Sequential test sampling for stochastic derivative-free optimization

A. Ding[*] F. Rinaldi [†] L. N. Vicente[‡]

September 18, 2025

Abstract

In many derivative-free optimization algorithms, a sufficient decrease condition decides whether to accept a trial step in each iteration. This condition typically requires that the potential objective function value decrease of the trial step, i.e., the true reduction in the objective function value that would be achieved by moving from the current point to the trial point, be larger than a multiple of the squared stepsize. When the objective function is stochastic, evaluating such a condition accurately can require a large estimation cost.

In this paper, we frame the evaluation of the sufficient decrease condition in a stochastic setting as a hypothesis test problem and solve it through a sequential hypothesis test. The two hypotheses considered in the problem correspond to accepting or rejecting the trial step. This test sequentially collects noisy sample observations of the potential decrease until their sum crosses either a lower or an upper boundary depending on the noise variance and the stepsize. When the noise of observations is Gaussian, we derive a novel sample size result, showing that the effort to evaluate the condition explicitly depends on the potential decrease, and that the sequential test terminates early whenever the sufficient decrease condition is away from satisfaction. Furthermore, when the potential decrease is $\Theta(\delta^r)$ for some $r \in (0, 2]$, the expected sample size decreases from $\Theta(\delta^{-4})$ to $O(\delta^{-2-r})$.

We apply this sequential test sampling framework to probabilistic-descent direct search. To analyze its convergence rate, we extend a renewal-reward supermartingale-based convergence rate analysis framework to an arbitrary probability threshold. By doing so, we are able to show that probabilistic-descent direct search has an iteration complexity of $O(n/\epsilon^2)$ for gradient norm. Our numerical experiments indicate the superiority of sequential hypothesis testing over fixed sampling when dealing with the evaluation of stochastic sufficient decrease conditions.

1 Introduction

In this paper, we consider an unconstrained optimization problem of the form

$$\min_{x \in \mathbb{R}^n} f(x),$$

[*]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015-1582, USA (and523@lehigh.edu).

[†]Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Italy (rinaldi@math.unipd.it).

[‡]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015-1582, USA (lnv@lehigh.edu).

where the function values of the objective function $f : \mathbb{R}^n \to \mathbb{R}$ are not available directly. Instead, we have access to $F(x, \xi)$ as a noisy observation of $f(x)$, where ξ is a random variable. We assume that the observed noise is unbiased

$$E_\xi[F(x, \xi)] = f(x). \tag{1.1}$$

We also assume that the derivatives of f are not available or that the cost of computing them is unaffordable. Furthermore, the objective function f is assumed to be continuously differentiable and bounded from below by f^*. Its gradient ∇f is assumed to be L_f-Lipschitz continuous. The non-availability of derivatives is a common scenario in many simulation-based optimization applications and is the main subject of derivative-free optimization (DFO). In DFO, the only information available about f is through zeroth-order oracles (in our case, stochastic zeroth-order oracles), and the cost of querying the zeroth-order oracle is expensive. The goal of DFO is to achieve good solutions with few zeroth-order oracle queries.

Trust-region, direct-search, and line-search methods are three popular classes of algorithms in DFO. Trust-region methods build local surrogate models of f and compute trial steps and decide its acceptance using those models, while direct search explores the space directly via a set of search directions or points, without explicit models. In the middle of the spectrum between with and without models, line-search methods approximate the steepest descent direction using finite differences and searches along it. For a more comprehensive understanding of these classes of DFO methods, interested readers are encouraged to consult sources such as, e.g., [8, 19]. Most instances of these algorithms rely on a decrease condition to ensure that each step taken by the algorithm makes meaningful progress toward reducing the objective function. In particular, probabilistic-descent direct search is a direct-search algorithm proposed in [15], that instead of using a positive spanning set (see, for example, [8]) to ensure a descent direction (which may require at least $n + 1$ function evaluations), incorporates randomness into the algorithm to obtain a descent direction probabilistically using only one point. It is shown in [15] that, when the objective function is deterministic, this algorithm has a zeroth-order oracle complexity of $O(n/\epsilon^2)$ for gradient norm. In this paper, we consider probabilistic-descent direct search for the purpose of applying sequential hypothesis testing to the evaluation of sufficient decrease conditions in stochastic DFO.

1.1 A brief literature review of stochastic derivative-free optimization

We start by giving a brief overview of the main results in the literature on stochastic derivative-free optimization. To our knowledge, they all require, under the standard assumption of noise exhibiting finite variance, a sample size of $\Theta(\delta^{-4})$ function estimates per iteration, where δ is a stepsize or a trust-region radius (see, e.g., [23] and references therein for further details on this matter).

A trust-region method is designed in [18] to address optimization problems involving noisy objective functions. The authors establish convergence guarantees under conditions where the objective function f exhibits adequate smoothness properties (specifically, possessing a Lipschitz continuous gradient) and when the noise is independently drawn from a distribution with zero mean and finite variance. ASTRO-DF, proposed in [24] and refined in [17], is an adaptive sampling trust-region method designed for objective functions that maintain Lipschitz continuous gradients and can be accessed through a Monte Carlo oracle. The framework in [24] assumes noise that is independently distributed with zero mean, finite variance, and a bounded 4νth moment (where $\nu \geq 2$), and develops an almost sure convergence result.

Furthermore, [7] provides another significant contribution by examining a trust-region algorithm for unconstrained stochastic optimization. This work focuses on random models derived from a smooth objective function using stochastic observations of either the function itself or its gradient. The convergence analysis and rates for such methodologies are further detailed in [5] using martingale theory, which is applicable to a wide class of stochastic algorithms including direct search. The theoretical frameworks developed in [5, 6, 7] represent extensions of the probabilistic trust-region DFO approach originally outlined in [3] for deterministic functions. Each of these trust-region algorithms requires functions to possess some level of smoothness (such as Lipschitz continuous gradients) and relies on the ability to construct probabilistically accurate gradient approximations.

A comprehensive review of stochastic direct-search variants is provided in the survey [12, Chapter 4] for both smooth and non-smooth objective functions. StoMADS, proposed in [2], is a stochastic variant of the mesh adaptive direct search (MADS) algorithm. StoMADS generates an asymptotically dense set of search directions and is proved in [2] using martingale theory to converge to a Clarke stationary point of a locally Lipschitz continuous function with probability one. In another line of research, the work in [11] considers stochastic direct-search methods of directional type and presents its convergence rate analysis by utilizing the supermartingale-based framework in [5]. In [23], a new probabilistic tail-bound condition for function estimation is introduced under which stochastic direct-search and trust-region methods are shown to converge globally. Reduction in sample complexity is obtained under stronger assumptions than the standard finite noise variance. More specifically, under a bounded $q/(1-q)$ moment assumption, using $c\delta^q$ with $q \in (1, 2]$ as a threshold for decrease in the acceptance test, the authors give a $O(\delta^{-2q})$ sample complexity. Furthermore, under the assumption of using a common number generator framework and correlated errors (satisfied when, e.g., the noise is modeled as a Gaussian process), the authors give a $O(\delta^{2-2q})$ instead. As noted in [23, Remark 5.1], the improvement in the number of samples per iteration does not however necessarily lead to a reduction in the overall computational cost of the considered algorithmic frameworks. More specifically, using a decrease threshold of $c\delta^q$, while reducing the number of samples needed to certify a step, it increases the iteration complexity. In fact, in the case of smooth objectives with stochastic oracles, an iteration complexity of $O(n\epsilon^{-q/(q-1)})$ for gradient norm, with $q \in (1, 2]$, was proved in [12] for a direct-search scheme similar to the one given in [23].

It is finally important to highlight that all the methods mentioned above are fixed-sampling schemes, which always pay the worst-case cost. This basically means that, in order to satisfy the assumptions needed for convergence, one always needs to take the prescribed number of samples, no matter how obvious the decision is. In [1], the authors gave a sequential sampling strategy for a stochastic direct-search scheme for which a sequential hypothesis test has also been given. They claim a $O((\log T)^{\frac{2}{3}} T^{\frac{2}{3}})$ regret bound with respect to a sample budget T for smooth and strongly convex objectives, which would translate, neglecting the polylog term, into an overall complexity of $O(n\epsilon^{-6})$ for the gradient norm.

1.2 Our contribution

In Section 2, we introduce a new way of testing the satisfaction of a sufficient decrease condition in stochastic derivative-free optimization by framing it as a hypothesis test problem and solving it through the means of a sequential hypothesis test. The test makes a decision between two hypotheses, essentially corresponding to accepting or rejecting the sufficient decrease condition, outputting the probabilities of making correct and incorrect decisions. For the purpose of establishing a standard non-convex iteration complexity result, such probabilities need to satisfy certain bounds dependent on the algorithmic stepsize to ensure enough correctness. Specifically, this test sequentially collects noisy observations of the potential decrease, by calling a zeroth-order oracle (at the current and trial points) until their sum crosses either a lower or an upper bound depending on the function variance and the stepsize. When the function noise is Gaussian, we show that the size of the sample required to estimate the decrease drops significantly when the potential decrease is far from a multiple of the square of the stepsize, in which case we observe an early termination of the test. Furthermore, when the potential decrease is $\Theta(\delta^r)$ for some $r \in (0, 2]$, we show that the expected sample size decreases from the known $\Theta(\delta^{-4})$ to $O(\delta^{-2-r})$.

In Section 3, we apply this sequential test sampling framework to probabilistic-descent direct search when the function is stochastic. We first show that the expected decrease of an auxiliary merit function is sufficiently large when compared with the square of the stepsize and that the stepsize does not approach zero at non-stationary points. We then extend an existing renewal-reward supermartingale-based convergence rate analysis framework to a general case where the probability defining the Bernoulli process in (B.1) is arbitrary. Finally, we conclude that the iteration complexity of probabilistic-descent direct search is $O(n/\epsilon^2)$. In Section 4, our numerical results indicate the superiority of the sequential hypothesis test against a fixed sample test in evaluating the sufficient decrease condition when the function is stochastic.

The use of our sequential test sampling framework, guarantees a similar iteration complexity as the deterministic case while guaranteeing, under the Gaussian noise assumption, a reduced sample complexity with respect to the standard sample complexity obtained considering the finite variance noise assumption. Furthermore, the sequential test sampling guarantees more flexibility than fixed-sampling schemes. In fixed-sample approaches, such as those considered in [23], the same order of samples must indeed be taken at every iteration to satisfy the required convergence conditions, even when the trial step is clearly acceptable or clearly unacceptable, thus leading to potentially waste in terms of function evaluations. On the other side, the sequential test dynamically adapts the number of samples to the given scenario. Hence, when the decrease is clearly above or below the acceptance threshold, the test terminates after generating just a small number of samples. This way of adapting the number of generated samples gives a reduction of the expected per-iteration cost and provides a more efficient alternative to fixed-sampling strategies that always require, as mentioned before, the worst-case sample complexity.

