# Global Linear Convergence of Evolution Strategies on More Than Smooth Strongly Convex Functions

Youhei Akimoto, Anne Auger, Tobias Glasmachers, Daiki Morinaga

# GLOBAL LINEAR CONVERGENCE OF EVOLUTION STRATEGIES ON MORE THAN SMOOTH STRONGLY CONVEX FUNCTIONS

YOUHEI AKIMOTO *, ANNE AUGER †, TOBIAS GLASMACHERS ‡, AND DAIKI MORINAGA §

**Abstract.** Evolution strategies (ESs) are zeroth-order stochastic black-box optimization heuristics invariant to monotonic transformations of the objective function. They evolve a multivariate normal distribution, from which candidate solutions are generated. Among different variants, CMA-ES is nowadays recognized as one of the state-of-the-art zeroth-order optimizers for difficult problems. Albeit ample empirical evidence that ESs with a step-size control mechanism converge linearly, theoretical guarantees of linear convergence of ESs have been established only on limited classes of functions. In particular, theoretical results on convex functions are missing, where zeroth-order and also first-order optimization methods are often analyzed. In this paper, we establish almost sure linear convergence and a bound on the expected hitting time of an ES family, namely the $(1+1)_\kappa$-ES, which includes the (1+1)-ES with (generalized) one-fifth success rule and an abstract covariance matrix adaptation with bounded condition number, on a broad class of functions. The analysis holds for monotonic transformations of positively homogeneous functions and of quadratically bounded functions, the latter of which particularly includes monotonic transformation of strongly convex functions with Lipschitz continuous gradient. As far as the authors know, this is the first work that proves linear convergence of ES on such a broad class of functions.

**Key words.** Evolution strategies, Randomized Derivative Free Optimization, Black-box optimization, Linear Convergence, Stochastic Algorithms

**AMS subject classifications.** 65K05, 90C25, 90C26, 90C56, 90C59

## 1. Introduction.

We consider the unconstrained minimization of an objective function $f : \mathbb{R}^d \to \mathbb{R}$ without the use of derivatives where an optimization solver sees $f$ as a zeroth-order *black-box oracle* [12, 47, 48]. This setting is also referred to as derivative-free optimization [15]. Such problems can be advantageously approached by randomized algorithms that can typically be more robust to noise, non-convexity and irregularities of the objective function than deterministic algorithms. There has been recently a vivid interest in randomized derivative-free algorithms giving rise to several theoretical studies of randomized direct search methods [25], trust region [9,26] and model-based methods [13, 49]. We refer to [40] for an in-depth survey including the references of this paragraph and additional ones.

In this context, we investigate Evolution Strategies (ES), which are among the oldest randomized derivative-free or zeroth-order black-box methods [16, 50, 53]. They are widely used in applications in different domains [4, 11, 20–22, 27, 39, 44, 56, 57]. Notably a specific ES called covariance-matrix-adaptation ES (CMA-ES) [30] is among the best solvers to address *difficult* black-box problems. It is affine-invariant and implements complex adaptation mechanisms for the sampling covariance matrix and step-size. It performs well on many ill-conditioned, non-convex, non-smooth, and non-separable problems [29, 52]. ES are known to be difficult to analyze. Yet, given their importance in practice, it is essential to study them from a theoretical convergence

*Faculty of Engineering, Information and Systems, University of Tsukuba; RIKEN AIP, Tsukuba, Japan (akimoto@cs.tsukuba.ac.jp).

†Inria and CMAP, Ecole Polytechnique, IP Paris, France (anne.auger@inria.fr).

‡Institute for Neural Computation, Ruhr-University Bochum, Bochum, Germany (tobias.glasmachers@ini.rub.de).

§Department of Computer Science, University of Tsukuba; RIKEN AIP, Tsukuba, Japan (morinaga@bbo.cs.tsukuba.ac.jp).

perspective.

