}(X_i^t),
\end{aligned}$$

where the second equality is because $X_i^{t+1}-X_i^t=1$ iff $|x'|=i$ and $x'$ is added into the population meanwhile the solutions with $i$ 1-bits in $\xi_t$ are not deleted; $X_i^{t+1}-X_i^t=-1$ iff $|x'|\neq i$ and one solution with $i$ 1-bits in $\xi_t$ is deleted, the first inequality is because any solution with $i$ 1-bits is deleted with probability at least $p=\frac{3}{5(\mu+1)}$ by the condition Eq. (6), and the fourth equality is since a parent solution is uniformly selected from $\xi_t$ for mutation. We further derive an upper bound on $\frac{1}{\mu}\cdot\sum_{j=0}^n\mathrm{P}_{mut}(x^j,x^i)\cdot\mathrm{E}(X_j^t)$ as follows:

$$\begin{aligned}
&\frac{1}{\mu}\cdot\sum_{j=0}^n\mathrm{P}_{mut}(x^j,x^i)\cdot\mathrm{E}(X_j^t)\\
&\leq\mathrm{P}_{mut}(x^a,x^i)+\sum_{j=a+1}^{i-1}b^{a-j}\cdot\binom{n-j}{i-j}\left(\frac{1}{n}\right)^{i-j}\\
&\quad+b^{a-i}\cdot\left(\left(1-\frac{1}{n}\right)^n+\sum_{k=1}^{n-i}\binom{n-i}{k}\left(\frac{1}{n}\right)^k\right)+\sum_{j=i+1}^n b^{a-j}\\
&\leq\binom{n-a}{i-a}\left(\frac{1}{n}\right)^{i-a}+b^{a-i}\sum_{j=a+1}^{i-1}b^{i-j}\left(\frac{n-j}{n}\right)^{i-j}\\
&\quad+b^{a-i}\left(\frac{1}{e}+\sum_{k=1}^{n-i}\left(\frac{n-i}{n}\right)^k\right)+b^{a-i}\sum_{j=i+1}^n b^{i-j}\\
&\leq\left(\frac{n-a}{n}\right)^{i-a}+b^{a-i}\cdot\left(\sum_{j=a+1}^{i-1}b^{i-j}\left(\frac{n-a}{n}\right)^{i-j}\right.\\
&\quad\left.+\frac{1}{e}+\sum_{k=1}^{n-i}\left(\frac{n-a}{n}\right)^k+\sum_{j=i+1}^n b^{i-j}\right)\\
&\leq b^{a-i}\left(\left(\frac{1}{b}\cdot\frac{n}{n-a}\right)^{a-i}+\frac{1}{\frac{n}{b(n-a)}-1}+\frac{1}{e}+\frac{1}{\frac{n}{n-a}-1}+\frac{1}{b-1}\right)\\
&\leq b^{a-i}/2,
\end{aligned}$$

where the first inequality is derived by applying $\forall j\leq a:\ \mathrm{P}_{mut}(x^j,x^i)\leq\mathrm{P}_{mut}(x^a,x^i)$, $\sum_{j=0}^n\mathrm{E}(X_j^t)=\mu$, $\forall j>a:\ \mathrm{E}(X_j^t)\leq\mu b^{a-j}$ and some simple upper bounds on $\mathrm{P}_{mut}(x^j,x^i)$ for $j>a$, the fourth inequality is by $\forall 0<c<1:\sum_{k=1}^{\infty}c^k=\frac{c}{1-c}=\frac{1}{1/c-1}$, and the last is by $a=\frac{599n}{600}$, $b=20$ and $i>a$. Combining the above two formulas, we get

$$\mathrm{E}(X_i^{t+1}-X_i^t)\leq\frac{1-p}{2}\cdot b^{a-i}-p\cdot\mathrm{E}(X_i^t),$$

which implies that

$$\begin{aligned}
\mathrm{E}(X_i^{t+1})&\leq\frac{1-p}{2}\cdot b^{a-i}+(1-p)\cdot\mathrm{E}(X_i^t)\\
&\leq\left(\frac{1}{2\mu}+1\right)\cdot\frac{5\mu+2}{5(\mu+1)}\cdot\mu b^{a-i}\leq\mu b^{a-i},
\end{aligned}$$

where the second inequality is by $p=\frac{3}{5(\mu+1)}$ and $\mathrm{E}(X_i^t)\leq\mu b^{a-i}$, and the last inequality holds with $\mu\geq 2$. Thus, our claim that $\forall t\geq 0,\forall i>a:\mathrm{E}(X_i^t)\leq\mu b^{a-i}$ holds.

Based on Eq. (7) and Markov's inequality, we get, for any $t\geq 0$, $\mathrm{P}(X_n^t\geq 1)\leq\mathrm{E}(X_n^t)\leq\mu b^{a-n}$. Note that $X_n^t$ is the number of optimal solutions in the population after $t$ iterations. Let $T=b^{(n-a)/2}$. Then, the probability of finding the optimal solution $1^n$ in $T$ iterations is

$$\mathrm{P}(\exists t\leq T,X_n^t\geq 1)\leq\sum_{t=0}^T\mathrm{P}(X_n^t\geq 1)\leq T\cdot\mu b^{a-n}=\mu\cdot b^{(a-n)/2},$$

which is exponentially small for $\mu\in poly(n)$. This implies that the expected running time for finding the optimal solution is exponential. □