2 Sequential hypothesis testing framework for stochastic DFO

A hypothesis test problem is a statistical framework used to decide between two competing hypotheses about a population based on observed samples. Two hypotheses are called the null hypothesis H_0 and the alternative hypothesis H_1. This problem is widely investigated in statistical inference.

Sequential hypothesis testing is a statistical method that uses sequential hypothesis tests to solve hypothesis test problems. It has a long history, which, according to its pioneer Wald [26], may date back to the work [9] of Dodge and Romig in 1929. Sigmund [25] points out that sequential hypothesis testing was developed to solve hypothesis test problems more efficiently, which also happens to be our goal.

In fact, many nonlinear optimization algorithms accept steps at a given iteration based on the satisfaction of a decrease condition on the value of the objective function. In derivative-free optimization, such a condition consists of imposing a simple decrease or a sufficient decrease on the objective function related to the size of a step. We are going to apply sequential hypothesis testing to enhance the satisfaction of the sufficient decrease condition when the function is stochastic.

2.1 Testing a sufficient decrease condition

Denote the current iterate of a DFO algorithm by x, the current stepsize by δ, and the current candidate point by $x + \delta d$, where d is a certain direction. Suppose that the candidate point $x + \delta d$ is accepted if the sufficient decrease condition

$$f(x) - f(x + \delta d) - c\delta^2 \geq 0 \tag{2.1}$$

is satisfied. Our goal is to reformulate its evaluation as a hypothesis test problem which can be used by the algorithm at stake.

For this purpose, denote by $F(x, \xi^x)$ one random observation of $f(x)$ and by $F(x + \delta d, \xi^d)$ one random observation of $f(x + \delta d)$. The notation ξ^x and ξ^d is used to clarify that $F(x, \xi^x)$ and $F(x + \delta d, \xi^d)$ are sampled independently. Let us define a random variable

$$Y = c\delta^2 - (F(x, \xi^x) - F(x + \delta d, \xi^d)). \tag{2.2}$$

We have from the unbiased noise assumption (1.1) that

$$E[Y] = c\delta^2 - (f(x) - f(x + \delta d)). \tag{2.3}$$

Since we do not know whether the sufficient decrease condition (2.1) is satisfied, it follows from (2.3) that two hypotheses to be considered in an algorithm are

$$H_0 : E[Y] \leq 0$$
$$H_1 : E[Y] > 0.$$

Now it becomes clear that our hypothesis test problem is to decide whether the mean of a random variable is positive or not through observations. Since we are not directly interested in the value of $E[Y]$, our hypothesis test problem is different from a parameter estimation problem. A very accurate estimate of $E[Y]$ can be a burden when $E[Y]$ is far from 0. If the first few observations tell us that $E[Y]$ may be far from 0, then we may not want to obtain a very accurate estimate of it.

2.2 The hypothesis test problem and sequential hypothesis tests

To illustrate how to test the sufficient decrease condition in a sequential hypothesis testing framework, we first state our hypothesis test problem.

Problem 2.1 (Hypothesis test problem) *Let Y be a random variable. The mean of the random variable Y is denoted by $\mu \in \mathbb{R}$ and is unknown. The hypothesis test problem is to decide between two hypotheses*

$$\begin{aligned} H_0 &: \mu \leq 0 \\ H_1 &: \mu > 0 \end{aligned}$$

on the basis of m independent observations $Y^1, \ldots, Y^m$ drawn from Y, where m is a random variable called stopping rule and defined on every sample $\omega = (Y^1, Y^2, \ldots)$.

Then we give the definition of a sequential hypothesis test, which is adopted from [28].

Definition 1 (Sequential hypothesis test) *A sequential hypothesis test consists of a stopping rule $m(\omega)$ and a decision rule to decide H_0 or H_1 in a hypothesis test problem.*

A sequential hypothesis test is usually expected to end with a finite number of observations almost surely. We give the following definition with respect to this property.

Definition 2 *We say that a sequential hypothesis test ends properly if*

$$P(m(\omega) < \infty) = 1. \tag{2.4}$$

Finally, for Problem 2.1, we give the following definition regarding the accuracy property of a given sequential hypothesis test.

Definition 3 (C-accurate sequential hypothesis test) *For Problem 2.1, we say that a sequential hypothesis test is C-accurate if its error probabilities satisfy*

$$P(H_1 \text{ is accepted} \,|\, \mu \leq 0) \leq \frac{1}{2} \tag{2.5}$$

$$P(H_0 \text{ is accepted} \,|\, \mu > 0) \leq \frac{C}{\mu}. \tag{2.6}$$

Two important properties of a sequential hypothesis test are its probability of accepting each hypothesis and its expected number of observations used or expected sample size. For $j = 0, 1$, we define that the acceptance region $S_j = \{\omega : H_j \text{ is accepted}\}$ of a sequential hypothesis test is the sample set where H_j is accepted. We denote the expected sample size by $E_\mu[m] = E[m|\mu]$ and the probability of accepting H_j by $P_\mu(S_j) = P(S_j|\mu)$ for $j = 0, 1$, when the mean of Y has the value of μ. If $\mu > 0$, then $P_\mu(S_0)$ is the error probability in (2.6). Similarly $P_\mu(S_1)$ is the error probability in (2.5) when $\mu \leq 0$.

A C-accurate sequential hypothesis test draws inferences of the mean of a random variable and delivers accuracy conditions (2.5) and (2.6) for its error probabilities. Conditions (2.5) and (2.6) represent a certain level of accuracy requirement for the error of the solution of Problem 2.1 and will be used in Section 3 to prove a convergence rate or complexity result.

2.3 The proposed sequential hypothesis test

Now we propose a sequential hypothesis test (see Test 2.1) for Problem 2.1 and study its properties.

```
**Test 2.1 (A sequential hypothesis test)**
    Specify $\{a_l\}$ and $\{b_l\}$ such that $a_l, b_l \in [-\infty, \infty]$, and $a_l \geq b_l$ for each $l$.
    **Repeat** for $l = 1, 2, \ldots$
        Draw a new i.i.d. observation $Y^l$ from $Y$.
    **Until** $\sum_{i=1}^{l} Y^i \geq a_l$ or $\sum_{i=1}^{l} Y^i \leq b_l$.
    Record the number of used samples with $m(\omega) = l$.
    Decide that $H_0$ is true if $\sum_{i=1}^{m} Y^i \leq b_m$.
    Decide that $H_1$ is true if $\sum_{i=1}^{m} Y^i \geq a_m$.
```

Therefore, Test 2.1 continues to draw observations from Y until one of the two termination conditions is satisfied. To check whether Test 2.1 fits the definition of a sequential hypothesis test, we notice that the stopping rule in Test 2.1 is

$$m(\omega) = \inf\{l \geq 1 : \sum_{i=1}^{l} Y^i \geq a_l \text{ or} \sum_{i=1}^{l} Y^i \leq b_l\}.$$

Then Test 2.1 decides that H_0 or H_1 is accepted based on which termination condition is satisfied. Specifically, it decides that H_0 is true if $\sum_{i=1}^{m} Y_i$ is smaller than b_m and that H_1 is true if $\sum_{i=1}^{m} Y_i$ is larger than a_m.

To see the generality of Test 2.1, we notice that a sampling procedure with a predetermined fixed sample size at each iteration, which is what most stochastic DFO algorithms employ, is also an instance of Test 2.1. We describe such a sampling procedure in the following test.

```
**Test 2.2 (A test with fixed sample size $m$)**
    Draw $m$ i.i.d. observation $Y^1, Y^2, \ldots, Y^m$ from $Y$.
    Decide that $H_0$ is true if $\sum_{i=1}^{m} Y^i \leq 0$.
    Decide that $H_1$ is true if $\sum_{i=1}^{m} Y^i > 0$.
```

We can easily check that Test 2.2 is an instance of Test 2.1, by choosing the parameters in Test 2.2 as $a_l = \infty$ and $b_l = -\infty$ for $l < m$ and $a_m = b_m = 0$. Test 2.2 is usually referred to as a fixed sample size hypothesis test in the literature of statistics. It can be shown through Markov's inequality that Test 2.2 delivers the accuracy condition (2.6) when at least $m = \sigma^2 C^{-2}$ samples are used. When C is $\Theta(\delta^2)$, which is commonly required in stochastic DFO, Test 2.2 requires a sample set of size $\Theta(\delta^{-4})$, which is known to be what most stochastic DFO algorithms need per iteration for standard convergence and convergence rates. This may further explain that most stochastic DFO algorithms using Test 2.2 require a sample size of $\Theta(\delta^{-4})$ per iteration.

To make sure that condition (2.4) holds and Test 2.1 ends properly, we need to choose $\{a_l\}$ and $\{b_l\}$ such that

$$P(\{b_l \leq \sum_{i=1}^{l} Y_i \leq a_l, \ \forall l\}) = 0. \tag{2.7}$$

Notice that $\sum_{i=1}^{l} Y_i$ is a one-dimensional random walk with i.i.d. increments. If $\{a_l\}$ and $\{b_l\}$ are two bounded sequences and Y is not almost surely zero, then condition (2.7) holds, and hence also (2.4).

Test 2.1 is C-accurate if we choose the appropriate parameters $\{a_l\}$ and $\{b_l\}$ such that conditions (2.5) and (2.6) hold. To investigate when condition (2.5) holds, the following lemma tells us that $P_\mu(S_j)$ of any Test 2.1 is monotone with respect to μ. The proof consists of two steps. In the first step, given any $\Delta\mu > 0$ and any sequential test Test 2.1, we define an ancillary sequential hypothesis test, called Test 2.1*, such that $P_{\mu+\Delta\mu}(S_j) = P_\mu(\tilde{S}_j)$, where S_j and $\tilde{S}_j$ are the acceptance regions of H_j in Test 2.1 and Test 2.1* accordingly. In the second step, we note that Test 2.1* makes each sample harder for H_0 and easier for H_1 than Test 2.1, and therefore, that $P_\mu(\tilde{S}_0) \leq P_\mu(S_0)$ and $P_\mu(\tilde{S}_1) \geq P_\mu(S_1)$.