We focus on the arguably simplest and oldest adaptive ES, denoted (1+1)-ES. It samples a candidate solution from a Gaussian distribution whose step-size (standard deviation) is adapted. The candidate solution is accepted if and only if it is better than the current one (see pseudo-code Algorithm 2.1). The algorithm shares some similarities with simplified direct search whose complexity analysis has been presented in [38]. Yet the (1+1)-ES is comparison-based and thus invariant to strictly increasing transformations of the objective function. Simplified direct search can be thought of as a variant of mesh adaptive direct search [1, 6]. Arguably, in contrast to direct search, a sufficient decrease condition cannot be guaranteed. This causes some difficulties for the analysis. The (1+1)-ES is rotational invariant, while direct search candidate solutions are created along a predefined set of vectors. While the CMA-ES should always be preferred for practical applications over the (1+1)-ES variant analyzed here, this latter variant achieves faster linear convergence on well-conditioned problems when compared to algorithms with established complexity analysis (see [54, Table 6.3 and Figure 6.1] and [8, Figure B.4] where the random pursuit algorithm and the (1+1)-ES algorithms are compared, and also Appendix A).

Prior theoretical studies of the (1+1)-ES with 1/5 success rule have established the global linear convergence on differentiable positively homogeneous functions (composed with a strictly increasing function) with a single optimum [7, 8]. Those results establish the almost sure linear convergence from all initial states. They however do not provide the dependency of the convergence rate with respect to the dimension. A more specific study on the sphere function $f(x) = \frac{1}{2}\|x\|^2$ establishes lower and upper bounds on the expected hitting time of an $\epsilon$-ball of the optimum in $\Theta(\log(d\|m_0 - x^*\|/\epsilon))$, where $x^*$ is the optimum of the function, $m_0$ is the initial solution, and $d$ is the problem dimension [3]. Prior to that, a variant of the $(1+1)$-ES with one-fifth success rule had been analyzed on the sphere and certain convex quadratic functions establishing bounds on the expected hitting time with overwhelming probability in $\Theta(\log(\kappa_f d\|m_0 - x^*\|/\epsilon))$, where $\kappa_f$ is the condition number (the ratio between the greatest and smallest eigenvalues) of the Hessian [33–36]. Recently, the class of functions where the convergence of the (1+1)-ES was proven has been extended to continuously differentiable functions. This analysis does not address the question of linear convergence, focusing only on convergence as such, which is possibly sublinear [23].

Our main contribution is as follows. For a generalized version of the (1+1)-ES with one-fifth success rule, we prove bounds on the expected hitting time akin to linear convergence, i.e., hitting an $\epsilon$-ball in $\Theta(\log\|m_0 - x^*\|/\epsilon)$ iterations on a quite general class of functions. This class of functions includes all composites of Lipschitz-smooth strongly convex functions with a strictly increasing transformation. This latter transformation allows to include some non-continuous functions, and even functions with non-smooth level sets. We additionally deduce linear convergence with probability one. Our analysis relies on finding an appropriate Lyapunov function with lower and upper-bounded expected drift. It is building on classical fundamental ideas presented by Hajek [28] and widely used to analyze stochastic hill-climbing algorithms on discrete search spaces [42].

**Notation.** Throughout the paper, we use the following notations. The set of natural numbers $\{1, 2, \dots, \}$ is denoted $\mathbb{N}$. Open, closed, and left open intervals on $\mathbb{R}$ are denoted by $(,)$, $[,]$, and $(,]$, respectively. The set of strictly positive real numbers is denoted by $\mathbb{R}_>$. The Euclidean norm on $\mathbb{R}^d$ is denoted by $\|\ \|$. Open and closed balls with center $c$ and radius $r$ are denoted as $\mathcal{B}(c, r) = \{x \in \mathbb{R}^d : \|x - c\| < r\}$ and $\bar{\mathcal{B}}(c, r) = \{x \in \mathbb{R}^d : \|x - c\| \leqslant r\}$, respectively. Lebesgue measures on $\mathbb{R}$ and $\mathbb{R}^d$ are both denoted by the same symbol $\mu$. A multivariate normal distribution with mean $m$ and covariance matrix $\Sigma$ is denoted by $\mathcal{N}(m, \Sigma)$. Its probability measure and its induced probability density under Lebesgue measure are denoted by $\Phi(\cdot; m, \Sigma)$ and $\varphi(\cdot; m, \Sigma)$. The indicator function of a set or condition $C$ is denoted by $1\{C\}$. We use Bachmann-Landau notations: $o(\cdot)$, $O(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$, $\omega(\cdot)$.

