General Drift Analysis with Tail Bounds

Per Kristian Lehre
School of Computer Science
University of Nottingham
Nottingham, NG8 1BB
United Kingdom

Carsten Witt
DTU Compute
Technical University of Denmark
2800 Kgs. Lyngby
Denmark

July 22, 2013

Abstract

Drift analysis is one of the state-of-the-art techniques for the runtime analysis of randomized search heuristics. In recent years, many different drift theorems, including additive, multiplicative and variable drift, have been developed, applied and partly generalized or adapted to particular processes. A comprehensive overview article was missing.

We provide not only such an overview but also present a universal drift theorem that generalizes virtually all existing drift theorems found in the literature. On the one hand, the new theorem bounds the expected first hitting time of optimal states in the underlying stochastic process. On the other hand, it also allows for general upper and lower tail bounds on the hitting time, which were not known before except for the special case of upper bounds in multiplicative drift scenarios. As a proof of concept, the new tail bounds are applied to prove very precise sharp-concentration results on the running time of the (1+1) EA on OneMax, general linear functions and LeadingOnes. Moreover, user-friendly specializations of the general drift theorem are given.

Contents

1 Introduction

Runtime analysis is a rather recent and increasingly popular approach in the theory of randomized search heuristics. Typically, the aim is to analyze the (random) time until one goal of optimization (optimum found, good approximation found etc.) is reached. This is equivalent to deriving the first hitting time for a set of states of an underlying (discrete-time) stochastic process.

Drift analysis has turned out as one of the most powerful techniques for runtime analysis. In a nutshell, drift is the expected progress of the underlying process from one time step to another. An expression for the drift is turned into an expected first hitting time via a drift theorem. An appealing property of such a theorem is that a local property (the one-step) drift is translated into a global property (the first hitting time).

Sasak and Hajek (1988) introduced drift analysis to the analysis of randomized search heuristics (more precisely, of simulated annealing), and He and Yao (2001) were the first to apply drift analysis to evolutionary algorithms. The latter paper presents a drift theorem that is nowadays called *additive drift*. Since then, numerous variants of drift theorems have been proposed, including upper and lower bounds in the scenario of *multiplicative drift* (Doerr, Johannsen, and Winzen, 2012; Lehre and Witt, 2012), variable drift (Johannsen, 2010; Mitavskiy, Rowe, and Cannings, 2009; Doerr, Fouz, and Witt, 2011; Rowe and Sudholt, 2012) and generalizations thereof, e. g., variable drift without monotonicity conditions (Doerr, Hota, and Kötzing, 2012; Feldmann and Kötzing, 2013). Moreover, considerable progress was made in the development of so-called distance functions used to model the process analyzed by drift analysis (Doerr and Goldberg, 2013; Witt, 2013). The powerful drift theorems available so far allow for the analysis of randomized search heuristics, in particular evolutionary algorithms and ant colony optimization, on example problems and problems from combinatorial optimization. See also the text books by Auger and Doerr (2011), Neumann and Witt (2010) and Jansen (2013) for detailed expositions of the state of the art in runtime analysis of randomized search heuristics.

At present, the exciting and powerful research done in drift analysis is scattered over the literature. Existing formulations of similar theorems may share many details but deviate in minor conditions. Notation is not always consistent. Several existing variants of drift theorems contain assumptions that might be convenient to formulate, e. g., Markovian properties and discrete or finite search spaces; however, it was not always clear what assumptions were really needed and whether the drift theorem was general enough. This is one reason why additional effort was spent on removing the assumption of discrete search spaces from multiplicative and variable drift theorems (Feldmann and Kötzing, 2013) – an effort, as we will show, was not really required.

Our work makes two main contributions to the area of drift analysis. The first one is represented by a "universal" formulation of a drift theorem that strives for as much generality as possible. We provably can identify all of the existing drift theorems mentioned above as special cases. While doing this, we propose a consistent notation and remove unnecessary assumptions such as discrete search spaces and Markov processes. In fact, we even identify another famous technique for the runtime analysis of randomized search heuristics, namely *fitness levels* (Sudholt, 2013) as a special case of our general theorem. **Caveat.** When we say "all" existing drift theorems, we exclude a specific but important scenario from our considerations. Our paper only considers the case that the drift is directed towards the target of optimization. The opposite case, i. e., scenarios where the process moves away from the target, is covered by the lower bounds from the so-called *simplified/negative drift* theorem (Oliveto and Witt, 2011), which states rather different conditions and implications. The conditions and generality of the latter theorem were scrutinized in a recent erratum (Oliveto and Witt, 2012).

The second contribution is represented by tail bounds, also called deviation bounds or concentration inequalities, on the hitting time. Roughly speaking, conditions are provided under which it is unlikely that the actual hitting time is above or below its expected value by a certain amount. Such tail bounds were not known before in drift analysis, except for the special case of upper tail bounds in multiplicative drift (Doerr and Goldberg, 2013). In particular, our drift theorem is the first to prove lower tails. We use these tail bounds in order to prove very sharp concentration bounds on the running time of a (1+1) EA on OneMax, general linear functions and LeadingOnes. Up to minor details, the following is shown for the running time T of the (1+1) EA on OneMax (and the same holds on all linear functions): the probability that T deviates (from above or below) from its expectation by an additive term of rn is $e^{-\Omega(r)}$ for any constant $r > 0$. With LeadingOnes, a deviation by $rn^{3/2}$ from the expected value is proved to have probability $e^{-\Omega(r)}$. Such sharp-concentration results are extremely useful from a practical point of view since they reveal that the process is "almost deterministic" such that very precise predictions of its actual running time can be made. Moreover, the concentration inequalities allow a change of perspective to tell what progress can be achieved within a certain time budget, see the recent line of work on fixed-budget computations (Jansen and Zarges, 2012; Doerr, Jansen, Witt, and Zarges, 2013).

This paper is structured as follows. Section 2 introduces notation and basics of drift analysis. Section 3 presents the general drift theorem, its proof and suggestions for user-friendly corollaries. Afterwards, specializations are discussed. Section 4 shows how the general drift theorem is related to known variable drift theorems, and Section 5 specializes our general theorem into existing multiplicative drift theorems. The fitness level technique, both for lower and upper bounds, is identified as a special case in Section 6. Section 7 is devoted to the tail bounds contained in the general drift theorem. It is shown how they can directly be applied to prove sharp-concentration results on the running time of the (1+1) EA on OneMax and general linear functions. Moreover, a more user-friendly special case of the theorem with tail bounds is proved and used to show sharp-concentration results w. r. t. LeadingOnes. We finish with some conclusions.

2 Preliminaries

Stochastic process. Throughout this paper, we analyze time-discrete stochastic processes represented by a sequence of non-negative random variables $(X_t)_{t\geq 0}$. For example, X_t could represent the number of zero- or one-bits of an (1+1) EA at generation t, a certain distance value of a population-based EA from an optimal population etc. In particular, X_t might aggregate several different random variables realized by a search heuristic at time t into a single one. We do not care whether the state space is discrete (e. g., all non-negative integers or even a finite subset thereof) or continuous. In discrete search spaces, the random variables X_t will have a discrete support; however, this is not important for the formulation of the forthcoming theorems.

First hitting time. We adopt the convention that the process should pass below some threshold $a \geq 0$ ("minimizes" its state) and define the first hitting time $T_a := \min\{t \mid X_t \leq a\}$. If the actual process seeks to maximize its state, typically a straightforward mapping allows us to stick to the convention of minimization. In a special case, we are interested in the hitting time T_0 of state 0; for example when a (1+1) EA is run on OneMax and were are interested in the first point of time where the number of zero-bits becomes zero. Note that T_a is a stopping time and that we tacitly assume that the stochastic process is adapted to its natural filtration $\mathcal{F}_t := (X_0, \dots, X_t)$, i. e., the information available up to time t.

Drift. The expected one-step change $\delta_t := E(X_t - X_{t+1} \mid \mathcal{F}_t)$ for $t \geq 0$ is called drift. Note that δ_t in general is a random variable since the outcomes of $X_0, \dots, X_t$ are random. Suppose we manage to bound δ_t from below by some $\delta^* > 0$ for all possible outcomes of δ_t, where $t < T$. Then we know that the process decreases its state ("progresses towards 0") in expectation by at least δ^* in every step, and the additive drift theorem (see Theorem 1 below) will provide a bound on T_0 that only depends on X_0 and δ^*. In fact, the very naturally looking result $E(T_0 \mid X_0) \leq X_0/\delta^*$ will be obtained. However, bounds on the drift might be more complicated. For example, a bound on δ_t might depend on X_t or states at even earlier points of time, e. g., if the progress decreases as the current state decreases. This is often the case in applications to evolutionary algorithms. It is not so often the case that the whole "history" is needed. Simple evolutionary algorithms and other randomized search heuristics are Markov processes such that simply $\delta_t = E(X_t - X_{t+1} \mid X_t)$. With respect to Markov processes on discrete search spaces, drift conditions traditionally use conditional expectations such as $E(X_t - X_{t+1} \mid X_t = i)$ and bound these for arbitrary $i > 0$ instead of directly bounding the random variable $E(X_t - X_{t+1} \mid X_t)$ on $X_t > 0$.

Caveat. As pointed out, the drift δ_t in general is a random variable and should not be confused with the "expected drift" $E(\delta_t) = E(E(X_t - X_{t+1} \mid \mathcal{F}_t))$, which rarely is available since it averages over the whole history of the stochastic process. Drift is based on the inspection of the progress from one step to another, taking into account every possible history. This one-step inspection often makes it easy to come up with bounds on δ_t. Drift theorems could also be formulated based on expected drift; however, this might be tedious to compute. See Jägersküpper

(2011) for one of the rare analyses of "expected drift", which we will not get into in this paper.

We now present the first formal drift theorem dealing with additive drift. It is based on a formulation by He and Yao (2001), from which we removed some unnecessary assumptions, more precisely the discrete search space and the Markov property. We only demand a bounded state space.

Theorem 1 (Additive Drift, following He and Yao (2001))**.** *Let $(X_t)_{t\geq 0}$, be a stochastic process over some bounded state space $S \subseteq \mathbb{R}_0^+$. Assume that $E(T_0 \mid X_0) < \infty$. Then:*

(i) If $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t > 0) \geq \delta_{\mathrm{u}}$ then $E(T_0 \mid X_0) \leq \frac{X_0}{\delta_{\mathrm{u}}}$.

(ii) If $E(X_t - X_{t+1} \mid \mathcal{F}_t) \leq \delta_\ell$ then $E(T_0 \mid X_0) \geq \frac{X_0}{\delta_\ell}$.

By applying the law of total expectation, Statement (i) implies $E(T_0) \leq \frac{E(X_0)}{\delta_{\mathrm{u}}}$ and analogously for Statement (ii).

For the sake of completeness, we also provide as simple proof using martingale theory, inspired by Lehre (2012). This proof is simpler than the original one by He and Yao (2001).

Proof of Theorem 1. We prove only the upper bound since the lower bound is proven symmetrically. We define $Y_t = X_0 + t\delta_{\mathrm{u}}$. Note that as long $t < T$, Y_t is a supermartingale w. r. t. $X_0, \ldots, X_t$, more precisely by induction

$$\begin{aligned} E(Y_{t+1} \mid X_0, \ldots, X_t) &= E(X_{t+1} + (t+1)\delta_{\mathrm{u}} \mid X_0, \ldots, X_t) \\ &\leq X_t - \delta + (t+1)\delta_{\mathrm{u}} = Y_t, \end{aligned}$$

where the inquality uses the drift condition. Since the state space is bounded and $E(T_0 \mid X_0) < \infty$, we can apply the optional stopping theorem and get $0 + E(T_0)\delta_{\mathrm{u}} = E(Y_{T_0} \mid X_0) \leq Y_0 = X_0$. Rearranging terms, the theorem follows. □

Summing up, additive drift is concerned with the very simple scenario that there is a progress of at least δ_{u} from all non-optimal states towards the optimum in (i) and a progress of at most δ_ℓ in (ii). Since the δ-values are not allowed to depend on X_t, one has to use the worst-case drift over all X_t. This might lead to very bad bounds on the first hitting time, which is why more general theorems (as mentioned in the introduction) were developed. It is interesting to note that these more general theorems are often proved based on Theorem 1 above by using an appropriate mapping from the original state space to a new one. Informally, the mapping "smoothes out" position-dependent drift into an (almost) position-independent drift. We will use the same approach in the following.