Lemma 2.1 *Consider Problem 2.1 and a sequential test Test 2.1, whose acceptance regions are S_j, for $j = 0, 1$. Then $P_\mu(S_0)$ is non-increasing with respect to μ and $P_\mu(S_1)$ is non-decreasing with respect to μ.*

Proof. It suffices to prove $P_\mu(S_0) \geq P_{\mu+\Delta\mu}(S_0)$ and $P_\mu(S_1) \leq P_{\mu+\Delta\mu}(S_1)$ for any $\Delta\mu > 0$. For the purpose of the proof, we define an ancillary sequential hypothesis test in the form of Test 2.1, called Test 2.1*, by selecting parameters $\{a_l - l\Delta\mu\}$ and $\{b_l - l\Delta\mu\}$, where $\{a_l\}$ and $\{b_l\}$ are the parameters of the given Test 2.1. Denote its acceptance region by $\tilde{S}_j$ and its sample size by $\tilde{m}$.

We first prove that $P_{\mu+\Delta\mu}(S_j) = P_\mu(\tilde{S}_j)$. For each $\omega = (y_1, y_2, \ldots)$, we make a change of variable and define $\tilde{\omega} = (\tilde{y}_1, \tilde{y}_2, \ldots) = (y_1 - \Delta\mu, y_2 - \Delta\mu, \ldots)$. From the definition of Test 2.1*, it follows that $m(\omega) = \tilde{m}(\tilde{\omega})$ and $\omega \in S_j$ if and only if $\tilde{\omega} \in \tilde{S}_j$. It also follows from this change of variable that $P_{\mu+\Delta\mu}(S_j) = P_\mu(\tilde{S}_j)$.

It then suffices to prove that $P_\mu(\tilde{S}_0) \leq P_\mu(S_0)$ and $P_\mu(\tilde{S}_1) \geq P_\mu(S_1)$. From the definition of S_0 and $\tilde{S}_0$, we have $\tilde{S}_0 \subseteq S_0$ and $P_\mu(\tilde{S}_0) \leq P_\mu(S_0)$. Similarly, we have $\tilde{S}_1 \supseteq S_1$ and $P_\mu(\tilde{S}_1) \geq P_\mu(S_1)$. □

Lemma 2.1 tells us that $P_\mu(S_1)$ is non-decreasing, which implies that the left-hand side of (2.5) is non-decreasing and admits its maximum when $\mu = 0$. Hence to satisfy condition (2.5), it suffices to make sure that the maximum of the left-hand side of (2.5) is no larger than 1/2. We formalize this reasoning in the following lemma.

Lemma 2.2 *For any sequential test Test 2.1, one has that (2.5) holds if and only if*

$$P(H_1 \textit{ is accepted} \,|\, \mu = 0) \leq \frac{1}{2}. \tag{2.8}$$

Proof. We have, by definition, $P_\mu(S_1) = P(S_1 \,|\, \mu) = P(H_1 \text{ is accepted} \,|\, \mu)$. Then it follows from Lemma 1 that $P(H_1 \text{ is accepted} \,|\, \mu)$ is non-decreasing with respect to μ

$$P(H_1 \text{ is accepted} \,|\, \mu \leq 0) \leq P(H_1 \text{ is accepted} \,|\, \mu = 0).$$

So, as a function of μ, $P(H_1 \text{ is accepted} \,|\, \mu \leq 0)$ reaches its maximum when $\mu = 0$. Therefore, (2.5) holds if and only if $P(H_1 \text{ is accepted} \,|\, \mu = 0) \leq 1/2$. □

Lemma 2.2 gives us an equivalent condition (2.8) for (2.5) when Test 2.1 is used. One way of ensuring (2.5) through the satisfaction of (2.8) is when the probability density function $\phi_\mu(y)$ of Y is symmetric and $a_l + b_l = 0$ for each $l > 0$.

Lemma 2.3 *Assume that $\phi_\mu(\mu + y) = \phi_\mu(\mu - y)$ for any y. Select a_l and b_l in Test 2.1 such that $a_l + b_l = 0$ holds for each $l > 0$. Then (2.5) holds.*

Proof. The symmetry of the probability density function $\phi_\mu(y)$ and $a_l + b_l = 0$ implies $P(H_1 \text{ is accepted} \,|\, \mu = 0) = 1/2$. Then (2.8) holds and condition (2.5) follows from Lemma 2.2. □

Now we focus our attention on the satisfaction of condition (2.6). When we use Test 2.1 in the context of stochastic DFO algorithms, the value of C in (2.6) can be small, and so we need to investigate how large a sample size needs to be for the satisfaction of (2.6). We will show in the next subsection that, if Y in Problem 2.1 follows a Gaussian distribution, we can select parameters in Test 2.1 so that condition (2.6) holds.

2.4 Sample size in the Gaussian case

In this subsection, we assume that Y in Problem 2.1 follows a Gaussian distribution. Let $\phi_\mu(y) = (2\pi\sigma^2)^{-1/2} \exp\left(-(y-\mu)^2/(2\sigma^2)\right)$ be the probability density function of Y. We notice that, when Y is Gaussian, Test 2.1 with constant parameters is equivalent to a sequential probability ratio test in [26]. We will specify the parameters in Test 2.1 so that Test 2.1 is C-accurate. Then we approximate its expected number of observations $E_\mu[m]$.

We first select constant parameters in Test 2.1 as $a_l = -b_l = c^0$ for each l, where $c^0 > 0$ is a real number. Lemma 2.3 shows that condition (2.5) is valid. Furthermore, in the Gaussian case, Test 2.1 is equivalent to a sequential probability ratio test in [26], which has four real number parameters $A > 1$, $0 < B < 1$, θ_0, and $\theta_1 > \theta_0$, given that our parameter satisfies $c^0 = \sigma^2 \log A/(\theta_1 - \theta_0)$ and $-c^0 = \sigma^2 \log B/(\theta_1 - \theta_0)$.

For this sequential probability ratio test, Wald bounded the left-hand side of condition (2.6) in [26, (3.42)], and it follows that

$$P(H_0 \text{ is accepted} \,|\, \mu > 0) \leq A^{-h},$$

where h denotes $2\mu/(\theta_1 - \theta_0)$. We can use his result for our Test 2.1 because of the equivalence between the two tests. Since we have equivalently $c^0 = \sigma^2 \log A/(\theta_1 - \theta_0)$, we can express A with respect to c^0 as $A = \exp\left(c^0(\theta_1 - \theta_0)/\sigma^2\right)$. After substituting h and A, we have for Test 2.1 that

$$P(H_0 \text{ is accepted} \,|\, \mu > 0) \leq e^{-\frac{2c^0}{\sigma^2}\mu}.$$

It then follows from Proposition 2 in Appendix A with $x = \mu/C$ that, given $C > 0$, we can select $c^0 \geq \sigma^2/(2eC)$ so that for any $\mu > C$, we have $\exp(-2c^0\mu/\sigma^2) \leq C/\mu$. Therefore, by setting parameters $a_l = -b_l = c^0 \geq \sigma^2/(2eC)$ in Test 2.1, we can ensure that condition (2.6) holds and Test 2.1 is C-accurate in the Gaussian case.

Calculating the exact value of the expected sample size is more elaborate. Wald [27] gave an approximate expected sample size in the Gaussian case. For parameters $A = \exp\left(c^0(\theta_1 - \theta_0)/\sigma^2\right)$, $B = 1/A$, and $c^0 = \sigma^2/(2eC)$ in [27, (3:43)] and in [27, (3:57)], the number is as follows

$$E_\mu[m] \approx \frac{\sigma^2}{2eC\mu} \frac{e^{\frac{\mu}{eC}} - 1}{e^{\frac{\mu}{eC}} + 1}. \tag{2.9}$$

To help clarify it, it follows from Proposition 1 in Appendix A with $x = \mu/(eC)$ that we can bound (2.9) as follows

$$\frac{\sigma^2}{2eC\mu}\frac{e^{\frac{\mu}{eC}}-1}{e^{\frac{\mu}{eC}}+1} \leq \frac{\sigma^2}{4e^2C^2}\min\left(1, \frac{eC}{|\mu|}\right). \tag{2.10}$$

In the context of stochastic DFO algorithms, the level of accuracy is $C = s\delta^2$ (for some constant $s > 0$) and $\mu = c\delta^2 - (f(x) - f(x + \delta d))$. After we substitute these values of C and μ into (2.10), we obtain the following proposition.

Proposition 2.1 *When evaluating the sufficient decrease condition (2.1) in a stochastic DFO algorithm using Test 2.1, if $Y = c\delta^2 - (F(x, \xi^x) - F(x + \delta d, \xi^d))$ is Gaussian with variance σ^2 and $a_l = -b_l = \sigma^2/(2es\delta^2)$, then one needs an expected sample size approximately bounded by*

$$\frac{\sigma^2}{4e^2s^2}\delta^{-4}\min\left(1, \frac{es\delta^2}{|c\delta^2 - (f(x) - f(x + \delta d))|}\right) \tag{2.11}$$

so that Test 2.1 is $s\delta^2$-accurate.

Proposition 2.1 gives us a novel sample size result, which explicitly depends on the potential decrease $f(x) - f(x + \delta d)$. Here we offer some interpretation and insight on the novel sample size quantity (2.11). Since $\min(a, b)$ is smaller than both the first term a and the second term b, it is clear from the first term that this quantity is $O(\delta^{-4})$. Therefore, the sample size in Test 2.1 is at least not larger than in Test 2.2 with respect to the power of δ. From the second term, the quantity (2.11) is $O(\delta^{-2}/|c\delta^2 - (f(x) - f(x + \delta d))|)$. In fact, when the potential decrease $f(x) - f(x + \delta d)$ is $\Theta(\delta)$, the quantity (2.11) drops to $O(\delta^{-3})$. Combining both terms, on the one hand, the quantity (2.11) achieves its maximum when $f(x) - f(x + \delta d)$ is close to $c\delta^2$, where it becomes difficult to distinguish the true hypothesis. On the other hand, Test 2.1 stops early and this quantity decreases significantly whenever $f(x) - f(x + \delta d)$ is far from $c\delta^2$, in which case the ratio $es\delta^2/|c\delta^2 - (f(x) - f(x + \delta d))|$ becomes small. This happens frequently when the algorithm is far from stationarity and $f(x) - f(x + \delta d)$ is $\Theta(\delta)$. Such an early termination effect and sample size advantage are commonly seen in the sequential hypothesis test theory [25, 27].

3 Convergence rate of probabilistic-descent direct search

In this section, to handle the stochastic setting, we propose a version of probabilistic-descent direct search in which its sufficient decrease condition is tested in a sequential hypothesis testing framework, through a sequential hypothesis test that ends properly and is C-accurate, where $C > 0$ is given at the beginning of each algorithm iteration. Before we introduce the algorithm, we formally state the assumption that such a sequential hypothesis test can be developed for any given $C > 0$.