## 2. Algorithm, Definitions and Objective Function Assumptions.

### 2.1. Algorithm: (1+1)-ES with Success-based Step-size Control.

We analyze a generalized version of the (1+1)-ES with one-fifth success rule presented in Algorithm 2.1, which implements one of the oldest approaches to adapt the step-size in randomized optimization methods [16, 50, 53]. The specific implementation was proposed in [37]. At each iteration, a candidate solution $x_t$ is sampled. It is centered in the current incumbent $m_t$ and follows a multivariate normal distribution with mean vector $m_t$ and covariance matrix equal to $\sigma_t^2 I_d$ where $I_d$ denotes the identity matrix. The candidate solution is accepted, that is $m_t$ becomes $x_t$, if and only if $x_t$ is better than $m_t$ (i.e. $f(x_t) \leqslant f(m_t)$). In this case, we say that the candidate solution is successful. The step-size $\sigma_t$ is adapted so as to maintain a probability of success to be approximately the target success probability denoted by $p_{\text{target}} := \frac{\log(1/\alpha_\downarrow)}{\log(\alpha_\uparrow/\alpha_\downarrow)}$. To do so, the step-size is increased by the increase factor $\alpha_\uparrow > 1$ in case of success (which is an indication that the step-size is likely to be too small) and decreased by the decrease factor $\alpha_\downarrow < 1$ otherwise. The covariance matrix $\Sigma_t$ of the sampling distribution of candidate solutions is adapted in the set $\mathcal{S}_\kappa$ of positive-definite symmetric matrices with determinant $\det(\Sigma) = 1$ and condition number $\mathrm{Cond}(\Sigma) \leqslant \kappa$. We do not assume any specific update mechanism for $\Sigma$, but we assume that the update of $\Sigma$ is invariant to any strictly increasing transformation of $f$. We call such an update comparison-based (see Lines 7 and 11 of Algorithm 2.1). Then, our algorithm behaves exact-equally on $f$ and on $g \circ f$ for all strictly increasing functions $g : \mathbb{R} \to \mathbb{R}$ $\big($i.e., $g(s) \lesseqgtr g(t) \Leftrightarrow s \lesseqgtr t\big)$. This defines a class of comparison-based randomized algorithms and we denote it as (1+1)-ES$_\kappa$. For $\kappa = 1$, it is simply denoted as (1+1)-ES.

**Algorithm 2.1** (1+1)-ES$_\kappa$ with success-based step-size adaptation

1: **input** $m_0 \in \mathbb{R}^d$, $\sigma_0 > 0$, $\Sigma_0 = I$, $f : \mathbb{R}^d \to \mathbb{R}$, **parameter** $\alpha_\uparrow > 1 > \alpha_\downarrow > 0$
2: **for** $t = 1, 2, \dots,$ *until stopping criterion is met* **do**
3: sample $x_t \sim m_t + \sigma_t \mathcal{N}(0, \Sigma_t)$
4: **if** $f\big(x_t\big) \leqslant f\big(m_t\big)$ **then**
5: $m_{t+1} \leftarrow x_t$ ▷ move to the better solution
6: $\sigma_{t+1} \leftarrow \sigma_t \alpha_\uparrow$ ▷ increase the step size
7: $\Sigma_{t+1} \in \mathcal{S}_\kappa$ ▷ adapt the covariance matrix
8: **else**
9: $m_{t+1} \leftarrow m_t$ ▷ stay where we are
10: $\sigma_{t+1} \leftarrow \sigma_t \alpha_\downarrow$ ▷ decrease the step size
11: $\Sigma_{t+1} \in \mathcal{S}_\kappa$ ▷ adapt the covariance matrix