3 General Drift Theorem

In this section, we present our general drift theorem. As pointed out in the introduction, we strive for a very general statement, which is partly at the expense of simplicity. More user-friendly specializations will be proved in the following sections. Nevertheless, the underlying idea of the complicated-looking general theorem is the same as in all drift theorems. We look into the one-step drift $E(X_t - X_{t+1} \mid \mathcal{F}_t)$ and assume we have a (upper or lower) bound $h(X_t)$ on the drift, which (possibly heavily) depends on X_t. Based on h, a new function g is defined with the aim of "smoothing out" the dependency, and the drift w. r. t. g (formally, $E(g(X_t) - g(X_{t+1}) \mid \mathcal{F}_t)$) is analyzed. Statements (*i*) and (*ii*) of the following Theorem 2 provide bounds on $E(T_a)$ based on the drift w. r. t. g. In fact, g is defined in a very similar way as in existing variable-drift theorems, such that Statements (*i*) and (*ii*) can be understood as generalized variable drift theorems for upper and lower bounds on the expected hitting time, respectively. Statement (*ii*) is also valid (but useless) if the expected hitting time is infinite. Sections 4–6 study specializations of these first two statements into existing variable and multiplicative drift theorems.

Statements (*iii*) and (*iv*) are concerned with tail bounds on the hitting time. Here moment-generating functions of the drift w. r. t. g come into play, formally $E(e^{-\lambda(g(X_t)-g(X_{t+1}))} \mid \mathcal{F}_t)$ is bounded. Again for the sake of generality, bounds on the moment generating function may depend on the current state X_t, as captured by the bounds $\beta_{\mathrm{u}}(X_t)$ and $\beta_\ell(X_t)$. We will see an example in Section 7 where the mapping g smoothes out the position-dependent drift into a (nearly) position-independent drift, while the moment-generating function of the drift w. r. t. g still heavily depends on the current position X_t.

Theorem 2 (General Drift Theorem)**.** *Let* $(X_t)_{t\geq 0}$*, be a stochastic process over some state space* $S \subseteq \{0\} \cup [x_{\min}, x_{\max}]$*, where* $x_{\min} \geq 0$*. Let* $h\colon [x_{\min}, x_{\max}] \to \mathbb{R}^+$ *be an integrable function and define* $g\colon \{0\} \cup [x_{\min}, x_{\max}] \to \mathbb{R}^{\geq 0}$ *by* $g(x) := \frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{x} \frac{1}{h(y)}\,\mathrm{d}y$ *for* $x \geq x_{\min}$ *and* $g(0) := 0$*. Let* $T_a = \min\{t \mid X_t \leq a\}$ *for* $a \in \{0\} \cup [x_{\min}, x_{\max}]$*. Then:*

(i) If $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t \geq x_{\min}) \geq h(X_t)$ *and* $E(g(X_t) - g(X_{t+1}) \mid \mathcal{F}_t; X_t \geq x_{\min}) \geq \alpha_{\mathrm{u}}$ *for some* $\alpha_{\mathrm{u}} > 0$ *then* $E(T_0 \mid X_0) \leq \frac{g(X_0)}{\alpha_{\mathrm{u}}}$*.*

(ii) If $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t \geq x_{\min}) \leq h(X_t)$ *and* $E(g(X_t) - g(X_{t+1}) \mid \mathcal{F}_t; X_t \geq x_{\min}) \leq \alpha_\ell$ *for some* $\alpha_\ell > 0$ *then* $E(T_0 \mid X_0) \geq \frac{g(X_0)}{\alpha_\ell}$*.*

(iii) If $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t > a) \geq h(X_t)$ *and there exists* $\lambda > 0$ *and a function* $\beta_{\mathrm{u}}\colon (a, x_{\max}] \to \mathbb{R}^+$ *such that* $E(e^{-\lambda(g(X_t)-g(X_{t+1}))} \mid \mathcal{F}_t; X_t > a) \leq \beta_{\mathrm{u}}(X_t)$ *then* $\Pr(T_a \geq t^* \mid X_0) < \left(\prod_{r=0}^{t^*-1} \beta_{\mathrm{u}}(X_r)\right) \cdot e^{\lambda(g(X_0)-g(a))}$ *for* $t^* > 0$*.*

(iv) *If* $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t > a) \leq h(X_t)$ *and there exists* $\lambda > 0$ *and a function* $\beta_\ell \colon (a, x_{\max}] \to \mathbb{R}^+$ *such that* $E(e^{\lambda(g(X_t) - g(X_{t+1}))} \mid \mathcal{F}_t; X_t > a) \leq \beta_\ell(X_t)$ *then* $\Pr(T_a < t^* \mid X_0 > a) \leq \left(\sum_{s=1}^{t^*-1} \prod_{r=0}^{s-1} \beta_\ell(X_r)\right) \cdot e^{-\lambda(g(X_0)-g(a))}$ *for* $t^* > 0$.

If additionally the set of states $S \cap \{x \mid x \leq a\}$ *is absorbing, then* $\Pr(T_a < t^* \mid X_0 > a) \leq \left(\prod_{r=0}^{t^*-1} \beta_\ell(X_r)\right) \cdot e^{-\lambda(g(X_0)-g(a))}$.

Special cases of *(iii)* **and** *(iv)***.** If $E(e^{\lambda(g(X_t)-g(X_{t+1}))} \mid \mathcal{F}_t; X_t > a) \leq \beta_{\mathrm{u}}$ for some position-independent β_{u}, then Statement *(iii)* boils down to $\Pr(T_a \geq t^* \mid X_0) < \beta_{\mathrm{u}}^{t^*} \cdot e^{\lambda(g(X_0)-g(a))}$; similarly for Statement *(iv)*.

On $x_{\min}$. Some specializations of Theorem 2 require a "gap" in the state space between optimal and non-optimal states, modelled by $x_{\min} > 0$. One example is multiplicative drift, see Theorem 7 in Section 5. Another example is the process defined by $X_0 \sim \mathrm{Unif}[0,1]$ and $X_t = 0$ for $t > 0$. Its first hitting time of state 0 cannot be derived by drift arguments since the lower bound on the drift towards the optimum within the interval $[0,1]$ has limit 0.

Proof of Theorem 2. The first two items follow from the classical additive drift theorem (Theorem 1). To prove the third one, we use ideas implicit in Hajek (1982) and argue

$$\begin{aligned}\Pr(T_a \geq t^* \mid X_0) \;&\leq\; \Pr(g(X_{t^*}) > g(a) \mid X_0) \;=\; \Pr(e^{\lambda g(X_{t^*})} > e^{\lambda g(a)} \mid X_0)\\ &<\; E(e^{\lambda g(X_{t^*}) - \lambda g(a)} \mid X_0),\end{aligned}$$

where the first inequality uses that $g(x)$ is non-decreasing, the equality that $x \mapsto e^x$ is a bijection, and the last inequality is Markov's inequality. Now,

$$\begin{aligned}E(e^{\lambda g(X_{t^*})} \mid X_0) &= E(e^{\lambda g(X_{t^*-1})} \cdot E(e^{-\lambda(g(X_{t^*-1})-g(X_{t^*}))} \mid X_0, \dots, X_{t^*-1}) \mid X_0)\\ &= e^{\lambda g(X_0)} \cdot \prod_{r=0}^{t^*-1} E(e^{-\lambda(g(X_{r-1})-g(X_r))} \mid X_0, \dots, X_r),\end{aligned}$$

where the last equality follows inductively (note that this does not assume independence of the $g(X_{r-1}) - g(X_r)$). Using the prerequisite from the third item, we get

$$E(e^{\lambda g(X_{t^*})} \mid X_0) \leq e^{\lambda g(X_0)} \prod_{r=0}^{t^*-1} \beta_{\mathrm{u}}(X_r),$$

altogether

$$\Pr(T_a \geq t^* \mid X_0) < e^{\lambda(g(X_0)-g(a))} \prod_{r=0}^{t^*-1} \beta_{\mathrm{u}}(X_r),$$

which proves the third item.

The fourth item is proved similarly as the third one. By a union bound,

$$\Pr(T_a < t^* \mid X_0 > a) \le \sum_{s=1}^{t^*-1} \Pr(g(X_s) \le g(a) \mid X_0)$$

for $t^* > 0$. Moreover,

$$\Pr(g(X_s) \le a \mid X_0) = \Pr(e^{-\lambda g(X_s)} \ge e^{-\lambda a} \mid X_0) \le E(e^{-\lambda g(X_s)+\lambda a} \mid X_0)$$

using again Markov's inequality. By the prerequisites, we get

$$E(e^{-\lambda g(X_s)} \mid X_0) \le e^{-\lambda g(X_0)} \prod_{r=0}^{s-1} \beta_\ell(X_r)$$

Altogether,

$$\Pr(T_a < t^*) \le \sum_{s=1}^{t^*-1} e^{-\lambda(g(X_0)+g(a))} \prod_{r=0}^{s-1} \beta_\ell(X_r).$$

If furthermore $S \cap \{x \mid x \le a\}$ is absorbing then $T_a < t^*$ is equivalent to $X_{t^*} \le a$. In this case,

$$\Pr(T_a < t^* \mid X_0) \le \Pr(g(X_{t^*}) \le g(a) \mid X_0) \le e^{-\lambda(g(X_0)+g(a))} \prod_{r=0}^{t^*-1} \beta_\ell(X_r).$$

□

Our drift theorem is very general and therefore complicated. In order to apply it, specializations might be welcome based on assumptions that typically are satisfied. The rest of this section discusses such simplifications; however, we do not yet apply them in this paper.

By making some additional assumptions on the function h, we get the following special cases.

Lemma 1. *Let $\lambda > 0$, and h be any real-valued, differentiable function. Define $g(x) := \int_{x_{\min}}^{x} 1/h(y)dy$.*

- *If $h'(x) \ge \lambda$ then $f_1(x) = e^{\lambda g(x)}$ is concave.*
- *If $h'(x) \le \lambda$ then $f_1(x) = e^{\lambda g(x)}$ is convex.*
- *If $h'(x) \ge -\lambda$ then $f_2(x) = e^{-\lambda g(x)}$ is convex.*

- *If $h'(x) \le -\lambda$ then $f_2(x) = e^{-\lambda g(x)}$ is concave.*

Proof. The double derivative of f_1 is

$$f_1''(x) = \frac{\lambda e^{\lambda g(x)}}{h(x)^2} \cdot (\lambda - h'(x)),$$

where the first factor is positive. If $h'(x) \ge \lambda$, then $f_1''(x) \le 0$, and f_1 is concave. If $h'(x) \le \lambda$, then $f_1''(x) \ge 0$, and f_1 is convex.