Assumption 3.1 *For any $C > 0$, there exists a C-accurate sequential hypothesis test for Problem 2.1 that ends properly.*

For the rest of this section, we will assume that Assumption 3.1 is satisfied. Probabilistic-descent direct search in stochastic setting is stated in Algorithm 1. The value of C is chosen as a multiple of Δ_k^2, where Δ_k is the stepsize, so C varies in each iteration and becomes small when Δ_k does. A C_k-accurate sequential hypothesis test will determine whether a candidate point is accepted (and the stepsize increased) or rejected (and the stepsize decreased).

Algorithm 1 Probabilistic-descent direct search based on sequential hypothesis testing

```
1: Initialization. Choose an initial point $x_0$, an initial stepsize $\delta_0$, $c > 0$, $\theta \in (0, 1)$, $\gamma \in (1, \infty)$.
2: for $k = 0, 1, \cdots$ do
3:   Uniformly select a random direction $D_k$ from the unit sphere.
4:   The candidate point is then $X_k + \Delta_k D_k$. Perform a $c\Delta_k^2(1-\theta^2)/2(\gamma^2-\theta^2)$-accurate
     sequential hypothesis test for Problem 2.1 with the random variable $Y_k = c\Delta_k^2 - (F(X_k, \xi^x) - F(X_k + \Delta_k D_k, \xi^d))$ (which requires estimating $\sigma$). The result of this test declares either $H_0$
     or $H_1$ accepted.
5:   if $H_0$ is accepted, then
6:     The candidate point is accepted. Set $X_{k+1} = X_k + \Delta_k D_k$ and $\Delta_{k+1} = \gamma\Delta_k$.
7:   else
8:     The candidate point is rejected. Set $X_{k+1} = X_k$ and $\Delta_{k+1} = \theta\Delta_k$.
9:   end if
10: end for
```

We need to introduce some notation to develop a convergence rate for Algorithm 1. Let us denote the decision result of the sequential hypothesis test at iteration k by

$$A_k = \mathbf{1}\{H_0 \text{ is accepted at iteration } k\}.$$

In Algorithm 1, the directions D_k and the decisions A_k are random. Denote the probability space of Algorithm 1 by $(\Omega, \mathcal{F}, P)$. As a consequence of the randomness of D_{k-1} and A_{k-1}, the current point X_k and the stepsize Δ_k are also random quantities. Let $\Phi_k = f(X_k) - f^* + \eta\Delta_k^2$, where $\eta > 0$ is a real number. To formalize conditioning on the past, let $\mathcal{F}_k$ denote the σ-algebra generated by $D_0, \ldots, D_{k-1}$ and $A_0, \ldots, A_{k-1}$ and let $\mathcal{F}_{k+1/2}$ denote the σ-algebra generated by $D_0, \ldots, D_k$ and $A_0, \ldots, A_{k-1}$. Under this definition of $\mathcal{F}_k$ and $\mathcal{F}_{k+1/2}$, we have the following conclusions: Δ_k, X_k, and Φ_k are $\mathcal{F}_k$-measurable; Δ_k, X_k, Φ_k, and D_k are $\mathcal{F}_{k+1/2}$-measurable; Δ_k, X_k, Φ_k, D_k, and A_k are $\mathcal{F}_{k+1}$-measurable.

To analyze Algorithm 1, we start by noting that the unbiased noise Assumption (1.1) implies

$$E[Y_k|\mathcal{F}_{k+1/2}] = c\Delta_k^2 - (f(X_k) - f(X_k + \Delta_k D_k)), \tag{3.1}$$

where Y_k was defined in Step 4 of Algorithm 1. Denote $S_k^1 = \{\omega : f(X_k) - f(X_k + \Delta_k D_k) \geq c\Delta_k^2\}$ and $S_k^2 = \{\omega : f(X_k) - f(X_k + \Delta_k D_k) < c\Delta_k^2\}$, which are two disjoint $\mathcal{F}_{k+1/2}$-measurable events. It follows from (3.1) that $S_k^1 = \{\omega : E[Y_k] \leq 0\}$ and $S_k^2 = \{\omega : E[Y_k] > 0\}$. We can then rewrite (2.5) and (2.6) in the context of Algorithm 1 in the following way

$$P(A_k = 0|\mathcal{F}_{k+1/2}, S_k^1) \leq \frac{1}{2} \tag{3.2}$$

$$P(A_k = 1|\mathcal{F}_{k+1/2}, S_k^2) \leq \frac{c\Delta_k^2(1-\theta^2)/2(\gamma^2-\theta^2)}{c\Delta_k^2 - (f(X) - f(X_k + \Delta_k D_k))}. \tag{3.3}$$

Note that $E[Y_k]$ plays here the role of μ in (2.5) and (2.6).

Our main goal is to bound the number of iterations T_ϵ defined as follows

$$T_\epsilon = \inf\{k \geq 0 : \|\nabla f(X_k)\| \leq \epsilon\}. \tag{3.4}$$

The first lemma tells us that the expected value decrease of Φ_k is bounded below by $\nu\Delta_k^2$, where $\nu > 0$ is a real number.

Lemma 3.1 *Let Assumption 3.1 hold and $\Phi_k = f(X_k) - f^* + \eta\Delta_k^2$. Then there exist $\eta = \frac{c}{\gamma^2-\theta^2} > 0$ and $\nu = \frac{c}{2}\frac{1-\theta^2}{\gamma^2-\theta^2}$ such that for all k we have*

$$E[\Phi_k - \Phi_{k+1}|\mathcal{F}_k] \geq \nu\Delta_k^2.$$

Proof. We start by proving $E[\Phi_k - \Phi_{k+1}|\mathcal{F}_{k+1/2}] \geq \nu\Delta_k^2$. Note that Δ_k, X_k, and D_k are $\mathcal{F}_{k+1/2}$-measurable. We express $E[\Phi_k - \Phi_{k+1}|\mathcal{F}_{k+1/2}]$ in terms of $P(A_k = 1|\mathcal{F}_{k+1/2})$ as follows

$$\begin{aligned}
E[\Phi_k - \Phi_{k+1}|\mathcal{F}_{k+1/2}] &= E[f(X_k) + \eta\Delta_k^2 - f(X_{k+1}) - \eta\Delta_{k+1}^2|\mathcal{F}_{k+1/2}] \\
&= E[\big(f(X_k) + \eta\Delta_k^2 - f(X_{k+1}) - \eta\Delta_{k+1}^2\big)\,\mathbf{1}(A_k = 1)|\mathcal{F}_{k+1/2}] \\
&\quad + E[\big(f(X_k) + \eta\Delta_k^2 - f(X_{k+1}) - \eta\Delta_{k+1}^2\big)\,\mathbf{1}(A_k = 0)|\mathcal{F}_{k+1/2}] \\
&= (f(X_k) - f(X_k + \Delta_k D_k) + \eta(1-\gamma^2)\Delta_k^2)P(A_k = 1|\mathcal{F}_{k+1/2}) \\
&\quad + \eta(1-\theta^2)\Delta_k^2 P(A_k = 0|\mathcal{F}_{k+1/2}).
\end{aligned}$$

Since $P(A_k = 0|\mathcal{F}_{k+1/2}) = 1 - P(A_k = 1|\mathcal{F}_{k+1/2})$, we have

$$\begin{aligned}
E[\Phi_k - \Phi_{k+1}|\mathcal{F}_{k+1/2}] &= (f(X_k) - f(X_k + \Delta_k D_k) + \eta(\theta^2-\gamma^2)\Delta_k^2)P(A_k = 1|\mathcal{F}_{k+1/2}) \\
&\quad + \eta(1-\theta^2)\Delta_k^2 \\
&= (f(X_k) - f(X_k + \Delta_k D_k) - c\Delta_k^2)P(A_k = 1|\mathcal{F}_{k+1/2}) \\
&\quad + c\frac{1-\theta^2}{\gamma^2-\theta^2}\Delta_k^2.
\end{aligned}$$

Note that $S_k^1 = \{\omega : f(X_k) - f(X_k + \Delta_k D_k) \geq c\Delta_k^2\}$ and $S_k^2 = \{\omega : f(X_k) - f(X_k + \Delta_k D_k) < c\Delta_k^2\}$ are two disjoint $\mathcal{F}_{k+1/2}$-measurable events such that $S_k^1 \cup S_k^2 = \Omega$. We prove $E[\Phi_k - \Phi_{k+1}|\mathcal{F}_{k+1/2}] \geq \nu\Delta_k^2$ for these two scenarios separately.

Case 1. First we consider the case $S_k^1 = \{\omega : f(X_k) - f(X_k + \Delta_k D_k) \geq c\Delta_k^2\}$ and it follows from $P(A_k = 1|\mathcal{F}_{k+1/2}) \geq 0$ that

$$E[\Phi_k - \Phi_{k+1}|\mathcal{F}_{k+1/2}, S_k^1] \geq c\frac{1-\theta^2}{\gamma^2-\theta^2}\Delta_k^2 \geq \nu\Delta_k^2.$$

Case 2. Otherwise, we consider $S_k^2 = \{\omega : f(X_k) - f(X_k + \Delta_k D_k) < c\Delta_k^2\}$. Then (3.3) guarantees

$$P(A_k = 1|\mathcal{F}_{k+1/2}, S_k^2) \leq \frac{c\Delta_k^2(1-\theta^2)/2(\gamma^2-\theta^2)}{c\Delta_k^2 - (f(X_k) - f(X_k + \Delta_k D_k))}.$$

It follows that

$$\begin{aligned}
E[\Phi_k - \Phi_{k+1}|\mathcal{F}_{k+1/2}, S_k^2] =&(f(X_k) - f(X_k + \Delta_k D_k) - c\Delta_k^2)P(A_k = 1|\mathcal{F}_{k+1/2}, S_k^2) + c\frac{1-\theta^2}{\gamma^2-\theta^2}\Delta_k^2 \\
\geq&(c\frac{1-\theta^2}{\gamma^2-\theta^2} - \frac{c}{2}\frac{1-\theta^2}{\gamma^2-\theta^2})\Delta_k^2 = \nu\Delta_k^2.
\end{aligned}$$

Since S_k^1 and S_k^2 partition Ω and are $\mathcal{F}_{k+1/2}$-measurable, we conclude that

$$E[\Phi_k - \Phi_{k+1} | \mathcal{F}_{k+1/2}] \geq \nu \Delta_k^2.$$

Note that Δ_k is $\mathcal{F}_k$-measurable. It follows from the tower property of conditional expectation that

$$\begin{aligned} E[\Phi_k - \Phi_{k+1} | \mathcal{F}_k] &= E[E[\Phi_k - \Phi_{k+1} | \mathcal{F}_{k+1/2}] | \mathcal{F}_k] \\ &\geq E[\nu \Delta_k^2 | \mathcal{F}_k] = \nu \Delta_k^2. \end{aligned}$$

□

Now we need to guarantee that the directions used in Algorithm 1 are of descent type with a sufficiently large probability. To do so, we need to generate a direction for which

$$\kappa_k = \frac{-\nabla f(X_k)^\top D_k}{\|\nabla f(X_k)\| \|D_k\|}$$

is sufficiently large. Note that the denominator of κ_k will not be zero because we will assume that $k < T_\epsilon$, where T_ϵ was defined in (3.4). From [15, Lemma B.2], it turns out that if $\mathcal{D}_k = \{D_k\}$ is uniformly generated from a unit sphere, the direction D_k will be enough descent ($\kappa_k \geq 1/(7\sqrt{n})$) for a sufficiently large probability.