Note that $\alpha_\uparrow$ and $\alpha_\downarrow$ are not meant to be tuned depending on the function properties. How to choose such constants for $\Sigma_t = I_d$ is well-known and is related to the so-called evolution window [51]. In practice, $\alpha_\downarrow = \alpha_\uparrow^{-1/4}$ is the most commonly used setting, which leads to $p_{\text{target}} = 1/5$. It has been shown to be close to optimal,

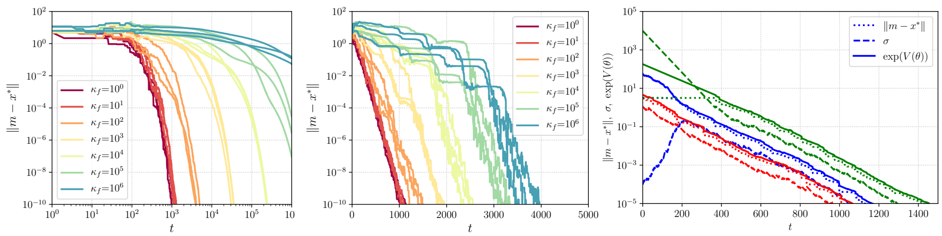


Fig. 2.1: Convergence of the (1+1)-ES (left) and the (1+1)-CMA-ES (middle) on 10 dimensional ellipsoidal function $f(x) = \frac{1}{2}\sum_{i=1}^{d} \kappa_f^{\frac{i-1}{d-1}} x_i^2$ with $\kappa_f = 10^0, 10^1, \ldots, 10^6$. The y-axis displays the distance to the optimum (and not the function value). We employ the covariance matrix adaptation mechanism proposed by [5], where $\sigma$ is adapted as in Algorithm 2.1 with $\alpha_{\uparrow} = e^{0.1}$ and $\alpha_{\downarrow} = e^{-0.025}$. Note the logarithmic scale of the time axis of the left plot vs. the linear time axis of the middle plot.
Right: Three runs of (1+1)-ES ($\alpha_{\uparrow} = e^{0.1}$ and $\alpha_{\uparrow} = e^{-0.025}$) on 10 dimensional spherical function $f(x) = \frac{1}{2}\|x - x^*\|^2$ with initial step-size $\sigma_0 = 10^{-4}$, 1, and $10^4$ (in blue, red, green, respectively). Plotted are the distance to the optimum (dotted line), the step-size (dashed line), and the potential function $V(\theta)$ defined in (4.5) (solid line) with $v = 4/d$, $\ell = \alpha_{\uparrow}^{-10}$, and $u = \alpha_{\downarrow}^{-10}$.

which gives nearly optimal (linear) convergence rate on the sphere function [16, 50]. Hereunder we write $\theta = (m, \sigma, \Sigma)$ as the state of the algorithm, $\theta_t = (m_t, \sigma_t, \Sigma_t)$ and the state-space is denoted by $\Theta$.

Figure 2.1 shows typical runs of the (1+1)-ES and a version of (1+1)-ES$_\kappa$ proposed in [5], which is known as the (1+1)-CMA-ES, on a 10-dimensional ellipsoidal function with different condition numbers $\kappa_f$ of the Hessian. It is empirically observed that $\Sigma_t$ in the (1+1)-CMA-ES approaches the inverse Hessian $\nabla^2 f(m_t)$ of the objective function up to the scalar factor if the objective function is convex quadratic. The runtime of (1+1)-ES scales linearly with $\kappa_f$ (notice the logarithmic scale of the horizontal axis), while the runtime of the (1+1)-CMA-ES suffers only an additive penalty, roughly proportional to the logarithm of $\kappa_f$. Once the Hessian is well approximated by $\Sigma$ (up to a scalar factor), it approaches the global optimum geometrically at the same rate for different values of $\kappa_f$.