Similarly, the double derivative of f_2 is

$$f_2''(x) = \frac{\lambda e^{-\lambda g(x)}}{h(x)^2} \cdot (\lambda + h'(x)),$$

where the first factor is positive. If $h'(x) \le -\lambda$, then $f_2''(x) \le 0$, and f_2 is concave. If $h'(x) \ge -\lambda$, then $f_2''(x) \ge 0$, and f_2 is convex. □

Corollary 1. *Let $(X_t)_{t\ge 0}$, be a stochastic process over some state space $S \subseteq \{0\} \cup [x_{\min}, x_{\max}]$, where $x_{\min} \ge 0$. Let $h\colon [x_{\min}, x_{\max}] \to \mathbb{R}^+$ be a differentiable function. Then the following statements hold for the first hitting time $T := \min\{t \mid X_t = 0\}$.*

(i) If $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t \ge x_{\min}) \ge h(X_t)$ and $h'(x) \ge 0$, then

$$E(T \mid X_0) \le \frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{X_0} \frac{1}{h(y)}\,\mathrm{d}y.$$

(ii) If $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t \ge x_{\min}) \le h(X_t)$ and $h'(x) \le 0$, then

$$E(T \mid X_0) \ge \frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{X_0} \frac{1}{h(y)}\,\mathrm{d}y.$$

(iii) If $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t \ge x_{\min}) \ge h(X_t)$ and $h'(x) \ge \lambda$ for some $\lambda > 0$, then

$$\Pr(T \ge t \mid X_0) < \exp\left(-\lambda\left(t - \frac{x_{\min}}{h(x_{\min})} - \int_{x_{\min}}^{X_0} \frac{1}{h(y)}\,\mathrm{d}y\right)\right).$$

(iv) a) If $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t \ge x_{\min}) \le h(X_t)$ and $h'(x) \le -\lambda$ for some $\lambda > 0$, then

$$\Pr(T < t \mid X_0 > 0) < \frac{e^{\lambda t} - e^{\lambda}}{e^{\lambda} - 1} \exp\left(-\frac{\lambda x_{\min}}{h(x_{\min})} - \int_{x_{\min}}^{X_0} \frac{\lambda}{h(y)}\,\mathrm{d}y\right).$$

Proof. Let $g(x) := x_{\min}/h(x_{\min}) + \int_{x_{\min}}^{x} 1/h(y)\,\mathrm{d}y$, and note that $g''(x) = -h'(x)/h(x)^2$.

For (i), it suffices to show that condition (i) Theorem 2 is satisfied for $\alpha_u := 1$. From the assumption $h'(x) \geq 0$, it follows that $g''(x) \leq 0$, hence g is a concave function. Jensen's inequality therefore implies that

$$\begin{aligned} E(g(X_t) - g(X_{t+1}) \mid \mathcal{F}_t; X_t \geq x_{\min}) &\geq g(X_t) - g(E(X_{t+1} \mid \mathcal{F}_t; X_t \geq x_{\min})) \\ &\geq \int_{X_t - h(X_t)}^{X_t} \frac{1}{h(y)}\,\mathrm{d}y \\ &\geq \frac{1}{h(X_t)} \cdot h(X_t) = 1, \end{aligned}$$

where the last inequality holds because h is a non-decreasing function.

For (ii), it suffices to show that condition (i) Theorem 2 is satisfied for $\alpha_\ell := 1$. From the assumption $h'(x) \leq 0$, it follows that $g''(x) \geq 0$, hence g is a convex function. Jensen's inequality therefore implies that

$$\begin{aligned} E(g(X_t) - g(X_{t+1}) \mid \mathcal{F}_t; X_t \geq x_{\min}) &\leq g(X_t) - g(E(X_{t+1} \mid \mathcal{F}_t; X_t \geq x_{\min})) \\ &\leq \int_{X_t - h(X_t)}^{X_t} \frac{1}{h(y)}\,\mathrm{d}y \\ &\leq \frac{1}{h(X_t)} \cdot h(X_t) = 1, \end{aligned}$$

where the last inequality holds because h is a non-increasing function.

For (iii), it suffices to show that condition (iii) of Theorem 2 is satisfied for $\beta_u := e^{-\lambda}$. By Lemma 1 and Jensen's inequality, it holds that

$$E(e^{-\lambda(g(X_t) - g(X_{t+1}))} \mid \mathcal{F}_t; X_t \geq x_{\min}) \leq e^{-\lambda r}$$

where

$$\begin{aligned} r &:= g(X_t) - g(E(X_{t+1} \mid \mathcal{F}_t; X_t \geq x_{\min})) \\ &\geq \int_{X_t - h(X_t)}^{X_t} \frac{1}{h(y)}\,\mathrm{d}y \\ &> \frac{1}{h(X_t)} \cdot h(X_t) = 1, \end{aligned}$$

where the last inequality holds because h is strictly monotonically increasing.

For (iv) a), it suffices to show that condition (iv) of Theorem 2 is satisfied for $\beta_\ell := e^{\lambda}$. By Lemma 1 and Jensen's inequality, it holds that

$$E(e^{\lambda(g(X_t) - g(X_{t+1}))} \mid \mathcal{F}_t; X_t \geq x_{\min}) \leq e^{\lambda r}$$

where

$$\begin{aligned} r &:= g(X_t) - g(E(X_{t+1} \mid \mathcal{F}_t; X_t \geq x_{\min})) \\ &\leq \int_{X_t - h(X_t)}^{X_t} \frac{1}{h(y)} \,\mathrm{d}y \\ &< \frac{1}{h(X_t)} \cdot h(X_t) = 1, \end{aligned}$$

where the last inequality holds because h is strictly monotonically decreasing. □

4 Variable Drift as Special Case

The purpose of this section is to show that known variants of variable drift theorems can be derived from our general Theorem 2.

4.1 Classical Variable Drift and Fitness Levels

A clean form of a variable drift theorem, generalizing previous formulations by Johannsen (2010) and Mitavskiy et al. (2009), was recently presented by Rowe and Sudholt (2012). We restate their theorem in our notation and carry out two generalizations that are obvious: we allow for a continuous state space instead of demanding a finite one and do not fix $x_{\min} = 1$.

Theorem 3 (Variable Drift; following Rowe and Sudholt (2012))**.** *Let $(X_t)_{t\geq 0}$, be a stochastic process over some state space $S \subseteq \{0\} \cup [x_{\min}, x_{\max}]$, where $x_{\min} > 0$. Let $h(x)$ be an integrable, monotone increasing function on $[x_{\min}, x_{\max}]$ such that $E(X_t - X_{t+1} \mid \mathcal{F}_t) \geq h(X_t)$ if $X_t \geq x_{\min}$. Then it holds for the first hitting time $T := \min\{t \mid X_t = 0\}$ that*

$$E(T \mid X_0) \leq \frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{X_0} \frac{1}{h(x)} \,\mathrm{d}x.$$

Proof. Since $h(x)$ is monotone increasing, $1/h(x)$ is decreasing and $g(x)$, defined in Theorem 2, is concave. By Jensen's inequality, we get

$$\begin{aligned} E(g(X_t) - g(X_{t+1}) \mid \mathcal{F}_t) \;&\geq\; g(X_t) - g(E(X_{t+1} \mid \mathcal{F}_t)) \\ = \int_{E(X_{t+1}\mid\mathcal{F}_t)}^{X_t} \frac{1}{h(x)} \,\mathrm{d}x \;&\geq\; \int_{X_t - h(X_t)}^{X_t} \frac{1}{h(x)} \,\mathrm{d}x, \end{aligned}$$

where the equality just expanded $g(x)$. Using that $1/h(x)$ is decreasing, it follows

$$\int_{X_t-h(X_t)}^{X_t} \frac{1}{h(x)}\,\mathrm{d}x \geq \int_{X_t-h(X_t)}^{X_t} \frac{1}{h(X_t)}\,\mathrm{d}y = \frac{h(X_t)}{h(X_t)} = 1.$$

Plugging in $\alpha_{\mathrm{u}} := 1$ in Theorem 2 completes the proof. □

Rowe and Sudholt (2012) also pointed out that variable drift theorems in discrete search spaces look very similar to bounds obtained from the fitness level technique (also called the method of f-based partitions, first formulated by Wegener, 2001). For the sake of completeness, we present the classical upper bounds by fitness levels w. r. t. the (1+1) EA here and prove them by drift analyis.

Theorem 4 (Classical Fitness Levels, following Wegener (2001))**.** *Consider the (1+1) EA maximizing some function f and a partition of the search space into non-empty sets $A_1, \dots, A_m$. Assume that the sets form an f-based partition, i. e., for $1 \leq i < j \leq m$ and all $x \in A_i$, $y \in A_j$ it holds $f(x) < f(y)$. Let p_i be a lower bound on the probability that a search point in A_i is mutated into a search point in $A_{i+1} \cup \dots \cup A_m$. Then the expected hitting time of A_m is at most*

$$\sum_{i=1}^{m-1} \frac{1}{p_i}.$$

Proof. At each point of time, the (1+1) EA is in a unique fitness level. Let Y_t the current fitness level at time t. We consider the process defined by $X_t = m - Y_t$. By definition of fitness levels and the (1+1) EA, X_t is non-increasing over time. Consider $X_t = k$ for $1 \leq k \leq m-1$. With probability p_{m-k}, the X-value decreases by at least 1. Consequently, $E(X_t - X_{t+1} \mid X_t = k) \geq p_{m-k}$. We define $h(x) = p_{m-\lceil x \rceil}$, $x_{\min} = 1$ and $x_{\max} = m-1$ and obtain an integrable, monotone increasing function on $[x_{\min}, x_{\max}]$. Hence, the upper bound on $E(T \mid X_0)$ from Theorem 3 becomes at most $\frac{1}{p_1} + \sum_{i=1}^{m-2} \frac{1}{p_{m-i}}$, which completes the proof. □

Recently, the fitness-level technique was considerably refined and supplemented by lower bounds (Sudholt, 2013). We will also identify these extensions as a special case of general drift in Section 6.

4.2 Non-monotone Variable Drift and Lower Bounds by Variable Drift

In many applications, a monotone increasing function $h(x)$ bounds the drift from below. For example, the expected progress towards the optimum of OneMax increases with the distance of the current search point from the optimum. However, recently Doerr et al. (2012) found that that certain ACO algorithms do not have this property and exhibit a non-monotone drift. To handle this case, they present a generalization of Johannsen's drift theorem that does not require $h(x)$ to be monotone. The most recent version of this theorem is presented in Feldmann and Kötzing (2013). Unfortunately, it turned out that the two generalizations suffer from a missing condition, relating positive and negative drift to each other. Adding the condition and removing an unnecessary assumption (more precisely, the continuity of $h(x)$) the theorem by Feldmann and Kötzing (2013) can be corrected as follows.

Theorem 5 (extending Feldmann and Kötzing (2013)). *Let $(X_t)_{t\geq 0}$, be a stochastic process over some state space $S \subseteq \{0\} \cup [x_{\min}, x_{\max}]$, where $x_{\min} > 0$. Suppose there exists two functions $h, d\colon [x_{\min}, x_{\max}] \to \mathbb{R}^+$, where h is integrable, and a constant $c \geq 1$ such that for all $t \geq 0$*

(1) $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t \geq x_{\min}) \geq h(X_t)$,

(2) $\frac{E((X_{t+1}-X_t)\cdot \mathbb{1}\{X_{t+1}>X_t\} \mid \mathcal{F}_t; X_t\geq x_{\min})}{E((X_t-X_{t+1})\cdot \mathbb{1}\{X_{t+1}<X_t\} \mid \mathcal{F}_t; X_t\geq x_{\min})} \leq \frac{1}{2c^2}$,

(3) $|X_t - X_{t+1}| \leq d(X_t)$ *if* $X_t \geq x_{\min}$,

(4) for all $x, y \geq x_{\min}$ *with* $|x-y| \leq d(x)$, *it holds* $h(\min\{x,y\}) \leq c h(\max\{x,y\})$.

Then it holds for the first hitting time $T := \min\{t \mid X_t = 0\}$ *that*

$$E(T \mid X_0) \leq 2c\left(\frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{X_0} \frac{1}{h(x)}\,\mathrm{d}x\right).$$

It is worth noting that Theorem 3 is not necessarily a special case of Theorem 5.