Lemma 3.2 *For $k < T_\epsilon$, it holds for $\tau \in [0, \sqrt{n}]$ that*

$$P(\kappa_k \geq \frac{\tau}{\sqrt{n}} | \mathcal{F}_k) \geq \frac{1}{2} - \frac{\tau}{\sqrt{2\pi}}.$$

For simplicity we select $\tau = \frac{1}{7}$ and it holds for $k < T_\epsilon$ that

$$P(\kappa_k \geq \frac{1}{7\sqrt{n}} | \mathcal{F}_k) \geq \frac{3}{7}.$$

Proof. See [15, Lemma B.2], where the result holds for any k. □

The next step in the convergence theory is to ensure that for each $k < T_\epsilon$, if D_k is an at least $1/(7\sqrt{n})$-descent direction and the stepsize Δ_k is smaller than a certain δ_ϵ, then the sufficient decrease condition is satisfied at iteration k. To do so, we need another assumption on the objective function, which is common in the DFO literature.

Assumption 3.2 *Suppose that the objective function f is bounded from below on $\mathbb{R}^n$. Suppose that f is smooth and its gradient ∇f is Lipschitz continuous with constant L_f.*

The next lemma is a consequence of assuming that f is smooth and its gradient ∇f is L_f-Lipschitz continuous.

Lemma 3.3 *Let Assumption 3.2 hold. For all $k < T_\epsilon$, if $\kappa_k \geq \frac{1}{7\sqrt{n}}$ and $\Delta_k \leq \delta_\epsilon = \frac{2}{7L_f + 14c} \frac{\epsilon}{\sqrt{n}}$, then*

$$f(X_k) - f(X_k + \Delta_k D_k) \geq c \Delta_k^2. \tag{3.5}$$

Proof. It is known (see, for example, [22, Lemma 1.2.3]) that Assumption 3.2 implies for any x, y from $\mathbb{R}^n$ that

$$f(x) - f(y) \geq \nabla f(x)^T(x - y) - \frac{L_f}{2}\|x - y\|^2.$$

After substituting $x = X_k$ and $y = X_k + \Delta_k D_k$, and using $\|D_k\| = 1$, we obtain

$$\begin{aligned} f(X_k) - f(X_k + \Delta_k D_k) &\geq -\nabla f(X_k)^\top(\Delta_k D_k) - \frac{L_f}{2}\Delta_k^2 \\ &= \kappa_k\|\nabla f(X_k)\|\Delta_k - \frac{L_f}{2}\Delta_k^2. \end{aligned} \tag{3.6}$$

It then follows from $\kappa_k \geq \frac{1}{7\sqrt{n}}$, $\|\nabla f(X_k)\| > \epsilon$, $\Delta_k \leq \delta_\epsilon$, and the definition of δ_ϵ that

$$\kappa_k\|\nabla f(X_k)\|\Delta_k \geq \frac{\epsilon}{7\sqrt{n}}\Delta_k \geq c\Delta_k^2 + \frac{L_f}{2}\Delta_k^2. \tag{3.7}$$

Combining (3.6) with (3.7) gives us (3.5). □

Lemma 3.3 gives a sufficient condition for the sufficient decrease condition to be satisfied. The next lemma gives a constant lower bound on the probability of accepting H_0 when the stepsize Δ_k is smaller than δ_ϵ. The proof relies on both the probability of having an at least $1/(7\sqrt{n})$-descent direction (Lemma 3.2) and the conditional probability of correctly identifying sufficient decrease (given by condition (3.2)).

Lemma 3.4 *Let Assumptions 3.1–3.2 hold. For any $\epsilon > 0$ and $k < T_\epsilon$, it holds*

$$P(A_k = 1|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\}) \geq \frac{3}{14}. \tag{3.8}$$

Proof. Note that Δ_k and X_k are $\mathcal{F}_k$-measurable. It follows from event inclusion that

$$P(A_k = 1|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\}) \ \geq \ P(\{A_k = 1\} \cap \{\kappa_k \geq \frac{1}{7\sqrt{n}}\}|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\}).$$

Then it follows from conditional probability formula $P(E_1E_2|E_3) = P(E_1|E_2E_3)P(E_2|E_3)$ that

$$\begin{aligned} P(A_k = 1|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\}) \ &\geq \ P(\{A_k = 1\} \cap \{\kappa_k \geq \frac{1}{7\sqrt{n}}\}|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\}) \\ &= P(A_k = 1|\mathcal{F}_k, \{\kappa_k \geq \frac{1}{7\sqrt{n}}\} \cap \{\Delta_k \leq \delta_\epsilon\}) \cdot P(\kappa_k \geq \frac{1}{7\sqrt{n}}|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\}). \end{aligned} \tag{3.9}$$

From Lemma 3.2, since Δ_k is $\mathcal{F}_k$-measurable, we know that

$$P(\kappa_k \geq \frac{1}{7\sqrt{n}}|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\}) \geq \frac{3}{7}. \tag{3.10}$$

From Lemma 3.3 we have

$$\left\{\{\kappa_k \geq \frac{1}{7\sqrt{n}}\} \cap \{\Delta_k \leq \delta_\epsilon\}\right\} \subseteq \{f(X_k) - f(X_k + \Delta_k D_k) \geq c\Delta_k^2\} = S_k^1. \tag{3.11}$$

From (3.2) we have

$$P(A_k = 1|\mathcal{F}_{k+1/2}, S_k^1) > \frac{1}{2}. \tag{3.12}$$

Together with the knowledge that both events in (3.11) above are $\mathcal{F}_{k+1/2}$-measurable, (3.12) implies

$$P(A_k = 1|\mathcal{F}_{k+1/2}, \{\kappa_k \geq \frac{1}{7\sqrt{n}}\} \cap \{\Delta_k \leq \delta_\epsilon\}) > \frac{1}{2}.$$

Using the tower property of conditional expectation, it follows that

$$\begin{aligned} &P(A_k = 1|\mathcal{F}_k, \{\kappa_k \geq \frac{1}{7\sqrt{n}}\} \cap \{\Delta_k \leq \delta_\epsilon\}) \\ &\quad = E[P(A_k = 1|\mathcal{F}_{k+1/2}, \{\kappa_k \geq \frac{1}{7\sqrt{n}}\} \cap \{\Delta_k \leq \delta_\epsilon\})|\mathcal{F}_k, \{\kappa_k \geq \frac{1}{7\sqrt{n}}\} \cap \{\Delta_k \leq \delta_\epsilon\}] \\ &\quad > E[\frac{1}{2}|\mathcal{F}_k, \{\kappa_k \geq \frac{1}{7\sqrt{n}}\} \cap \{\Delta_k \leq \delta_\epsilon\}] \\ &\quad = \frac{1}{2}. \end{aligned} \tag{3.13}$$

Applying (3.10) and (3.13) to (3.9) yields (3.8). □

The rest of the rate derivation makes use of the result [5, Theorem 2] which is rederived in the Appendix for a general probability $p > 1/2$ (see Theorem B.3). This result requires defining a submartingale W_k to model the behavior of the stepsize, which we now do in the context of Algorithm 1. Denote $W_0 = 0$. For $k \geq 0$, let W_{k+1} be a Bernoulli random variable taking values $\log \gamma$ or $\log \theta$ with probabilities described next. For $k \geq T_\epsilon$, the probabilities are

$$\begin{aligned} P(W_{k+1} = \log \gamma|\mathcal{F}_k) &= \frac{3}{14} \\ P(W_{k+1} = \log \theta|\mathcal{F}_k) &= \frac{11}{14}. \end{aligned}$$

For $k < T_\epsilon$, when $\Delta_k > \delta_\epsilon$, the probabilities are

$$\begin{aligned} P(W_{k+1} = \log \gamma|\mathcal{F}_k, \{\Delta_k > \delta_\epsilon\}) &= \frac{3}{14} \\ P(W_{k+1} = \log \theta|\mathcal{F}_k, \{\Delta_k > \delta_\epsilon\}) &= \frac{11}{14}. \end{aligned}$$

For $k < T_\epsilon$, when $\Delta_k \leq \delta_\epsilon$ and $A_k = 0$, we define $W_{k+1} = \log \theta$ with probability one

$$\begin{aligned} P(W_{k+1} = \log \gamma|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\} \cap \{A_k = 0\}) &= 0 \qquad (3.14) \\ P(W_{k+1} = \log \theta|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\} \cap \{A_k = 0\}) &= 1. \end{aligned}$$

For $k < T_\epsilon$, when $\Delta_k \leq \delta_\epsilon$ and $A_k = 1$, the probabilities are

$$\begin{aligned} P(W_{k+1} = \log \gamma|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\} \cap \{A_k = 1\}) &= \frac{3}{14} \frac{1}{P(A_k = 1|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\})} \qquad (3.15) \\ P(W_{k+1} = \log \theta|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\} \cap \{A_k = 1\}) &= 1 - \frac{3}{14} \frac{1}{P(A_k = 1|\mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\})}. \end{aligned}$$

Note that these last two probabilities are well defined because of Lemma 3.4. It then follows from (3.14), (3.15), and the formula of total probability that for $k < T_\epsilon$

$$\begin{aligned} P(W_{k+1} = \log \gamma | \mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\}) &= \frac{3}{14} \\ P(W_{k+1} = \log \theta | \mathcal{F}_k, \{\Delta_k \leq \delta_\epsilon\}) &= \frac{11}{14}. \end{aligned}$$

Then we can conclude that for any $k > 0$

$$\begin{aligned} P(W_{k+1} = \log \gamma | \mathcal{F}_k) &= \frac{3}{14} \\ P(W_{k+1} = \log \theta | \mathcal{F}_k) &= \frac{11}{14}. \end{aligned}$$

To be able to apply Proposition B.3 in Appendix B, the next step is to verify Assumptions B.1–B.2. The validity of Assumption B.2 results from Lemma 3.1, where $\nu\Delta_k^2$ plays the role of $h(\Delta_k)$ in Assumption B.2. To complete the convergence theory, it remains to show the validity of Assumption B.1, which provides a lower bound for W_{k+1} when $k < T_\epsilon$, by ensuring that if the stepsize Δ_{k+1} is smaller than a threshold (in our context this threshold is $\Delta_\epsilon = \delta_\epsilon\theta$), then the stepsize Δ_{k+1} must be no less than $\Delta_k \exp(W_{k+1})$.