In our analysis, we do not assume any specific $\Sigma$ update mechanism, hence it does not necessarily behave as shown in Figure 2.1. Our analysis is therefore the worst case analysis (for the upper bound of the runtime) and the best case analysis (for the lower bound of the runtime) among the algorithms in (1+1)-ES$_\kappa$.

### 2.2. Preliminary definitions.

#### 2.2.1. Spatial Suboptimality Function.

The algorithms studied in this paper are comparison-based and thus invariant to strictly increasing transformations of $f$. If the convergence of the algorithms is measured in terms of $f$, say by investigating the convergence or hitting time of the sequence $f(m_t)$, this will not reflect the invariance to monotonic transformations of $f$ because the first iteration $t_0$ such that $f(m_{t_0}) \leqslant \epsilon$ is not equal to the first iteration $t_0'$ such that $g(f(m_{t_0'})) \leqslant \epsilon$ for some $\epsilon > 0$. For this reason, we introduce a quality measure called *spatial suboptimality function* [23]. It is the $d$th root of the volume of the sub-levelset where the function value is better or equal to $f(x)$:

Definition 2.1 (Spatial Suboptimality Function). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a measurable function with respect to the Borel $\sigma$ algebra of $\mathbb{R}^d$ (simply referred to as measurable function in the sequel). Then the spatial suboptimality function $f_\mu : \mathbb{R}^d \to [0, +\infty]$ is defined as*

$$f_\mu(x) = \sqrt[d]{\mu\left(f^{-1}\left((-\infty, f(x)]\right)\right)} = \sqrt[d]{\mu\left(\left\{y \in \mathbb{R}^d \,\middle|\, f(y) \leqslant f(x)\right\}\right)} \ . \tag{2.1}$$

We remark that for any $f$, the suboptimality function $f_\mu$ is greater or equal to zero. For any $f$ and any strictly increasing function $g : \mathrm{Im}(f) \to \mathbb{R}$, $f$ and its composite $g \circ f$ have the same spatial suboptimality function such that hitting time of $f_\mu$ smaller than $\epsilon > 0$ will be the same for $f$ or $g \circ f$. Moreover, there exists a strictly increasing function $g$ such that $f_\mu(x) = g(f(x))$ holds $\mu$-almost everywhere [23, Lemma 1].

We will investigate the expected first hitting time of $\|m_t - x^*\|$ to $\epsilon > 0$. For this, we will bound the first hitting time of $\|m_t - x^*\|$ to $\epsilon$ by the first hitting time of $f_\mu(m_t)$ to a constant times $\epsilon$. To understand why, consider first a strictly convex quadratic function $f$ with Hessian $H$ and minimal solution $x^*$. We have $f_\mu(x) = V_d\big[2(f(x) - f(x^*))/\det(H)^{1/d}\big]^{1/2}$ for all $x \in \mathbb{R}^d$, where $V_d = \pi^{1/2}/\Gamma^{1/d}(d/2+1)$ is the $d$th root of the volume of the $d$-dimensional unit hyper-sphere [2]. This implies that the first hitting time of $f_\mu(m_t)$ translates to the first hitting time of $\sqrt{f(m_t) - f(x^*)}$. We have $\sqrt{\lambda_{\min}}\|x - x^*\| \leqslant \sqrt{f(x) - f(x^*)} \leqslant \sqrt{\lambda_{\max}}\|x - x^*\|$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimal and maximal eigenvalues of $H$. E.g., consider $f(x) = \|x - x^*\|^2 + 1$. Then the above condition also translates to the first hitting time of $\|m_t - x^*\|$. More generally, we will formalize an assumption on $f$ later on (Assumption A1), which allow us to bound $\|x - x^*\|$ by a constant times $f_\mu(x)$ from above and below (see (2.6)), implying that the first hitting time of $\|m_t - x^*\|$ to $\epsilon$ is bounded by that of $f_\mu(m_t)$ to $\epsilon$, times a constant.