Proof. Using the definition of g according to Theorem 2 and assuming $X_t \geq x_{\min}$, we compute the drift

$$E(g(X_t) - g(X_{t+1}) \mid \mathcal{F}_t) \;=\; E\left(\int_{X_{t+1}}^{X_t} \frac{1}{h(x)}\,\mathrm{d}x \mid \mathcal{F}_t\right)$$
$$=\; E\left(\int_{X_{t+1}}^{X_t} \frac{1}{h(x)}\,\mathrm{d}x \cdot \mathbb{1}\left\{X_{t+1} < X_t\right\} \mid \mathcal{F}_t\right) - E\left(\int_{X_t}^{X_{t+1}} \frac{1}{h(x)}\,\mathrm{d}x \cdot \mathbb{1}\left\{X_{t+1} > X_t\right\} \mid \mathcal{F}_t\right).$$

Item (4) from the prerequisites yields $h(x) \leq c h(X_t)$ if $X_t - d(X_t) \leq x < X_t$ and $h(x) \geq h(X_t)/c$ if $X_t < x \leq X_t + d(X_t)$. Using this and $|X_t - X_{t+1}| \leq d(X_t)$, the drift can be further bounded by

$$
\begin{aligned}
&E\left(\int_{X_{t+1}}^{X_t} \frac{1}{ch(X_t)} \,\mathrm{d}x \cdot \mathbb{1}\left\{X_{t+1} < X_t\right\} \;\middle|\; \mathcal{F}_t\right) - E\left(\int_{X_t}^{X_{t+1}} \frac{c}{h(X_t)} \,\mathrm{d}x \cdot \mathbb{1}\left\{X_{t+1} > X_t\right\} \;\middle|\; \mathcal{F}_t\right) \\
&\geq\; E\left(\int_{X_t}^{X_{t+1}} \frac{1}{2ch(X_t)} \,\mathrm{d}x \cdot \mathbb{1}\left\{X_{t+1} < X_t\right\} \;\middle|\; \mathcal{F}_t\right) \;=\; \frac{E((X_t - X_{t+1} \mid \mathcal{F}_t) \cdot \mathbb{1}\left\{X_{t+1} < X_t\right\})}{2ch(X_t)} \\
&\geq\; \frac{h(X_t)}{2ch(X_t)} \;=\; \frac{1}{2c},
\end{aligned}
$$

where the first inquality used the Item (2) from the prerequisites and the last one Item (1). Plugging in $\alpha_{\mathrm{u}} := 1/(2c)$ in Theorem 2 completes the proof. □

Finally, so far only a single variant dealing with upper bounds on variable drift and thus lower bounds on the hitting time seems to have been published. It was derived by Doerr, Fouz, and Witt (2011). Again, we present a variant without unnecessary assumptions, more precisely we allow continuous state spaces and use less restricted $c(x)$ and $h(x)$.

Theorem 6 (following Doerr, Fouz, and Witt (2011))**.** *Let $(X_t)_{t\geq 0}$, be a stochastic process over some state space $S \subseteq \{0\} \cup [x_{\min}, x_{\max}]$, where $x_{\min} > 0$. Suppose there exists two functions $c(x)$ and $h(x)$ on $[x_{\min}, x_{\max}]$ such that $h(x)$ is monotone increasing and integrable and for all $t \geq 0$,*

1. *$X_{t+1} \leq X_t$,*
2. *$X_{t+1} \geq c(X_t)$ for $X_t \geq x_{\min}$,*
3. *$E(X_t - X_{t+1} \mid \mathcal{F}_t) \leq h(c(X_t))$ for $X_t \geq x_{\min}$.*

Then it holds for the first hitting time $T := \min\{t \mid X_t = 0\}$ that

$$E(T \mid X_0) \geq \frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{X_0} \frac{1}{h(x)} \,\mathrm{d}x.$$

Proof. Using the definition of g according to Theorem 2, we compute the drift

$$
\begin{aligned}
E(g(X_t) - g(X_{t+1}) \mid \mathcal{F}_t) \;&=\; E\left(\int_{X_{t+1}}^{X_t} \frac{1}{h(x)} \,\mathrm{d}x \mid \mathcal{F}_t\right) \\
\leq\; E\left(\int_{X_{t+1}}^{X_t} \frac{1}{h(c(X_t))} \,\mathrm{d}x \mid \mathcal{F}_t\right),
\end{aligned}
$$

where we have used that $X_t \geq X_{t+1} \geq c(X_t)$ and that $h(x)$ is monotone increasing. The last integral equals

$$\frac{X_t - E(X_{t+1} \mid \mathcal{F}_t)}{h(c(X_t))} \,\mathrm{d}x \leq \frac{h(c(x))}{h(c(x))} \;=\; 1.$$

Plugging in $\alpha_\ell := 1$ in Theorem 2 completes the proof. □

5 Multiplicative Drift as Special Case

We continue by showing that Theorem 2 can be specialized in order to re-obtain other classical and recent variants of drift theorems. Of course, Theorem 2 is a generalization of additive drift (Theorem 1), which interestingly was used to prove the general theorem itself. The remaining important strand of drift theorems is thus represented by so-called multiplicative drift, which we focus on in this section.

The following theorem is the strongest variant of the multiplicative drift theorem (originally introduced by Doerr et al. (2012)), which can be found in Doerr and Goldberg (2013). In this section, we for the first time use a tail bound from our main theorem (more precisely, the third item in Theorem 2). Note that the multiplicative drift theorem requires $x_{\min}$ to be positive, i. e., a gap in the state space. Without the gap, no finite first hitting time can be proved from the prerequisites of multiplicative drift.

Theorem 7 (Multiplicative Drift; following Doerr and Goldberg (2013))**.** *Let $(X_t)_{t\geq 0}$, be a stochastic process over some state space $S \subseteq \{0\} \cup [x_{\min}, x_{\max}]$, where $x_{\min} > 0$. Suppose that there exists some δ, where $0 < \delta < 1$ such that $E(X_t - X_{t+1} \mid \mathcal{F}_t) \geq \delta X_t$. Then the following statements hold for the first hitting time $T := \min\{t \mid X_t = 0\}$.*

1. $E(T \mid X_0) \leq \frac{\ln(X_0/x_{\min})+1}{\delta}$.

2. $\Pr(T \geq \frac{\ln(X_0/x_{\min})+r}{\delta} \mid X_0) \leq e^{-r}$ *for all* $r > 0$.

In fact, our formulation is minimally stronger than the one by Doerr and Goldberg, who prove $\Pr(T > \frac{\ln(X_0/x_{\min})+r}{\delta} \mid X_0) \leq e^{-r}$.

Proof. Using the notation from Theorem 2, we choose $h(x) = \delta x$ and obtain $E(X_t - X_{t+1} \mid \mathcal{F}_t) \geq h(X_t)$ by the prerequisite on multiplicative drift. Moreover, $g(x) = x_{\min}/(\delta x_{\min}) + \int_{x_{\min}}^{x} 1/(\delta y)\,\mathrm{d}y = 1/\delta + \ln(x/x_{\min})/\delta$ for $x \geq x_{\min}$. Now we proceed as in the proof of Theorem 3. Since $\ln(x)$ is concave, Jensen's inequality yields $E(g(X_t) - g(X_{t+1}) \mid X_t) \geq g(X_t) - g(E(X_{t+1} \mid X_t)) \geq \ln(X_t/x_{\min})/\delta - \ln((1-\delta)X_t/x_{\min})/\delta = -\ln(1-\delta)/\delta \geq 1$, where the last inequality used $\ln(1-\delta) \leq -\delta$. Hence, using $\alpha_{\mathrm{u}} = 1$ and $g(X_0) \leq 1/\delta + \ln(X_0/x_{\min})/\delta$ in the first item of Theorem 2, we obtain the first claim of this theorem.

To prove the second claim, let $a := 0$ and consider

$$\begin{aligned} E(e^{-\delta(g(X_t)-g(X_{t+1}))} \mid \mathcal{F}_t; X_t \geq x_{\min}) &= E(e^{\ln(X_{t+1}/x_{\min})-\ln(X_t/x_{\min})}) \mid \mathcal{F}_t; X_t \geq x_{\min}) \\ &= E((X_{t+1}/X_t) \mid \mathcal{F}_t; X_t \geq x_{\min}) \leq 1-\delta, \end{aligned}$$

Hence, we can choose $\beta_{\mathrm{u}}(X_t) = 1-\delta$ for all $X_t \geq x_{\min}$ and $\lambda = \delta$ in the third item of Theorem 2 to obtain

$$\Pr(T \geq t^* \mid X_0) < (1-\delta)^{t^*} \cdot e^{\delta(g(X_0)-g(x_{\min}))} \leq e^{-\delta t^* + \ln(X_0/x_{\min})}.$$

Now the claim follows by choosing $t^* := (\ln(X_0/x_{\min}) + r)/\delta$. $\square$

Compared to the upper bound, the following lower-bound version includes a condition on the maximum step-wise progress and requires non-increasing sequences. It generalizes the version in Witt (2013) and its predecessor in Lehre and Witt (2012) in that it does not assume $x_{\min} \geq 1$.

Theorem 8 (Multiplicative drift, Lower Bound; following Witt (2013))**.** *Let $(X_t)_{t\geq 0}$, be a stochastic process over some state space $S \subseteq \{0\} \cup [x_{\min}, x_{\max}]$, where $x_{\min} > 0$. Suppose that there exist β, δ, where $0 < \beta, \delta \leq 1$ such that for all $t \geq 0$*

1. $X_{t+1} \leq X_t$,
2. $\Pr(X_t - X_{t+1} \geq \beta X_t) \ \leq \ \frac{\beta\delta}{1+\ln(X_t/x_{\min})}$.
3. $E(X_t - X_{t+1} \mid \mathcal{F}_t) \leq \delta X_t$

Then it holds for the first hitting time $T := \min\{t \mid X_t = 0\}$ that

$$E(T \mid X_0) \ \geq \ \frac{1 + \ln(X_0/x_{\min})}{\delta} \cdot \frac{1-\beta}{1+\beta}.$$

Proof. Using the definition of g according to Theorem 2, we compute the drift

$$\begin{aligned}
E(g(X_t) - g(X_{t+1}) \mid \mathcal{F}_t) \ &= \ E\left(\int_{X_{t+1}}^{X_t} \frac{1}{h(x)} \,\mathrm{d}x \mid \mathcal{F}_t\right) \\
\leq \ E\left(\int_{X_{t+1}}^{X_t} \frac{1}{h(x)} \,\mathrm{d}x \mid \mathcal{F}_t; X_{t+1} \geq (1-\beta)X_t\right) &\cdot \Pr(X_{t+1} \geq (1-\beta)X_t) \\
&+ g(X_t) \cdot (1 - \Pr(X_{t+1} \geq (1-\beta)X_t))
\end{aligned}$$

where we used the law of total probability and $g(X_{t+1}) \geq 0$. As in the proof of Theorem 7, we have $g(x) = (1 + \ln(x/x_{\min}))/\delta$. Plugging in $h(x) = \delta x$, using the bound on $\Pr(X_{t+1} \geq (1-\beta)X_t)$ and $X_{t+1} \leq X_t$, the drift is further bounded by

$$\begin{aligned}
&E\left(\int_{X_{t+1}}^{X_t} \frac{1}{\delta(1-\beta)X_t} \,\mathrm{d}x \mid \mathcal{F}_t\right) + \frac{\beta\delta}{1 + \ln(X_t/x_{\min})} \cdot \frac{1 + \ln(X_t/x_{\min})}{\delta} \\
&= \ \frac{E(X_t - X_{t+1} \mid \mathcal{F}_t)}{\delta(1-\beta)X_t} + \beta \ \leq \ \frac{\delta X_t}{\delta(1-\beta)X_t} + \beta \ \leq \ \frac{1+\beta}{1-\beta},
\end{aligned}$$

Using $\alpha_\ell = (1+\beta)/(1-\beta)$ and expanding $g(X_0)$, the proof is complete. $\square$

6 Fitness Levels Lower and Upper Bounds as Special Case

We pick up the consideration of fitness levels again and prove the following lower-bound theorem due to Sudholt (2013) by drift analysis. See Sudholt's paper for possibly undefined or unknown terms.