Lemma 3.5 *One has for all k*

$$\mathbf{1}(k < T_\epsilon)\Delta_{k+1} \geq \mathbf{1}(k < T_\epsilon) \min\left(\Delta_k e^{W_{k+1}}, \delta_\epsilon\theta\right). \tag{3.16}$$

Proof. Note that it follows from Algorithm 1 and the definition of A_k that $A_k = 0$ if and only if $\Delta_{k+1}/\Delta_k = \theta$. When $k < T_\epsilon$ and $\Delta_k \leq \delta_\epsilon$, since we have set $W_{k+1} = \log\theta$ with probability one when $A_k = 0$, it turns out that, with probability one, whenever the value of $\log(\Delta_{k+1}/\Delta_k)$ is $\log\theta$, one has $W_{k+1} = \log\theta$. When $A_k = 1$, $\Delta_{k+1}/\Delta_k = \gamma$, and $k < T_\epsilon$, W_{k+1} can take the values $\log\gamma$ and $\log\theta$. Then we conclude that when $k < T_\epsilon$ and $\Delta_k \leq \delta_\epsilon$, one has

$$\log(\Delta_{k+1}/\Delta_k) \geq W_{k+1}. \tag{3.17}$$

Note that $\Delta_{k+1} < \delta_\epsilon\theta$ implies $\Delta_k < \delta_\epsilon$, which in turn implies $\Delta_{k+1} \geq \Delta_k e^{W_{k+1}}$ when $k < T_\epsilon$ because of (3.17). We have covered all possible cases, and the proof is concluded. □

We are now ready to apply Theorem B.3 to Algorithm 1 with $\Phi_0 = f(x_0) - f^* + \eta\delta_0^2$.

Theorem 3.6 *Let Assumptions 3.1–3.2 hold. Let θ and γ be chosen such that $3\log\gamma + 11\log\theta > 0$. Then*

$$E[T_\epsilon] \leq 1 + \frac{14\log\gamma}{3\log\gamma + 11\log\theta} \frac{(\gamma^2 - \theta^2)(f(X_0) - f^*) + c\Delta_0^2}{\frac{c}{2}(1-\theta^2)\theta^2} \frac{(7L_f + 14c)^2}{4} \frac{n}{\epsilon^2}.$$

Theorem 3.6 ensures an expected worst-case complexity bound of $O(n/\epsilon^2)$ for Algorithm 1.

4 A numerical experiment

In this section, we will report the numerical performance of Test 2.1 against Test 2.2 in an experiment running probabilistic-descent direct search under Gaussian noise. Recall from Section 2.3 that Test 2.1 is a sequential hypothesis test and Test 2.2 is a fixed sample test. Specifically, we ran Algorithm 1 with Test 2.1 or Test 2.2 to solve its hypothesis test problem, given a total budget of 10000 sample evaluations.

4.1 Experiment setup

We used 38 problems suggested in [13] from the CUTEst dataset [14, 16], which exhibit different features in terms of non-linearity, non-convexity, and partial separability. For each problem, we generated problem instances with different dimensions, which gives us a problem set of 91 problem instances in total. Please see Table 4.1 for the problem names and corresponding dimensions of our problem set. To inject noise in the objective functions, we selected the noise term $F(X, \xi) - f(x)$ to follow an independent identically distributed Gaussian distribution with variance $\sigma^2 \in \{0.01, 1\}$. Our choice of additive Gaussian noise is consistent with the assumption of Proposition 2.1. Other choices are certainly possible, and we discuss them in Section 5. For performance evaluation, we used data profiles [20] and performance profiles [10], selecting the tolerance parameter, which defines a problem being solved, as $\tau = 0.1$. Each problem instance was ran 10 times in the experiment and considered as 10 problem instances in all the profiles.

Name	Dimension	Name	Dimension
ARGLINA	10, 50, 100	ARGTRIGLS	10, 50, 100
ARWHEAD	100	BDEXP	100
BOXPOWER	10, 100	BROWNAL	10, 100
COSINE	10, 100	CURLY10	100
DIXON3DQ	10, 100	DQRTIC	10, 50, 100
ENGVAL1	2, 50, 100	EXTROSNB	5, 10, 100
FLETBV3M	10, 100	FLETCBV3	10, 100
FLETCHBV	10, 100	FLETCHCR	10, 100
FREUROTH	2, 10, 50, 100	INDEFM	10, 50, 100
MANCINO	10, 20, 30, 50, 100	MOREBV	10, 50, 100
NONCVXU2	10, 100	NONCVXUN	10, 100
NONDIA	10, 50, 100	NONDQUAR	100
PENALTY2	10, 50, 100	POWER	10, 50, 100
QING	100	QUARTC	25, 100
SENSORS	10, 100	SINQUAD	5, 50, 100
SCURLY10	10, 100	SCURLY20	100
SPARSINE	10, 50, 100	SPARSQUR	10, 50, 100
SSBRYBND	10, 50, 100	TRIDIA	10, 50, 100
TRIGON1	10, 100	TOINTGSS	10, 50, 100

Table 4.1: Names and corresponding dimensions of the 91 CUTEst problem instances in the problem set.

We now specify the parameters of Algorithm 1. We used the initial point provided by CUTEst dataset, chose the initial stepsize $\delta_0 = 1$, and selected $c = 0.5$. However, the choice of $\theta \in (0, 1)$ and $\gamma \in (1, \infty)$ can greatly affect the performance of the algorithm. Note that it is required from Theorem 3.6 that $3 \log \gamma + 11 \log \theta > 0$, otherwise, the stepsize δ of Algorithm 1 may still converge to 0 even when the iterates approach a non-stationary point. We also notice that the sample size per iteration increases rapidly as the stepsize δ decreases to 0. This phenomenon occurs for both Test 2.1 and Test 2.2, and is more severe in Test 2.2 with a sample size of $\Theta(\delta^{-4})$. For the purposes of both algorithm performance and a fair comparison, we want to select a large θ so that the stepsize may stay away from 0 unless necessary. Based on the above observations, we selected $\theta = 0.95$ and $\gamma = 1.3$ in our numerical experiment when using either Test 2.1 or Test 2.2. We tried different θ and γ when the algorithm uses either tests and observed no significant changes in the relative performance of the methods.

To choose the parameters in Test 2.1 and Test 2.2, let us suppose that at an iteration k of Algorithm 1, σ_k^2 is the variance of Y_k, or at least a known upper bound thereof. The number $C_k = c\Delta_k^2(1-\theta^2)/2(\gamma^2-\theta^2)$ in Algorithm 1 is known at each iteration k. For Test 2.2, we selected the fixed sample size $m = \sigma_k^2 C_k^{-2}$. For Test 2.1, we selected the test lower and upper bounds $a_l = -b_l = \sigma_k^2/(2eC_k)$.

4.2 Experiment results

The results are reported in Figures 4.1 and 4.2 for the two noise variances. We tried larger budgets, different values of τ in the profiles, and smaller variances, but we observed no significant changes in the relative performance of the methods. For both cases, it can be clearly seen that using the sequential test (Test 2.1) outperforms using the fixed sample test (Test 2.2). The performance gap increases as the noise variance increases.

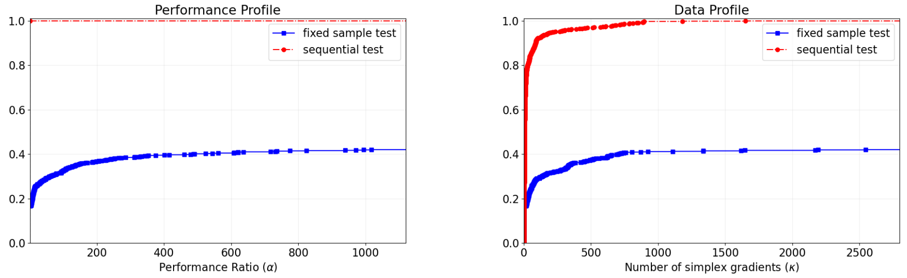

Figure 4.1: Performance and data profiles for Algorithm 1 using sequential test (Test 2.1) or fixed sample test (Test 2.2) for noise variance $\sigma^2 = 1$ for the problem set of Table 4.1.

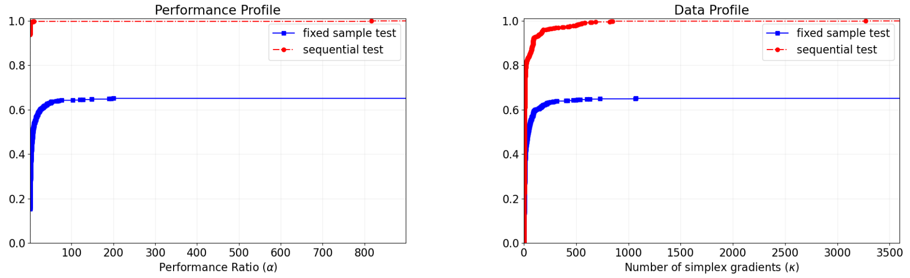

Figure 4.2: Performance and data profiles for Algorithm 1 using sequential test (Test 2.1) or fixed sample test (Test 2.2) for noise variance $\sigma^2 = 0.01$ for the problem set of Table 4.1.

5 Concluding remarks

In this paper, we introduced sequential hypothesis testing for solving a stochastic DFO problem. Specifically, we formulated the evaluation of the sufficient decrease condition (2.1) as a hypothesis test problem (Problem 2.1) and solved it through a sequential hypothesis test (Test 2.1).

Given an additive Gaussian noise assumption, we estimated the sample size of Test 2.1 in Proposition 2.1, which indicated a possible early termination under a reduced sample size. In particular, when the potential decrease $f(x) - f(x + \delta d)$ is $\Theta(\delta^r)$ for some $r \in (0, 2]$, we showed that the expected sample size can decrease from the literature's $\Theta(\delta^{-4})$ to $O(\delta^{-2-r})$. We applied the sequential test framework to probabilistic-descent direct search and derived an iteration complexity of $O(n\epsilon^{-2})$ in Theorem 3.6. Our numerical results showed that the use of sequential hypothesis test (Test 2.1) significantly outperforms the use of its fixed sample counterpart.