**2.2.2. Success Probability.** The success probability, i.e., the probability of sampling a candidate solution $x_t$ with an objective function better than or equal to that of the current solution $m_t$, plays an important role in the analysis of the (1+1)-ES$_\kappa$ with success-based step-size control mechanism. We present here several useful definitions related to the success probability.

We start with the definition of the *success domain with rate $r$* and the *success probability with rate $r$*. The probability to sample in the $r$-success domain is called success probability with rate $r$. When $r = 0$ we simply talk about success probability.[1]

Definition 2.2 (Success Domain). *For a measurable function $f : \mathbb{R}^d \to \mathbb{R}$ and $m \in \mathbb{R}^d$ such that $f_\mu(m) < \infty$, the $r$-success domain at $m$ with $r \in [0, 1]$ is defined as*

$$S_r(m) = \{x \in \mathbb{R}^d \mid f_\mu(x) \leqslant (1 - r) f_\mu(m)\} \ . \tag{2.2}$$

Definition 2.3 (Success Probability). *Let $f$ be a measurable function and let $m_0 \in \mathbb{R}^d$ be the initial search point satisfying $f_\mu(m_0) < \infty$. For any $r \in [0, 1]$ and any $m \in S_0(m_0)$, the success probability with rate $r$ at $m$ under the normalized step-size*

[1] For $r = 0$, the success domain $S_0(m)$ is not necessarily equivalent to the sub-levelset $S_0'(m) := \{x \in \mathbb{R}^d \mid f(x) \leqslant f(m)\}$, where it always holds that $S_0'(m) \subseteq S_0(m)$. However, since it is guaranteed that $\mu(S_0(m) \setminus S_0'(m)) = 0$ by [23, Lemma 1], due to the absolute continuity of $\Phi(;0,\Sigma)$ for $\Sigma \in \mathcal{S}_\kappa$, the success probability with rate $r = 0$ is equivalent to $\Pr_{z\sim\mathcal{N}(0,\Sigma)}\left[m + f_\mu(m) \cdot \bar{\sigma} z \in S_0'(m)\right]$, with $\bar{\sigma}$ defined in (2.3).

*$\bar\sigma$ is defined as*

$$p_r^{\mathrm{succ}}(\bar\sigma; m, \Sigma) = \Pr_{z\sim\mathcal{N}(0,\Sigma)}\left[m + f_\mu(m)\bar\sigma z \in S_r(m)\right] \ . \tag{2.3}$$

Definition 2.3 introduces the notion of *normalized step-size* $\bar\sigma$ and the success probability is defined as a function of $\bar\sigma$ rather than the actual step-size $\sigma = f_\mu(m)\bar\sigma$. This is motivated by the fact that as $m$ approaches the global optimum $x^*$ of $f$, the step-size $\sigma$ needs to shrink for the success probability to be constant. If the objective function is $f(x) = \frac{1}{2}\|x - x^*\|^2$ and the covariance matrix is the identity matrix, then the success probability is fully controlled by $\bar\sigma_t = \sigma_t/f_\mu(m_t) \propto \sigma_t/\|m_t - x^*\|$ and is independent of $m_t$. This statement can be formalized in the following way.