Theorem 9 (Theorem 3 in Sudholt (2013))**.** *Consider an algorithm $\mathcal{A}$ and a partition of the search space into non-empty sets $A_1, \dots, A_m$. For a mutation-based EA $\mathcal{A}$ we again say that $\mathcal{A}$ is in A_i or on level i if the best individual created so far is in A_i. Let the probability of $\mathcal{A}$ traversing from level i to level j in one step be at most $u_i \cdot \gamma_{i,j}$ and $\sum_{j=i+1}^{m} \gamma_{i,j} = 1$. Assume that for all $j > i$ and some $0 \le \chi \le 1$ it holds*

$$\gamma_{i,j} \ge \chi \sum_{k=j}^{m} \gamma_{i,k}. \tag{1}$$

Then the expected hitting time of A_m is at least

$$\begin{aligned} &\sum_{i=1}^{m-1} \Pr(\mathcal{A} \textit{ starts in } A_i) \cdot \left(\frac{1}{u_i} + \chi \sum_{j=i+1}^{m-1} \frac{1}{u_j} \right) \\ &\ge \sum_{i=1}^{m-1} \Pr(\mathcal{A} \textit{ starts in } A_i) \cdot \chi \sum_{j=i}^{m-1} \frac{1}{u_j}. \end{aligned}$$

Proof. Since $\chi \le 1$, the second lower bound follows immediately from the first one, which we prove in the following. To adopt the perspective of minimization, we say that $\mathcal{A}$ is on distance level $m - i$ if the best individual created so far is in A_i. Let X_t be the algorithm's distance level at time t. We define the potential function g mapping distance levels to non-negative numbers (which then form a new stochastic process) by

$$g(m-i) = \frac{1}{u_i} + \chi \sum_{j=i+1}^{m-1} \frac{1}{u_j}$$

for $1 \le i \le m - 1$. Defining $u_m := \infty$, we extend the function to $g(0) = 0$. Our aim is to prove that the drift

$$\Delta_t(m-i) := E(g(m-i) - g(X_{t+1}) \mid X_t = m - i)$$

has expected value at most 1. Then the theorem follows immediately using additive drift (Theorem 1) along with the law of total probability to condition on the starting level.

To analyze the drift, consider the case that the distance level decreases from $m-i$ to $m-\ell$, where $\ell > i$. We obtain

$$g(m-i) - g(m-\ell) = \frac{1}{u_i} - \frac{1}{u_\ell} + \chi \sum_{j=i+1}^{\ell} \frac{1}{u_j},$$

which by the law of total probability (and as the distance level cannot increase) implies

$$\begin{aligned}\Delta_t(m-i) &= \sum_{\ell=i+1}^{m} u_i \cdot \gamma_{i,\ell} \left(\frac{1}{u_i} - \frac{1}{u_\ell} + \chi \sum_{j=i+1}^{\ell} \frac{1}{u_j} \right) \\ &= 1 + u_i \sum_{\ell=i+1}^{m} \gamma_{i,\ell} \left(-\frac{1}{u_\ell} + \chi \sum_{j=i+1}^{\ell} \frac{1}{u_j} \right),\end{aligned}$$

where the last equality used $\sum_{\ell=i+1}^{m} \gamma_{i,\ell} = 1$. If we can prove that

$$\sum_{\ell=i+1}^{m} \gamma_{i,\ell} \chi \sum_{j=i+1}^{\ell} \frac{1}{u_j} \le \sum_{\ell=i+1}^{m} \gamma_{i,\ell} \cdot \frac{1}{u_\ell} \tag{2}$$

then $\Delta_t(m-i) \le 1$ follows and the proof is complete. To show this, observe that

$$\sum_{\ell=i+1}^{m} \gamma_{i,\ell} \chi \sum_{j=i+1}^{\ell} \frac{1}{u_j} = \sum_{j=i+1}^{m} \frac{1}{u_j} \cdot \chi \sum_{\ell=j}^{m} \gamma_{i,\ell}$$

since the term $\frac{1}{u_j}$ appears for all terms $\ell = j, \dots, m$ in the outer sum, each term weighted by $\gamma_{i,\ell}\chi$. By (1), we have $\chi \sum_{\ell=j}^{m} \gamma_{i,\ell} \le \gamma_{i,j}$, and (2) follows. □

We remark here without going into the details that also the refined upper bound by fitness levels (Theorem 4 in Sudholt, 2013) can be proved using general drift.

7 Applying the Tail Bounds

So far we have mostly derived bounds on the expected first hitting time using Statements (*i*) and (*ii*) of our general drift theorem. This section is devoted to applications of the tail bounds (Statements (*iii*) and (*iv*)).

7.1 OneMax and Linear Functions

In this subsection, we study a classical benchmark problem, namely the (1+1) EA on OneMax. We start by deriving very precise bounds on first the expected optimization time and then prove tail bounds. The lower bounds obtained will imply results for a much larger function class. Note that Doerr et al. (2011) already proved the following result.

Statement 1 (Theorems 3 and 5 in Doerr et al., 2011)**.** *The expected optimization time of the (1+1) EA on* OneMax *is at most* $en \ln n - c_1 n + O(1)$ *and at least* $en \ln n - c_2 n$ *for certain constants* $c_1, c_2 > 0$.

The constant c_2 is not made explicit by Doerr et al. (2011), whereas the constant c_1 is stated as 0.369. However, unfortunately this value is due to a typo in the very last line of their proof – c_1 should have been 0.1369 instead. We correct this mistake in a self-contained proof. Furthermore, we improve the lower bound using variable drift. It is worth noting that Doerr et al. (2011) used variable drift as well, but overestimated the drift function. Here we use a more precise (upper) bound on the drift.

Lemma 2. *Let* X_t *denote the number of zeros of the current search point of the (1+1) EA on* OneMax. *Then*

$$\left(1-\frac{1}{n}\right)^{n-x}\frac{x}{n} \le E(X_t - X_{t+1} \mid X_t = x) \le \left(\left(1-\frac{1}{n}\right)\left(1+\frac{x}{(n-1)^2}\right)\right)^{n-x}\frac{x}{n}.$$

Proof. The lower bound considers the expected number of flipping zero-bits, assuming that no one-bit flips. The upper bound is obtained in the proof of Lemma 6 in Doerr et al. (2011) and denoted by $S_1 \cdot S_2$, but is not made explicit in the lemma. □

Theorem 10. *The expected optimization time of the (1+1) EA on* OneMax *is at most* $en \ln n - 0.1369n + O(1)$ *and at least* $en \ln n - 5.9338n - O(\log n)$.

Proof. Note that with probability $1 - 2^{-\Omega(n)}$ we have $\frac{(1-\epsilon)n}{2} \le X_0 \le \frac{(1+\epsilon)n}{2}$ for an arbitrary constant $\epsilon > 0$. Hereinafter, we assume this event to happen, which only adds an error term of absolute value $2^{-\Omega(n)} \cdot n \log n = 2^{-\Omega(n)}$ to the expected optimization time.

In order to apply the variable drift theorem (more precisely, Theorem 3 for the upper and Theorem 6 for the lower bound), we manipulate and estimate the expressions from Lemma 2 to make them easy to integrate. To prove the upper bound on the optimization time, we observe

$$
\begin{aligned}
E(X_t - X_{t+1} \mid X_t = x) &\geq \left(1-\frac{1}{n}\right)^{n-x}\frac{x}{n} \\
&= \left(1-\frac{1}{n}\right)^{n-1}\cdot\left(1-\frac{1}{n}\right)^{-x}\cdot\frac{x}{n}\cdot\left(1-\frac{1}{n}\right) \\
&\geq e^{-1+\frac{x}{n}}\cdot\frac{x}{n}\cdot\left(1-\frac{1}{n}\right) =: h_\ell(x).
\end{aligned}
$$

Now, by the variable drift theorem, the optimization time T satisfies

$$
\begin{aligned}
E(T \mid X_0) &\leq \frac{1}{h_\ell(1)} + \int_1^{(1+\epsilon)n/2}\frac{1}{h_\ell(x)}\,\mathrm{d}x \leq \left(en + \int_1^{(1+\epsilon)n/2} e^{1-\frac{x}{n}}\cdot\frac{n}{x}\right)\left(1-\frac{1}{n}\right)^{-1} \\
&\leq \left(en - en\left[E_1(x/n)\right]_1^{(1+\epsilon)n/2}\right)\left(1+O\!\left(\frac{1}{n}\right)\right),
\end{aligned}
$$

where $E_1(x) := \int_x^\infty \frac{e^{-t}}{t}\,\mathrm{d}t$ denotes the exponential integral (for $x > 0$). The latter is estimated using the series representation $E_1(x) = -\ln x - \gamma + \sum_{k=1}^\infty \frac{(-x)^k}{kk!}$, with $\gamma = 0.577\ldots$ being the Euler-Mascheroni constant (see Abramowitz and Stegun, 1964, Equation 5.1.11). We get for sufficiently small ϵ that

$$
-\left[E_1(x/n)\right]_1^{(1+\epsilon)n/2} = E_1(1/n) - E_1((1+\epsilon)/2) \leq -\ln(1/n) - \gamma + O(1/n) - 0.559774.
$$

Altogether,

$$
E(T \mid X_0) \leq en\ln n + en(1-0.559774-\gamma) + O(\log n) \leq en\ln n - 0.1369n + O(\log n)
$$

which proves the upper bound.

For the lower bound on the optimization time, we need according to Theorem 6 a monotone process (which is satisfied) and a function c bounding the progress towards the optimum. We use $c(x) = x - \log x - 1$. Since each bit flips with probability $1/n$, we get

$$
\begin{aligned}
\Pr(X_{t+1} \leq X_t - \log(X_t) - 1) &\leq \binom{X_t}{\log(X_t)+1}\left(\frac{1}{n}\right)^{\log(X_t)+1} \\
&\leq \left(\frac{eX_t}{n\log(X_t)+n}\right)^{\log(X_t)+1}.
\end{aligned}
$$

The last bound takes its maximum at $X_t = 2$ within the interval $[2,\ldots,n]$ and is $O(n^2)$ then. For $X_t = 1$, we trivially have $X_{t+1} \geq c(X_t) = 0$. Hence, by assuming

$X_{t+1} \geq c(X_t)$ for all $t = O(n \log n)$, we only introduce an additive error of value $O(\log n)$.

Next the upper bound on the drift from Lemma 2 is manipulated. We get for some sufficiently large constant $c^* > 0$ that

$$\begin{aligned} E(X_t - X_{t+1} \mid X_t = x) &\leq \left(\left(1 - \frac{1}{n}\right)\left(1 + \frac{x}{(n-1)^2}\right)\right)^{n-x} \cdot \frac{x}{n} \\ &\leq e^{-1+\frac{x}{n}+\frac{x(n-x)}{n^2}} \cdot \frac{x}{n} \cdot \left(\frac{1 + x/(n-1)^2}{1 + x/(n^2)}\right)^{n-x} \\ &\leq e^{-1+\frac{2x}{n}} \cdot \frac{x}{n} \cdot \left(1 + \frac{c^*}{n}\right) =: h^*(x), \end{aligned}$$

where we used $1 + x \leq e^x$ twice. The drift theorem requires a function $h_{\mathrm{u}}(x)$ such that $h^*(x) \leq h_{\mathrm{u}}(c(x)) = h_{\mathrm{u}}(x - \log x - 1)$. Introducing the substitution $y := y(x) := x - \log x - 1$ and its inverse function $x(y)$, we choose $h_{\mathrm{u}}(y) := h^*(x(y))$. We obtain

$$\begin{aligned} E(T \mid X_0) &\geq \left(\frac{1}{h^*(x(1))} + \int_1^{(1-\epsilon)n/2} \frac{1}{h^*(x(y))} \,\mathrm{d}y\right)\left(1 - O\left(\frac{1}{n}\right)\right) \\ &\geq \left(\frac{1}{h^*(2)} + \int_{x(1)}^{x((1-\epsilon)n/2)} \frac{1}{h^*(x)}\left(1 - \frac{1}{x}\right) \mathrm{d}x\right)\left(1 - O\left(\frac{1}{n}\right)\right) \\ &\geq \left(\frac{en}{2} + \int_2^{(1-\epsilon)n/2} e^{1-\frac{2x}{n}} \cdot \frac{n}{x}\left(1 - \frac{1}{x}\right) \mathrm{d}x\right)\left(1 - O\left(\frac{1}{n}\right)\right) \\ &= \left(\frac{en}{2} + \int_2^{(1-\epsilon)n/2} e^{1-\frac{2x}{n}} \cdot \frac{n}{x} \,\mathrm{d}x - \int_2^{(1-\epsilon)n/2} e^{1-\frac{2x}{n}} \cdot \frac{n}{x^2} \,\mathrm{d}x\right)\left(1 - O\left(\frac{1}{n}\right)\right) \end{aligned}$$

where the second inequality uses integration by substitution and $x(1) = 2$, the third one $x(y) \leq y$, and the last one partial integration.