The sampling procedure in the form of Test 2.2 is widely used in many stochastic DFO algorithms. It is referred to in the literature of statistics as a hypothesis test with a fixed sample size. In contrast, a sequential hypothesis test uses a random sample size, which can be regarded as a relaxation of a fixed sample size, to allow early termination of the test when the hypothesis test problem is not difficult (in our case, when the mean μ of Y in (2.2) is far from 0). From this point of view, a test with a fixed sample size is some form of a restricted sequential hypothesis test. Therefore, sequential tests are generally expected to use less samples than their fixed sample size counterparts for equally good performance.

We would like to make two observations regarding the noise setting considered in this paper. The first one is that, in our numerical results in Section 4, we only show that Test 2.1 performs well in a relatively limited setting where the noise is additive Gaussian. However, we can still use Test 2.1 for other noise types as long as an upper bound of the noise variance of Y is known. Regardless of the distribution of Y, the random walk $\sum_i Y^i$ is typically approximated by a Brownian motion process in the study of sequential analysis [25, Chapter 3.1]. Therefore, we believe that Test 2.1 will still work well for other noises beyond Gaussian. The other observation is that, in this paper, we assumed knowledge of the noise variance or at least an upper bound of it. For a more practical algorithmic implementation, we may need to estimate the noise variance using some estimation techniques (see, for instance, [4, 21]).

Acknowledgments

This work is partially supported by the U.S. Air Force Office of Scientific Research (AFOSR) award FA9550-23-1-0217 and the U.S. Office of Naval Research (ONR) award N000142412656. We thank Damiano Zeffiro and Trang H. Tran for insights and discussions about sequential sampling.

References

[1] J. Achddou, O. Cappe, and A. Garivier, *Stochastic direct search method for blind resource allocation*, Transactions on Machine Learning Research Journal, (2024).

[2] C. Audet, K. J. Dzahini, M. Kokkolaras, and S. Le Digabel, *Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates*, Computational Optimization and Applications, 79 (2021), pp. 1–34.

[3] A. S. Bandeira, K. Scheinberg, and L. N. Vicente, *Convergence of trust-region methods based on probabilistic models*, SIAM Journal on Optimization, 24 (2014), pp. 1238–1264.

[4] A. S. Berahas, R. H. Byrd, and J. Nocedal, *Derivative-free optimization of noisy functions via quasi-Newton methods*, SIAM Journal on Optimization, 29 (2019), pp. 965–993.

[5] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg, *Convergence rate analysis of a stochastic trust-region method via supermartingales*, INFORMS Journal on Optimization, 1 (2019), pp. 92–119.

[6] C. Cartis and K. Scheinberg, *Global convergence rate analysis of unconstrained optimization methods based on probabilistic models*, Mathematical Programming, 169 (2018), pp. 337–375.

[7] R. Chen, M. Menickelly, and K. Scheinberg, *Stochastic optimization using a trust-region method and random models*, Mathematical Programming, 169 (2018), pp. 447–487.

[8] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*, SIAM, Philadelphia, 2009.

[9] H. F. Dodge and H. G. Romig, *A method of sampling inspection*, The Bell System Technical Journal, 8 (1929), pp. 613–631.

[10] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Mathematical Programming, 91 (2002), pp. 201–213.

[11] K. J. Dzahini, *Expected complexity analysis of stochastic direct-search*, Computational Optimization and Applications, 81 (2022), pp. 179–200.

[12] K. J. Dzahini, F. Rinaldi, C. W. Royer, and D. Zeffiro, *Direct-search methods in the year 2025: Theoretical guarantees and algorithmic paradigms*, EURO Journal on Computational Optimization, (2025), p. 100110.

[13] T. Giovannelli, O. Sohab, and L. N. Vicente, *The limitation of neural nets for approximation and optimization*, Journal of Global Optimization, (2024). Published Online.

[14] N. I. Gould, D. Orban, and P. L. Toint, *CUTEst: a constrained and unconstrained testing environment with safe threads for mathematical optimization*, Computational Optimization and Applications, 60 (2015), pp. 545–557.

[15] S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang, *Direct search based on probabilistic descent*, SIAM Journal on Optimization, 25 (2015), pp. 1515–1541.

[16] S. Gratton and P. L. Toint, *S2MPJ and CUTEst optimization problems for matlab, python and julia*, Optimization Methods and Software, 40 (2025), pp. 871–903.

[17] Y. Ha and S. Shashaani, *Iteration complexity and finite-time efficiency of adaptive sampling trust-region methods for stochastic derivative-free optimization*, IISE Transactions, 57 (2025), pp. 541–555.

[18] J. Larson and S. C. Billups, *Stochastic derivative-free optimization using a trust region framework*, Computational Optimization and Applications, 64 (2016), pp. 619–645.

[19] J. Larson, M. Menickelly, and S. M. Wild, *Derivative-free optimization methods*, Acta Numerica, 28 (2019), pp. 287–404.

[20] J. J. Moré and S. M. Wild, *Benchmarking derivative-free optimization algorithms*, SIAM Journal on Optimization, 20 (2009), pp. 172–191.

[21] ——, *Estimating computational noise*, SIAM Journal on Scientific Computing, 33 (2011), pp. 1292–1314.

[22] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87, Springer Science & Business Media, New York, 2013.

[23] F. Rinaldi, L. N. Vicente, and D. Zeffiro, *Stochastic trust-region and direct-search methods: A weak tail bound condition and reduced sample sizing*, SIAM Journal on Optimization, 34 (2024), pp. 2067–2092.

[24] S. Shashaani, F. S. Hashemi, and R. Pasupathy, *ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization*, SIAM Journal on Optimization, 28 (2018), pp. 3145–3176.

[25] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*, Springer Science & Business Media, New York, 2013.

[26] A. Wald, *Sequential tests of statistical hypotheses*, in Breakthroughs in Statistics: Foundations and Basic Theory, Springer, New York, 1992, pp. 256–298.

[27] ——, *Sequential Analysis*, Courier Corporation, 2004.

[28] A. Wald and J. Wolfowitz, *Optimum character of the sequential probability ratio test*, The Annals of Mathematical Statistics, 19 (1948), pp. 326–339.

A Appendix (Auxiliary results)

Proposition A.1 *Let $A > 1$ be a real number. Then $\frac{A^x-1}{x(A^x+1)} \leq \frac{\log A}{2}$ holds for any $x \in \mathbb{R}$.*

Proof. To prove $\frac{A^x-1}{x(A^x+1)} \leq \frac{\log A}{2}$ for any $x \in \mathbb{R}$, since $\frac{A^{-x}-1}{-x(A^{-x}+1)} = \frac{A^x-1}{x(A^x+1)}$ is a symmetric function, it suffices to prove it when $x \geq 0$. We rearrange and rewrite it as follows

$$xA^x \log A + x \log A + 2 - 2A^x \geq 0.$$

The derivatives of $g(x) = xA^x \log A + x \log A + 2 - 2A^x$ are

$$\begin{aligned} g'(x) &= (xA^x \log A + 1 - A^x) \log A \\ g''(x) &= (xA^x \log A)(\log A)^2 \geq 0. \end{aligned}$$

It follows that $g'(x)$ is non-decreasing and $g'(x) \geq g'(0) = 0$. Then $g(x)$ is non-decreasing and $g(x) \geq g(0) = 0$. □

Proposition A.2 *Let $t > 0$ be a real number. Then $\frac{1}{t^x} \leq \frac{1}{x}$ holds for any $x \geq 1$ if and only if $t \geq e^{\frac{1}{e}}$.*

Proof. Since $\frac{1}{t^x} \leq \frac{1}{x}$ holds for $x = 2$, we only need to consider $t \geq \sqrt{2}$. Define $f(x) = t^x - x$. Its derivatives are $f'(x) = t^x \log t - 1$ and $f''(x) = t^x(\log t)^2 > 0$. The solution x^* of $f'(x) = 0$ is $x^* = -\frac{\log\log t}{\log t}$. If $x^* \leq 1$, then $f(x)$ achieves a minimum over $[1, \infty)$ at $x = 1$. We can verify that $f(1) > 0$. If $x^* \geq 1$, then $f(x)$ achieves a minimum over $[1, \infty)$ at x^*. The minimum value is $f(x^*) = \frac{1+\log\log t}{\log t}$. Therefore, we have that $f(x) \geq 0$ holds for any $x \in [1, \infty)$ if and only if $t \geq e^{\frac{1}{e}}$. □

B Appendix (Renewal–Reward Martingale Process)

We begin by describing how our Algorithm 1 can be analyzed through the lens of a renewal-reward process. A renewal event is said to occur within Algorithm 1 whenever the stepsize meets or exceeds a predetermined threshold, denoted by Δ_ϵ. We will show that, after one renewal, the next renewal will happen in a finite time interval, and this expected returning time is constant. Each of these renewal events is associated with a random reward, whose expectation is bounded below by a positive function $h(\Delta_\epsilon)$. Once defined in this way, the total accumulated reward across multiple renewals can be viewed as a submartingale that grows through these random increments. Since the total available reward is bounded, we will be able to deduce an expected stopping time for the algorithm mechanism. This line of analysis, which treats the underlying process as a renewal-reward martingale, was first developed in [5], where it was used to establish the expected global convergence rate of a stochastic trust-region method.

To adapt the framework [5] to our Algorithm 1, we need to remove one of the assumptions from [5, Assumption 1(ii)], which originally stipulated that $p > 1/2$. We also need to show that the main result in [5, Theorem 2] still holds under more general conditions where the sequence W_{k+1} can take any positive and negative values a and b, respectively, instead of ±1 as the authors used in [5].