LEMMA 2.4. *If $f(x) = \frac{1}{2}\|x - x^*\|^2$, then letting $e_1 = (1, 0, \ldots, 0)$, we have*

$$p_r^{\mathrm{succ}}(\bar\sigma; m, \mathrm{I}) = \Pr_{z\sim\mathcal{N}(0,\mathrm{I})}\left[m + f_\mu(m)\bar\sigma z \in S_r(m)\right] = \Pr_{z\sim\mathcal{N}(0,\mathrm{I})}\left[\|e_1 + V_d\bar\sigma z\| \leqslant (1 - r)\right] \ .$$

*Proof.* The suboptimality function is the $d$-th rooth of the volume of a sphere of radius $\|x - x^*\|$. Hence $f_\mu(x) = V_d\|x - x^*\|$. Then, the proof follows the derivation in Section 3 in [3]. ∎

Therefore, $\bar\sigma$ is more discriminative than $\sigma$ itself. In general, the optimal step-size is not necessarily proportional to neither $\|m_t - x^*\|$ nor $f_\mu(m_t)$.

Since the success probability under a given normalized step-size depends on $m$ and $\Sigma$, we define the upper and lower success probability as follows.

DEFINITION 2.5 (Lower and Upper Success Probability). *Let $\mathcal{X}_a^b = \{x \in \mathbb{R}^d : a < f_\mu(x) \leqslant b\}$. Given the normalized step-size $\bar\sigma > 0$, the lower and upper success probabilities are defined as*

$$p_{(a,b]}^{\mathrm{lower}}(\bar\sigma) = \inf_{m\in\mathcal{X}_a^b}\inf_{\Sigma\in\mathcal{S}_\kappa} p_0^{\mathrm{succ}}(\bar\sigma; m, \Sigma) \ , \quad p_{(a,b]}^{\mathrm{upper}}(\bar\sigma) = \sup_{m\in\mathcal{X}_a^b}\sup_{\Sigma\in\mathcal{S}_\kappa} p_0^{\mathrm{succ}}(\bar\sigma; m, \Sigma) \ .$$

A central quantity for our analysis is the limit for $\bar\sigma$ to 0 of the success probability $p_0^{\mathrm{succ}}(\bar\sigma; m, \Sigma)$. Intuitively, if this limit is too small for a given $m$ (compared to $p_{\mathrm{target}}$), because the ruling principle of the algorithm is to decrease the step-size if the probability of success is smaller than $p_{\mathrm{target}}$, the step-size will keep decreasing, causing undesired convergence. Following Glasmachers [23], we introduce the concepts of *p-improbability* and *p-criticality*. They are defined in [23] by the probability of sampling a better point from the isotropic normal distribution in the limit of the step-size to zero. Here, we define $p$-improvability and $p$-criticality for a general multivariate normal distribution.

DEFINITION 2.6 ($p$-improvability and $p$-criticality). *Let $f$ be a measurable function. The function $f$ is called $p$-improvable at $m \in \mathbb{R}^d$ under the covariance matrix $\Sigma \in \mathcal{S}_\kappa$ if there exists $p \in (0, 1]$ such that*

$$p = \liminf_{\bar\sigma\to+0} p_0^{\mathrm{succ}}(\bar\sigma; m, \Sigma) \ . \tag{2.4}$$

*Otherwise, it is called $p$-critical.*

The connection to the classical definition of the critical points for continuously differentiable functions is summarized in the following proposition, which is an extension of Lemma 4 in [23], taking a non-identity covariance matrix into account.