With respect to the first integral in the last bound, the only difference compared to the upper bound is the 2 in the exponent of $e^{-1+\frac{2x}{n}}$, such that we can proceed analogously to the above and obtain $-enE_1(2x/n) + C$ as anti-derivative. The anti-derivative of the second integral is $2eE_1(2x/n) - e^{1-2x/n}\frac{n}{x} + C$.

We obtain

$$E(T \mid X_0) \geq \left(\frac{en}{2} + \left[-(2e + en)E_1(2x/n) + e^{1-2x/n}\frac{n}{x}\right]_2^{(1-\epsilon)n/2}\right)\left(1 - O\left(\frac{1}{n}\right)\right)$$

Now, for sufficiently small ϵ,

$$-\left[E_1(2x/n)\right]_2^{(1-\epsilon)n/2} \geq -\ln(4/n) - \gamma - O(1/n) - 0.21939 \geq \ln n - 2.18291 - O(1/n)$$

and

$$\left[e^{1-2x/n}\frac{n}{x}\right]_2^{(1-\epsilon)n/2} \ge 1.9999 - \frac{en}{2} - O(1/n).$$

Altogether,

$$E(T \mid X_0) \ge en\ln n - 5.9338n - O(\log n)$$

as suggested. □

Knowing the expected optimization time very precisely, we now can derive sharp bounds. Note that the following upper concentration inequality in Theorem 11 is not new but is already implicit in the work on multiplicative drift analysis by Doerr et al. (2012). In fact, a very similar upper bound is even available for all linear functions (Witt, 2013). By constrast, the lower concentration inequality is a novel non-trivial result.

Theorem 11. *The optimization time of the (1+1) EA on* OneMax *is at least* $en\ln n - cn - ren$*, where* c *is a constant, with probability at least* $1-e^{-r/2}$ *for any* $r\ge 0$*. It is at most* $en\ln n + ren$ *with probability at least* $1-e^{-r}$*.*

Proof of Theorem 11, upper tail. The upper tail can be easily derived from the multiplicative drift theorem with tail bounds (Theorem 7). Let X_t denote the number of zeros at time t. By Lemma 2, we can choose $\delta := 1/(en)$. Then the upper bound follows since $X_0 \le n$ and $x_{\min} = 1$. □

We are left with the lower tail. The aim is to prove it using Theorem 2.(iv), which includes a bound on the moment-generating function of the drift of g. We first set up the h (and thereby the g) used for our purposes. Obviously, $x_{\min} := 1$.

Lemma 3. *Consider the (1+1) EA on* OneMax *and let the random variable* X_t *denote the current number of zeros at time* $t \ge 0$*. Then*

$$h(x) := e^{-1+\frac{2\lceil x\rceil}{n}} \cdot \frac{\lceil x\rceil}{n} \cdot \left(1+\frac{c^*}{n}\right),$$

where $c^* > 0$ *is a sufficiently large constant, satisfies the condition* $E(X_t - X_{t+1} \mid X_t = i) \le h(i)$ *for* $i \in \{1,\dots,n\}$*.*

Moreover, define $g(i) := x_{\min}/h(x_{\min}) + \int_{x_{\min}}^{i} 1/h(y)\,\mathrm{d}y$ *and* $\Delta_t := g(X_t) - g(X_{t+1})$*. Then* $g(i) = \sum_{j=1}^{i} 1/h(j)$ *and*

$$\Delta_t \le \sum_{j=X_{t+1}+1}^{X_t} \frac{e^{1-2X_{t+1}/n}\cdot n}{j}.$$

Proof. According to Lemma 2, $h^*(x) := ((1-\frac{1}{n})(1+\frac{x}{(n-1)^2})^{n-x}\frac{x}{n}$ is an upper bound on the drift. We obtain $h(x) \geq h^*(x)$ using the simple estimations exposed in the proof of Theorem 10, lower bound part.

The representation of $g(i)$ as a sum follows immediately from h due to the ceilings. The bound on Δ_t follows from h by estimating $e^{-1+\frac{2\lceil x\rceil}{n}} \cdot \left(1+\frac{c^*}{n}\right) \geq e^{-1+2x/n}$. $\square$

The next lemma provides a bound on the moment-generating function (mgf.) of the drift of g, depending on the current state. Note that we do not need the whole filtration based on $X_0, \dots, X_t$ but only X_t since we are dealing with a Markov chain.

Lemma 4. *Let $\lambda := 1/(en)$ and $i \in \{1, \dots, n\}$. Then*

$$E(e^{\lambda\Delta_t} \mid X_t = i) \;\leq\; 1 + \lambda + \frac{2\lambda}{i} + o\!\left(\frac{\lambda}{\log n}\right).$$

Proof. We distinguish between three major cases.

Case 1: $i = 1$. Then $X_{t+1} = 0$, implying $\Delta_t \leq en$, with probability at most $(1/n)(1-1/n)^{n-1} = (1/(en))(1+1/(n-1))$ and $X_{t+1} = i$ otherwise. We get

$$\begin{aligned} E(e^{\lambda\Delta_t} \mid X_t = i) \;&\leq\; \frac{1}{en}\cdot e^1 + \left(1 - \frac{1}{en}\right) + O\!\left(\frac{1}{n^2}\right) \\ &\leq\; 1 + \frac{e-1}{en} + O\!\left(\frac{1}{n^2}\right) \;\leq\; 1 + \lambda + \frac{(e-2)\lambda}{i} + o\!\left(\frac{\lambda}{\ln n}\right). \end{aligned}$$

Case 2: $2 \leq i \leq \ln^3 n$. Let $Y := i - X_{t+1}$ and note that $\Pr(Y \geq 2) \leq (\ln^6 n)/n^2$ since the probability of flipping a zero-bit is at most $(\ln^3 n)/n$. We further subdivide the case according to whether $Y \geq 2$ or not.

Case 2a: $2 \leq i \leq \ln^3 n$ and $Y \geq 2$. The largest value of Δ_t is taken when $Y = i$. Using Lemma 3 and estimating the i-th Harmonic number, we have $\lambda\Delta_t \leq (\ln i) + 1 \leq 3(\ln\ln n) + 1$. The contribution to the mgf. is bounded by

$$E(e^{\lambda\Delta_t} \cdot \mathbb{1}\{X_{t+1} \leq i-2\} \mid X_t = i) \;\leq\; e^{3\ln\ln n + 1} \cdot \left(\frac{\ln^6 n}{n^2}\right) = o\!\left(\frac{\lambda}{\ln n}\right).$$

Case 2b: $2 \leq i \leq \ln^3 n$ and $Y < 2$. Then $X_{t+1} \geq X_t - 1$, which implies $\Delta_t \leq en(\ln(X_t) - \ln(X_{t+1}))$. We obtain

$$\begin{aligned} &E(e^{\lambda\Delta_t} \cdot \mathbb{1}\{X_{t+1} \geq i-1\} \mid X_t = i) \;\leq\; E(e^{\ln(i/X_{t+1})}) \;\leq\; E(e^{\ln(1+\frac{i-X_{t+1}}{i-1})}) \\ &\;=\; E\!\left(1 + \frac{Y}{i-1}\right), \end{aligned}$$

where the first inequality estimated $\sum_{i=j+1}^{k} \frac{1}{i} \le \ln(k/j)$ and the second one used $X_{t+1} \ge i-1$. From Lemma 2, we get $E(Y) \le \frac{i}{en}(1+O((\ln^3 n)/n))$ for $i \le \ln^3 n$. This implies

$$
\begin{aligned}
&E\left(1+\frac{i-X_{t+1}}{i-1}\right) \le 1+\frac{i}{en(i-1)}\left(1+O\left(\frac{\ln^3 n}{n}\right)\right) \\
&= 1+\frac{1}{en}\cdot\left(1+\frac{1}{i-1}\right)\left(1+O\left(\frac{\ln^3 n}{n}\right)\right) = 1+\lambda+\frac{2\lambda}{i}+o\left(\frac{\lambda}{\ln n}\right),
\end{aligned}
$$

using $i/(i-1) \le 2$ in the last step. Adding the bounds from the two subcases proves the lemma in Case 2.

Case 3: $i > \ln^3 n$. Note that $\Pr(Y \ge \ln n) \le \binom{n}{\ln n}\left(\frac{1}{n}\right)^{\ln n} \le 1/(\ln n)!$. We further subdivide the case according to whether $Y \ge \ln n$ or not.

Case 3a: $i > \ln^3 n$ and $Y \ge \ln n$. Since $\Delta_t \le en(\ln n + 1)$, we get

$$
E(e^{\lambda\Delta_t}\cdot \mathbb{1}\left\{X_{t+1} \le i-\ln^3 n\right\} \mid X_t = i) \le \frac{1}{(\ln n)!}\cdot e^{\ln n+1} = o\left(\frac{\lambda}{\ln n}\right)
$$

Case 3b: $i > \ln^3 n$ and $Y < \ln n$. Then, using Lemma 3 and proceeding similarly as in Case 2b,

$$
\begin{aligned}
&E(e^{\lambda\Delta_t}\cdot \mathbb{1}\left\{X_{t+1} > i-\ln n\right\} \mid X_t = i) \\
&\le E(e^{\lambda \exp(1-2(i-\ln n)/n)\cdot n\ln(i/X_{t+1})} \mid X_t = i) = E\left(\left(1+\frac{i-X_{t+1}}{X_{t+1}}\right)^{\exp((-2i+\ln n)/n)}\right).
\end{aligned}
$$

Using $i > \ln^3 n$ and Jensen's inequality, the last expectation is at most

$$
\begin{aligned}
&\left(1+E\left(\frac{i-X_{t+1}}{X_{t+1}}\right)\right)^{\exp((-2i+\ln n)/n)} \le \left(1+E\left(\frac{Y}{i-\ln n}\right)\right)^{\exp((-2i+\ln n)/n)} \\
&\le \left(1+E\left(\frac{Y}{i(1-1/\ln^2 n)}\right)\right)^{\exp((-2i+\ln n)/n)},
\end{aligned}
$$

where the last inequality used again $i > \ln^3 n$. Since $E(Y) \le e^{-1+2i/n}\frac{i}{n}(1+c^*/n)$, we conclude

$$
\begin{aligned}
&E(e^{\lambda\Delta_t}\cdot \mathbb{1}\left\{X_{t+1} > i-\ln n\right\} \mid X_t = i) \le \left(1+\frac{e^{2i/n}}{en(1-1/\ln^2 n)}\right)^{\exp((-2i+\ln n)/n)} \\
&\le \left(1+\frac{1}{en(1-1/\ln^2 n)}\right)\left(1+O\left(\frac{\ln n}{n^2}\right)\right) \le 1+\lambda+o\left(\frac{\lambda}{\ln n}\right),
\end{aligned}
$$

where we used $(1+ax)^{1/a} \le 1+x$ for $x \ge 0$ and $a \ge 1$. Adding up the bounds from the two subcases, we have proved the lemma in Case 3.