Consider a stochastic process $\{(\Phi_k, \Delta_k)\}$ defined on some probability space, where Φ_k takes values in the interval $[0, \infty)$ and Δ_k takes values in the interval $(0, \infty)$, for all $k \geq 0$. Let $\{W_k\}$ be another sequence of random variables defined on the same probability space as $\{(\Phi_k, \Delta_k)\}$, initialized by $W_0 = 0$. For all $k \geq 0$, the conditional distribution of W_{k+1}, given the σ-algebra $\mathcal{F}_k$ generated by $\{(\Phi_0, \Delta_0, W_0), \ldots, (\Phi_k, \Delta_k, W_k)\}$, is described by

$$\begin{aligned} P(W_{k+1} = a|\mathcal{F}_k) &= p, \\ P(W_{k+1} = b|\mathcal{F}_k) &= 1 - p, \end{aligned} \tag{B.1}$$

where $a > 0$ and $b < 0$ are two constants and p is the probability of taking the value a. When $a = 1$ and $b = -1$, this construction coincides with the specific case discussed in [5]. From the above definition, it follows that $\{W_k\}$ are mutually independent and W_k is also independent of the sequence $\{(\Phi_j, \Delta_j)\}_{j=0}^{k-1}$ for all k. Lastly, let $\{T_\epsilon\}_{\epsilon>0}$ be a family of stopping times with respect to $\{\mathcal{F}_k\}_{k\geq 0}$, parameterized by some quantity $\epsilon > 0$. As in [5], we impose the following assumptions on $\{(\Phi_k, \Delta_k)\}$ and T_ϵ when $k < T_\epsilon$.

Assumption B.1 *There exists a constant $\Delta_\epsilon > 0$ such that the following holds for all $k \geq 0$*

$$\mathbf{1}(k < T_\epsilon)\Delta_{k+1} \geq \mathbf{1}(k < T_\epsilon)\min(\Delta_k e^{W_{k+1}}, \Delta_\epsilon),$$

where W_{k+1} satisfies $E[W_{k+1}|\mathcal{F}_k] > 0$ (which means $p(a-b)+b > 0$).

Assumption B.2 *There exists a nondecreasing function $h(\cdot) : [0,\infty) \to (0,\infty)$ such that*

$$E(\Phi_k - \Phi_{k+1}|\mathcal{F}_k)\mathbf{1}(k < T_\epsilon) \geq h(\Delta_k)\mathbf{1}(k < T_\epsilon).$$

Assumption B.1 tells us that for $k < T_\epsilon$, the stepsize Δ_k tends to increase and return to the threshold Δ_ϵ when it is smaller than Δ_ϵ. Assumption B.2 tells us that for $k < T_\epsilon$, the stochastic process $\Phi_0 - \Phi_{k+1}$ acts like a submartingale and the expected martingale difference $E(\Phi_k - \Phi_{k+1}|\mathcal{F}_k)$ is at least $h(\Delta_k)$.

In order to define a renewal process, we first define an auxiliary process $\{Z_k\}_{k=0}^\infty$ by letting $Z_0 = \log \frac{\Delta_\epsilon}{\Delta_0}$ and setting

$$Z_{k+1} = \min(Z_k + W_{k+1}, \log \frac{\Delta_\epsilon}{\Delta_0}),$$

or, equivalently,

$$\Delta_0 e^{Z_{k+1}} = \min(\Delta_0 e^{Z_k + W_{k+1}}, \Delta_\epsilon).$$

We then define the renewal process $\{A_n\}_{n=0}^\infty$ by letting $A_0 = 0$ and setting $A_n = \inf\{m > A_{n-1} : Z_m = \log \frac{\Delta_\epsilon}{\Delta_0}\}$. From Assumption B.1, we have that

$$\mathbf{1}(k < T_\epsilon)\Delta_{k+1} \geq \mathbf{1}(k < T_\epsilon)\min(\Delta_k e^{W_{k+1}}, \Delta_\epsilon) \geq \mathbf{1}(k < T_\epsilon)\Delta_0 e^{Z_{k+1}},$$

where we have used a simple inductive argument to obtain the second inequality. The interarrival times of this renewal process are defined for all $n \geq 1$ by

$$\tau_n = A_n - A_{n-1}.$$

The first main step in the analysis will be to bound the expected value of the interarrival time τ_n (see Lemma B.2). For this purpose, one needs to bound $E[\bar{\tau}]$, where $\bar{\tau} = \inf\{n \geq 0 : \bar{Z}_n \geq 0\}$, using the structure of the process W_k (see Lemma B.1 below).

Lemma B.1 *Let Assumption B.1 hold. Define the process $\bar{Z}_0 = b < 0$, $\bar{Z}_{k+1} = \bar{Z}_k + W_{k+1}$ for all $k \geq 0$. Then*

$$E[\bar{\tau}] \leq \frac{a-b}{pa - pb + b}. \tag{B.2}$$

Proof. For ease of notation, let $k \wedge \bar{\tau} = \min\{k, \bar{\tau}\}$ and $v = p(a-b)+b > 0$. Note that

$$E[W_{k+1}|\mathcal{F}_k] = v. \tag{B.3}$$

Consider the stochastic process defined by $R_0 = \bar{Z}_0$ and for $k \geq 1$

$$R_k = \bar{Z}_{k\wedge\bar{\tau}} - \sum_{j=0}^{k\wedge\bar{\tau}-1} v.$$

We first prove that $E[R_{k+1}|\mathcal{F}_k] = R_k$, which means R_k is a martingale with respect to $\{\mathcal{F}_k\}$. To see this, we first note that

$$R_{k+1} - R_k = \bar{Z}_{(k+1)\wedge\bar{\tau}} - \bar{Z}_{k\wedge\bar{\tau}} - \sum_{j=0}^{(k+1)\wedge\bar{\tau}-1} v + \sum_{j=0}^{k\wedge\bar{\tau}-1} v$$

and

$$E[R_{k+1} - R_k|\mathcal{F}_k] = E[(R_{k+1} - R_k)\mathbf{1}(\bar{\tau} > k)|\mathcal{F}_k] + E[(R_{k+1} - R_k)\mathbf{1}(\bar{\tau} \le k)|\mathcal{F}_k]. \tag{B.4}$$

We now show that $E[R_{k+1} - R_k|\mathcal{F}_k] = 0$. First, since $((k \wedge \bar{\tau}) - ((k+1) \wedge \bar{\tau}))\,\mathbf{1}(\bar{\tau} \le k) = 0$, we have

$$\begin{aligned} E[(R_{k+1} - R_k)\mathbf{1}(\bar{\tau} \le k)|\mathcal{F}_k] &= E\left[\left(\bar{Z}_{(k+1)\wedge\bar{\tau}} - \bar{Z}_{k\wedge\bar{\tau}} - \sum_{j=0}^{(k+1)\wedge\bar{\tau}-1} v + \sum_{j=0}^{k\wedge\bar{\tau}-1} v\right)\mathbf{1}(\bar{\tau} \le k)|\mathcal{F}_k\right] \\ &= E\left[0 \cdot \mathbf{1}(\bar{\tau} \le k)|\mathcal{F}_k\right] \\ &= 0. \end{aligned} \tag{B.5}$$

Secondly, from $\bar{Z}_{k+1} = \bar{Z}_k + W_{k+1}$ and (B.3), we have

$$\begin{aligned} E[(R_{k+1} - R_k)\mathbf{1}(\bar{\tau} > k)|\mathcal{F}_k] =&E\left[\left(\bar{Z}_{(k+1)\wedge\bar{\tau}} - \bar{Z}_{k\wedge\bar{\tau}} - \sum_{j=0}^{(k+1)\wedge\bar{\tau}-1} v + \sum_{j=0}^{k\wedge\bar{\tau}-1} v\right)\mathbf{1}(\bar{\tau} > k)|\mathcal{F}_k\right] \\ =&E\left[\left(\bar{Z}_{k+1} - \bar{Z}_k - v\right)\mathbf{1}(\bar{\tau} > k)|\mathcal{F}_k\right] \\ =&E\left[\left(W_{k+1} - v\right)\mathbf{1}(\bar{\tau} > k)|\mathcal{F}_k\right] \\ =&0. \end{aligned} \tag{B.6}$$

After summing up (B.5) and (B.6), since R_k is $\mathcal{F}_k$-measurable, we have from (B.4) that

$$E[R_{k+1}|\mathcal{F}_k] = R_k.$$

Since R_k is a martingale with respect to $\{\mathcal{F}_k\}$, we immediately have $E[R_k] = R_0$. Note that $\bar{Z}_{k\wedge\bar{\tau}} \le a$ for each $k \ge 0$ due to the definition of $\bar{\tau}$ and W_k. We then obtain from the definition of R_k that

$$E\left(\sum_{j=0}^{(k\wedge\bar{\tau})-1} v\right) = E[\bar{Z}_{k\wedge\bar{\tau}}] - E[R_k] \le a - R_0 = a - b. \tag{B.7}$$

Now, due to $v > 0$ and k being eventually larger than $\bar{\tau}$, observe that

$$0 < \sum_{j=0}^{(k\wedge\bar{\tau})-1} v \nearrow \sum_{j=0}^{\bar{\tau}-1} v$$

as $k \to \infty$. Note that this conclusion holds even on the event $\{\bar{\tau} = \infty\}$. Therefore, by the monotone convergence theorem and (B.7),

$$E\left(\sum_{j=0}^{\bar{\tau}-1} v\right) = \lim_{k\to\infty} E\left(\sum_{j=0}^{(k\wedge\bar{\tau})-1} v\right) \le a - b.$$

Finally, using Wald's identity, we have

$$E[\bar{\tau}]v = E\left(\sum_{j=0}^{\bar{\tau}-1} v\right) \leq a - b,$$

which implies (B.2) and concludes the proof. □

We will use the upper bound for the constant $E[\bar{\tau}]$ in the above lemma to give an upper bound for the constant $E[\tau_n]$.

Lemma B.2 *Let Assumption B.1 hold. Let τ_n be defined as before. Then for all n,*

$$E[\tau_n] = p + (1 + E[\bar{\tau}])(1 - p). \tag{B.8}$$

Proof. Note that $E[\tau_n] = E[E[\tau_n | Z_{A_{n-1}}]] = E[E[\tau_1 | Z_0]] = E[\tau_1]$ and it suffices to verify this proposition for $n = 1$.

By conditioning on W_1, we have that

$$E[\tau_1] = 1 \cdot P(W_1 = a) + (1 + E[\bar{\tau}])P(W_1 = b).$$

This identity follows because the distribution of τ_1 conditioned on $Z_1 = \log \frac{\Delta_\epsilon}{\Delta_0} + b$ is the same as the distribution of $\bar{\tau}$. Thus, we simplify this expression to conclude that (B.8) holds. □

The following proposition is proved in [5, Theorem 2] with $E[\tau_n] = p/(2p-1)$. The argument in [5, Theorem 2] works for any constant $E[\tau_n]$ and there is no need to repeat the proof here.

Theorem B.3 *Let Assumptions B.1–B.2 hold. Then*

$$E[T_\epsilon - 1] \leq E[\tau_n] \cdot \frac{\Phi_0}{h(\Delta_\epsilon)},$$

where Φ_0 is a given positive number, $E[\tau_n]$ satisfies (B.2), and h is a given function in Assumption B.2.