Altogether,

$$E(e^{\lambda\Delta_t} \mid X_t = i) \le 1 + \lambda + \frac{2\lambda}{i} + o\left(\frac{\lambda}{\ln n}\right).$$

for all $i \in \{1, \dots, n\}$. $\square$

The bound on the mgf. of Δ_t derived in the previous Lemma 4 is particularly large for $i = O(1)$, i. e., if the current state number X_t is small. If $X_t = O(1)$ held during the whole optimization process, we could not prove the lower tail in Theorem 11 from the lemma. However, it is easy to see that $X_t = i$ only holds for an expected number of at most en/i steps. Hence, most of the time during the optimization the term $2\lambda/i$ from the lemma is negligible, and the position-dependent $\beta_\ell(X_t)$-term from Theorem 2.(iv) comes into play. We make this precise in the following final proof, where we iteratively bound the probability of the process being at "small" states.

Proof of Theorem 11, lower tail. With overwhelming probability $1-2^{-\Omega(n)}$, $X_0 \ge (1-\epsilon)n/2$ for an arbitrarily small constant $\epsilon > 0$, which we assume to happen. We consider phases in the optimization process. Phase 1 starts with initialization and ends before the first step where $X_t < e^{\frac{\ln n - 1}{2}} = \sqrt{n} \cdot e^{-1/2}$. Phase i, where $i > 1$, follows Phase $i-1$ and ends before the first step where $X_t < \sqrt{n} \cdot e^{-i/2}$. Obviously, the optimum is not found before the end of Phase $\ln(n)$; however, this does not tell us anything about the optimization time yet.

We say that Phase i is *typical* if it does not end before time $eni - 1$. We will prove inductively that the probability of one of the first i phases not being typical is at most $c'e^{\frac{i}{2}}/\sqrt{n} = c'e^{\frac{i-\ln n}{2}}$ for some constant $c' > 0$. This implies the theorem since an optimization time of at least $en\ln n - cn - ren$ is implied by the event that Phase $\ln n - \lceil r - c/e \rceil$ is typical, which has probability at least $1 - c'e^{\frac{-r+c/e+1}{2}} = 1 - e^{\frac{-r}{2}}$ for $c = e(2\ln c' + 1)$.

Fix some $k > 1$ and assume for the moment that all phases up to and including Phase $k-1$ are typical. Then for $1 \le i \le k-1$, we have $X_t \ge \sqrt{n}e^{-i/2}$ in Phase i, i. e., when $en(i-1) \le t \le eni - 1$. We analyze the event that additionally Phase k is typical, which subsumes the event $X_t \ge \sqrt{n}e^{-k/2}$ throughout Phase k. According to Lemma 4, we get

$$E\bigl(e^{\lambda\Delta_t} \mid X_t\bigr) \le 1 + \lambda + \frac{2\lambda e^{i/2}}{\sqrt{n}} + o\left(\frac{\lambda}{\ln n}\right) = e^{\lambda + \frac{2\lambda e^{i/2}}{\sqrt{n}} + o\left(\frac{\lambda}{\ln n}\right)}$$

in Phase i, where $1 \le i \le k$, and therefore for $\lambda := \frac{1}{en}$

$$\prod_{t=0}^{enk-1} E\bigl(e^{\lambda\Delta_t} \mid X_t\bigr) \le e^{\lambda enk + \frac{2\lambda en}{\sqrt{n}}\sum_{i=1}^{k} e^{i/2} + enk\cdot o\left(\frac{\lambda}{\ln n}\right)} \le e^{k + \frac{6e^{k/2}}{n\sqrt{n}} + o(1)} \le e^{k+o(1)},$$

where we used that $k \le \ln n$. From Theorem 2.(*iv*) for $a = \sqrt{n}e^{-k/2}$ and $t^* = enk - 1$ we obtain

$$\Pr(T_a < t^*) \le e^{k+o(1)-\lambda(g(X_0)-g(e^{-k/2}/\sqrt{n}))}.$$

From the proof of of Theorem 10, lower bound part, we already know that $g(X_0) \ge en\ln n - c''n$ for some constant $c'' > 0$ (which is assumed large enough to subsume the $-O(\log n)$ term). Moreover, $g(x) \le en(\ln x + 1)$ according to Lemma 3. We get

$$\Pr(T_a < t^*) \le e^{k+o(1)-\ln n+O(1)-k/2+(\ln n)/2} = e^{\frac{k-\ln n+O(1)}{2}} = c'''\frac{e^{k/2}}{\sqrt{n}},$$

for some sufficiently large constant $c''' > 0$, which proves the desired bound on the probability of Phase k not being typical (without making statements about the earlier phases). The probability that all phases up to and including Phase k are typical is then at least $1 - (\sum_{i=1}^{k} c'''e^{i/2})/\sqrt{n} \ge 1 - c'e^{k/2}/\sqrt{n}$ for an appropriate constant $c' > 0$. This completes the proof. □

Finally, we deduce a concentration inequality w. r. t. linear functions, i. e., functions of the kind $f(x_1, \dots, x_n) = w_1x_1 + \dots + w_nx_n$, where $w_i \ne 0$. This class of functions contains OneMax and has been subject of intense research the last 15 years.

Theorem 12. *The optimization time of the (1+1) EA on an arbitrary linear function with non-zero weights is at least $en\ln n - cn - ren$, where c is a constant, with probability at least $1-e^{-r/2}$ for any $r \ge 0$. It is at most $en\ln n+(1+r)en+O(1)$ with probability at least $1 - e^{-r}$.*

Proof. The upper tail is proved in Theorem 5.1 in Witt (2013). The lower bound follows from the lower tail in Theorem 11 in conjunction with the fact that the optimization time within the class of linear functions is stochastically smallest for OneMax (Theorem 6.2 in Witt, 2013). □

7.2 Simplifications of the Tail Bounds

The third and fourth condition of Theorem 2 involve a moment-generating function, which may be tedious to compute. Greatly inspired by Hajek (1982) and Lehre (2012), we show that bounds on moment-generating function follow from more user-friendly conditions. They are based on stochastic dominance of random variables, which is represented by the symbol $\prec$ in the following theorem.

Theorem 13. *Let $(X_t)_{t\ge0}$, be a stochastic process over some state space $S \subseteq \{0\} \cup [x_{\min}, x_{\max}]$, where $x_{\min} \ge 0$. Let $h\colon [x_{\min}, x_{\max}] \to \mathbb{R}^+$ be an integrable* *function. Suppose there exist a random variable Z and some $\lambda > 0$ such that $|\int_{X_{t+1}}^{X_t} 1/h(x)\,\mathrm{d}x| \prec Z$ for $X_t \geq x_{\min}$ and $E(e^{\lambda Z}) = D$. Then the following two statements hold for the first hitting time $T := \min\{t \mid X_t = 0\}$.*

(i) If $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t \geq x_{\min}) \geq h(X_t)$ then for any $\delta > 0$, $\eta := \min\{\lambda, \delta\lambda^2/(D-1-\lambda)\}$ and $t^* > 0$ it holds

$$\Pr(T \geq t^* \mid X_0) \leq e^{\eta(\int_{x_{\min}}^{X_0} 1/h(x)\,\mathrm{d}x - (1-\delta)t^*)}.$$

(ii) If $E(X_t - X_{t+1} \mid \mathcal{F}_t; X_t \geq x_{\min}) \leq h(X_t)$ then for any $\delta > 0$, $\eta := \min\{\lambda, \delta\lambda^2/(D-1-\lambda)\}$ and $t^* > 0$ it holds

$$\Pr(T < t^* \mid X_0) \leq \frac{e^{\eta((1+\delta)t^* - \int_{x_{\min}}^{X_0} 1/h(x)\,\mathrm{d}x)}}{\eta(1+\delta)}.$$

Furthermore, if state 0 is absorbing then

$$\Pr(T < t^* \mid X_0) \leq e^{\eta((1+\delta)t^* - \int_{x_{\min}}^{X_0} 1/h(x)\,\mathrm{d}x)}$$

Stochastic dominance. Theorem 13 assumes a stochastic dominance of the kind $|\int_{X_{t+1}}^{X_t} 1/h(x)\,\mathrm{d}x| \prec Z$. This is implied by $|X_{t+1} - X_t|(1/\inf_{x \geq x_{\min}} h(x)) \prec Z$.

Proof. As in Theorem 2, define $g\colon \{0\} \cup [x_{\min}, x_{\max}] \to \mathbb{R}^{\geq 0}$ by $g(x) := \frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{x} \frac{1}{h(y)}\,\mathrm{d}y$ for $x \geq x_{\min}$ and $g(0) := 0$. Let $\Delta_t := g(X_t) - g(X_{t+1})$ and note that $\Delta_t = \int_{X_{t+1}}^{X_t} \frac{1}{h(x)}\,\mathrm{d}x$. To satisfy the third condition of Theorem 2, we note

$$\begin{aligned} E(e^{-\eta\Delta_t}) &= 1 - \eta E(\Delta_t) + \sum_{k=2}^{\infty} \frac{\eta^k E(\Delta_t^k)}{k!} \leq 1 - \eta E(\Delta_t) + \eta^2 \sum_{k=2}^{\infty} \frac{\eta^{k-2} E(|\Delta_t|^k)}{k!} \\ &\leq 1 - \eta E(\Delta_t) + \eta^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2} E(|\Delta_t|^k)}{k!} = 1 - \eta + \frac{\eta^2}{\lambda^2}(e^{\lambda Z} - \lambda E(Z) - 1), \end{aligned}$$

where we have used $E(\Delta_t) \geq 1$ (proved in Theorem 2) and $\lambda \geq \eta$. Since $|\Delta_t| \prec Z$, also $E(Z) \geq 1$. Using $e^{\lambda Z} = D$ and $\eta \leq \delta\lambda^2/(D-1-\lambda)$, we obtain

$$E(e^{-\eta\Delta_t}) \leq 1 - \eta + \delta\eta = 1 - (1-\delta)\eta \leq e^{-\eta(1-\delta)}.$$

Setting $\beta_{\mathrm{u}} := e^{-\eta(1-\delta)}$ and using η as the λ of Theorem 2 proves the first statement.

For the second statement, analogous calculations prove

$$E(e^{\eta\Delta_t}) \leq 1 + (1+\delta)\eta \leq e^{\eta(1+\delta)}.$$

We set $\beta_\ell := e^{\eta(1+\delta)}$, use η as the λ of Theorem 2.(iv) and note that

$$\frac{e^{\lambda(1+\delta)t} - e^{\lambda(1+\delta)}}{e^{\lambda(1+\delta)} - 1} \le \frac{e^{\lambda(1+\delta)t}}{\lambda(1+\delta)},$$

which was to be proven. If additionally an absorbing state 0 is assumed, the stronger upper bound follows from the corresponding statement in Theorem 2.(iv). □

7.3 LeadingOnes

Doerr et al. (2013) have proved tail bounds on the optimization time of the (1+1) EA on LeadingOnes. Their result represents a fundamentally new contribution, but suffers from the fact it is dependent on a very specific structure and closed formula for the optimization time. Using Theorem 13, we will prove similarly strong tail bounds without needing this exact formula.

As in Doerr et al. (2013), we are actually interested in a more general statement. Let $T(a)$ denote the number of steps until an LeadingOnes-value of at least a is reached, where $0 \le a \le n$. A key observation is that the drift of the LeadingOnes-value can be determined exactly. Here and hereinafter, let $X_t := \max\{0, a - \text{LeadingOnes}(x_t)\}$ denote the distance in LeadingOnes-value of the search point at time t from the target a.

Lemma 5. *For all $i > 0$, $E(X_t - X_{t+1} \mid X_t = i) = (2 - 2^{-n+a-i+1}) \cdot (1 - 1/n)^{a-i} \cdot (1/n)$.*

Proof. (This is taken from Doerr et al. (2013)). The leftmost zero-bit is at position $a-i+1$. To increase the LeadingOnes-value (it cannot decrease), it is necessary to flip this bit and not to flip the first $a-i$ bits, which is reflected by the last two terms in the lemma. The first term is due to the expected number of free-rider bits. Note that there can be between 0 and $n-a+i-1$ such bits. By the usual argumentation using a geometric distribution, the expected number of free riders in an improving step equals

$$\sum_{k=0}^{n-a+i-1} k \cdot \left(\frac{1}{2}\right)^{\min\{n-a+i-1,k+1\}} = 1 - 2^{-n+a-i+1},$$

hence the expected progress in an improving step is $2 - 2^{-n+a-i+1}$. □

We can now supply the tail bounds.

Theorem 14. *Let $T(a)$ the time for the (1+1) EA to reach a* LeadingOnes*-value of at least a. Moreover, let $r \ge 0$. Then*

1. $E(T(a)) = \frac{n^2-n}{2}\left(\left(1+\frac{1}{n-1}\right)^a - 1\right)$.

2. For $0 < a \le n - \log n$, *we have*

$$T(a) \le \frac{n^2}{2}\left(\left(1+\frac{1}{n-1}\right)^a - 1\right) + r$$

with probability at least $1 - e^{-\Omega(rn^{-3/2})}$.

3. For $\log^2 n - 1 \le a \le n$, *we have*

$$T(a) \ge \frac{n^2-n}{2}\left(\left(1+\frac{1}{n-1}\right)^a - 1 - \frac{2\log^2 n}{n}\right) - r$$

with probability at least $1 - e^{-\Omega(rn^{-3/2})} - e^{-\Omega(\log^2 n)}$.

Proof. The first statement is already contained in Doerr et al. (2013) and proved without drift analysis.

We now turn to the second statement. From Lemma 5, $h(x) = (2-2/n)(1-1/n)^{a-x}/n$ is a lower bound on the drift $E(X_t - X_{t+1} \mid X_t = x)$ if $x \ge \log n$. To bound the change of the g-function, we observe that $h(x) \ge 1/(en)$ for all $x \ge 1$. This means that $X_t - X_{t+1} = k$ implies $g(X_t) - g(X_{t+1}) \le enk$. Moreover, to change the LeadingOnes-value by k, it is necessary that

- the first zero-bit flips (which has probability $1/n$)
- $k-1$ free-riders occur.

The change does only get stochastically larger if we assume an infinite supply of free-riders. Hence, $g(X_t) - g(X_{t+1})$ is stochastically dominated by a random variable $Z = enY$, where Y

- is 0 with probability $1 - 1/n$ and
- follows the geometric distribution with parameter $1/2$ otherwise (where the support is $1, 2, \dots$).

The mgf. of Y therefore equals

$$E(e^{\lambda Y}) = \left(1 - \frac{1}{n}\right)e^0 + \frac{1}{n}\frac{1/2}{e^{-\lambda} - (1-1/2)} \le 1 + \frac{1}{n(1-2\lambda)},$$

where we have used $e^{-\lambda} \ge 1 - \lambda$. For the mgf. of Z it follows

$$E(e^{\lambda Z}) = E(e^{\lambda enY}) \le 1 + \frac{1}{n(1-2en\lambda)},$$

hence for $\lambda := 1/(4en)$ we get $D := E(e^{\lambda Z}) = 1 + 2/n = 1 + 8e\lambda$, which means $D - 1 - \lambda = (8e-1)\lambda$. We get

$$\eta := \frac{\delta\lambda^2}{D-1-\lambda} = \delta(8e-1)\lambda = \frac{\delta(8e-1)}{4en}$$

(which is less than λ if $\delta \le 8e-1$) . Choosing $\delta := n^{-1/2}$, we obtain $\eta = Cn^{-3/2}$ for $C := (8e-1)/(4e)$.

We set $t^* := (\int_{x_{\min}}^{X_0} 1/h(x)\,\mathrm{d}x + r)/(1-\delta)$ in the first statement of Theorem 13. The integral within t^* can be bounded according to

$$\begin{aligned} U &:= \int_{x_{\min}}^{X_0} \frac{1}{h(x)}\,\mathrm{d}x \le \sum_{i=1}^{a} \frac{1}{(2-2/n)(1-1/n)^{a-i}/n} \\ &= \left(\frac{1}{2} + \frac{1}{2n-2}\right) \cdot n \cdot \frac{(1+1/(n-1))^a - 1}{1/(n-1)} = \frac{n^2}{2}\left(\left(1+\frac{1}{n-1}\right)^a - 1\right) \end{aligned}$$

Hence, using the theorem we get

$$\Pr(T \ge t^*) = \Pr(T \ge (U+r)/(1-\delta)) \le e^{-\eta r} \le e^{-Crn^{-3/2}}.$$

Since $U \le en^2$ and $1/(1-\delta) \le 1 + 2\delta = 1 + 2n^{-1/2}$, we get

$$\Pr(T \ge U + 2en^{3/2} + 2r) \le e^{-Crn^{-3/2}}.$$

Using the upper bound on U derived above, we obtain

$$\Pr\left(T \ge \frac{n^2}{2}\left(\left(1+\frac{1}{n-1}\right)^a - 1\right) + r\right) \le e^{-\Omega(rn^{-3/2})}$$

as suggested.

Finally, we prove the third statement of this theorem in a quite symmetrical way to the second one. We can choose $h(x) := 2(1-1/n)^{a-x}/n$ as an upper lower bound on the drift $E(X_t - X_{t+1} \mid X_t = x)$. The estimation of the $E(e^{\lambda Z})$ still applies. We set $t^* := (\int_{x_{\min}}^{X_0} 1/h(x)\,\mathrm{d}x - r)/(1-\delta)$. Moreover, we assume $X_0 \ge n - \log^2 n - 1$, which happens with probability at least $1 - e^{-\Omega(\log^2 n)}$. Note that

$$\begin{aligned} L &:= \int_{x_{\min}}^{X_0} \frac{1}{h(x)}\,\mathrm{d}x \ge \sum_{i=1}^{a-\log^2 n} \frac{1}{2(1-1/n)^{a-i}/n} \\ &= \frac{n^2-n}{2}\left(\left(1+\frac{1}{n-1}\right)^a - \left(1+\frac{1}{n-1}\right)^{\log^2 n}\right) \\ &\ge \frac{n^2-n}{2}\left(\left(1+\frac{1}{n-1}\right)^a - 1 - \frac{\log^2 n}{n}\right), \end{aligned}$$

where the last inequality used $e^x \leq 1 + 2x$ for $x \leq 1$. The second statement of Theorem 13 yields (since state 0 is absorbing)

$$\Pr(T < t^*) = \Pr(T < (U - r)/(1 + \delta)) \leq e^{-\eta r} \leq e^{-Crn^{-3/2}}.$$

Now, since

$$\frac{L - r}{1 + \delta} \geq (L - r) - \delta(L - r) \geq L - r - en^{3/2},$$

(using $L \leq en^2$), we get the third statement by analogous calculations as above. □

8 Conclusions

We have presented a general and versatile drift theorem with tail bounds. The new theorem can be understood as a general variable drift theorem and can be specialized into all existing variants of variable, additive and multiplicative drift theorems we found in the literature as well as the fitness-level technique. Moreover, it provides for lower and upper tail bounds, which were not available before in the context of variable drift. We used the tail bounds to prove sharp concentration inequalities on the optimization time of the (1+1) EA on OneMax, linear functions and LeadingOnes. The proofs also give general advice on how to use the tail bounds and we provide simplified (specialized) versions of the corresponding statements.

We believe that the research presented here helps consolidating the area of drift analysis. The general formulation of drift analysis increases our understanding of the power of the technique and also its limitations. The tail bounds were heavily awaited in order to prove more practically relevant statements on the optimization time besides the plain expected time. We expect to see further applications of our theorem in the future.

References

Abramowitz, M. and I. A. Stegun (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* National Bureau of Standards. `http://people.math.sfu.ca/~cbm/aands/abramowitz_and_stegun.pdf`.

Auger, A. and B. Doerr (Eds.) (2011). *Theory of Randomized Search Heuristics: Foundations and Recent Developments.* World Scientific Publishing.

Doerr, B., M. Fouz, and C. Witt (2011). Sharp bounds by probability-generating functions and variable drift. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2011)*, pp. 2083–2090. ACM Press.

Doerr, B. and L. A. Goldberg (2013). Adaptive drift analysis. *Algorithmica 65*(1), 224–250.

Doerr, B., A. Hota, and T. Kötzing (2012). Ants easily solve stochastic shortest path problems. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2012)*, pp. 17–24. ACM Press.

Doerr, B., T. Jansen, C. Witt, and C. Zarges (2013). A method to derive fixed budget results from expected optimisation times. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2013)*. ACM Press. To appear.

Doerr, B., D. Johannsen, and C. Winzen (2012). Multiplicative drift analysis. *Algorithmica 64*(4), 673–697.

Feldmann, M. and T. Kötzing (2013). Optimizing expected path lengths with ant colony optimization using fitness proportional update. In *Proceedings of Foundations of Genetic Algorithms (FOGA 2013)*, pp. 65–74. ACM Press.

Hajek, B. (1982). Hitting and occupation time bounds implied by drift analysis with applications. *Advances in Applied Probability 14*, 502–525.

He, J. and X. Yao (2001). Drift analysis and average time complexity of evolutionary algorithms. *Artificial Intelligence 127*(1), 57–85. Erratum in Artif. Intell. 140(1/2): 245-248 (2002).

Jägersküpper, J. (2011). Combining markov-chain analysis and drift analysis - the (1+1) evolutionary algorithm on linear functions reloaded. *Algorithmica 59*(3), 409–424.

Jansen, T. (2013). *Analyzing Evolutionary Algorithms - The Computer Science Perspective*. Natural Computing Series. Springer.

Jansen, T. and C. Zarges (2012). Fixed budget computations: A different perspective on run time analysis. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2012)*, pp. 1325–1332. ACM Press.

Johannsen, D. (2010). *Random combinatorial structures and randomized search heuristics*. Ph. D. thesis, International Max Planck Research School for Computer Science, Universität des Saarlandes.

Lehre, P. K. (2012). Drift analysis (tutorial). In *Companion to GECCO 2012*, pp. 1239–1258. ACM Press.

Lehre, P. K. and C. Witt (2012). Black-box search by unbiased variation. *Algorithmica 64*(4), 623–642.

Mitavskiy, B., J. E. Rowe, and C. Cannings (2009). Theoretical analysis of local search strategies to optimize network communication subject to preserving the total number of links. *International Journal of Intelligent Computing and Cybernetics 2*(2), 243–284.

Neumann, F. and C. Witt (2010). *Bioinspired Computation in Combinatorial Optimization – Algorithms and Their Computational Complexity.* Natural Computing Series. Springer.

Oliveto, P. S. and C. Witt (2011). Simplified drift analysis for proving lower bounds in evolutionary computation. *Algorithmica 59*(3), 369–386.

Oliveto, P. S. and C. Witt (2012). Erratum: Simplified drift analysis for proving lower bounds in evolutionary computation. Technical report, `http://arxiv.org/abs/1211.7184`.

Rowe, J. E. and D. Sudholt (2012). The choice of the offspring population size in the (1,λ) EA. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2012)*, pp. 1349–1356. ACM Press.

Sasak, G. H. and B. Hajek (1988). The time complexity of maximum matching by simulated annealing. *Journal of the Association of Computing Machinery 35*, 387–403.

Sudholt, D. (2013). A new method for lower bounds on the running time of evolutionary algorithms. *IEEE Transactions on Evolutionary Computation 17*(3), 418–435.

Wegener, I. (2001). Theoretical aspects of evolutionary algorithms. In *Proceedings of the 28th International Colloquium on Automata, Languages and Programming (ICALP 2001)*, Volume 2076 of *Lecture Notes in Computer Science*, pp. 64–78. Springer.

Witt, C. (2013). Tight bounds on the optimization time of a randomized search heuristic on linear functions. *Combinatorics, Probability & Computing 22*(2), 294–318.