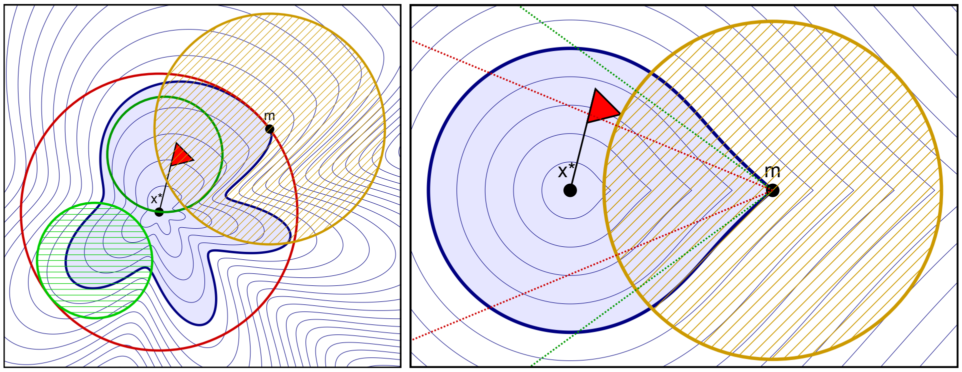


Fig. 2.2: The sampling distribution is indicated by the mean $m$ and the shaded orange circle, indicating one standard deviation. The blue set is the sub-levelset $S_0(m)$ of points improving upon $m$. **Left:** Illustration of property A1 in Subsection 2.3. The blue set is enclosed in the red (outer) ball of radius $C_u f_\mu(m)$ and contains the dark green (inner) ball of radius $C_\ell f_\mu(m)$. The shaded light green ball indicates the worst case situation captured by the bound, namely that the small ball is positioned within the large ball at maximal distance to $m$. **Right:** Illustration of property A2 in Subsection 2.3. Although the level set has a kink at $m$, there exists a cone centered at $m$ covering a probability mass of $p^{\text{limit}}$ of improving steps (inside $S_0(m)$) for small enough step size $\sigma$ (green outline). It contains a smaller cone (red outline) covering a probability mass of $p^{\text{target}}$.

PROPOSITION 2.7. *Let $f = g \circ h$ be a measurable function where $g$ is any strictly increasing function and $h$ is continuously differentiable. Then, $f$ is $p$-improvable with $p = 1/2$ at any regular point $m$ where $\nabla h(m) \neq 0$ under any $\Sigma \in \mathcal{S}_\kappa$. Moreover, if $h$ is twice continuously differentiable at a critical point $m$ where $\nabla h(m) = 0$ and at least one eigenvalue of $\nabla^2 f(m)$ is non-zero, under any $\Sigma \in \mathcal{S}_\kappa$, $m$ is $p$-improvable with $p = 1$ if $\nabla^2 h(m)$ has only non-positive eigenvalues, $p$-critical if $\nabla^2 h(m)$ has only non-negative eigenvalues, and $p$-improvable with some $p > 0$ if $\nabla^2 h(x)$ has at least one strictly negative eigenvalue.*

*Proof.* Note that $p_0^{\text{succ}}(\bar\sigma; m, \Sigma)$ on $f$ is equivalent to $p_0^{\text{succ}}(\bar\sigma; m, I_d)$ on $f'(x) = f(m + \sqrt{\Sigma}(x - m))$. Therefore, it suffices to show that the claims hold for $\Sigma = I_d$ on $f'$, which is proven in Lemma 4 in [23]. ∎

**2.3. Main Assumptions on the Objective Functions.** Given positive real numbers $a$ and $b$ satisfying $0 \leqslant a < b \leqslant +\infty$, and a measurable objective function, let $\mathcal{X}_a^b$ be the set defined in Definition 2.5.

We pose two core assumptions on the objective functions under which we will derive an upper bound on the expected first hitting time of $[0, \epsilon]$ by $f_\mu(m_t)$ (Theorem 4.5) provided $a \leqslant \epsilon \leqslant f_\mu(m_0) \leqslant b$. First, we require to be able to embed and include balls of radius scaling with $f_\mu(m)$ into the sublevel sets of $f$. We do not require this to hold on the whole search space but, for a set $\mathcal{X}_a^b$.

A1 We assume that $f$ is a measurable function and that there exists $\mathcal{X}_a^b$ such that for any $m \in \mathcal{X}_a^b$, there exist an open ball $\mathcal{B}_\ell$ with radius $C_\ell f_\mu(m)$ and a closed ball $\bar{\mathcal{B}}_u$ with radius $C_u f_\mu(m)$ such that it holds $\mathcal{B}_\ell \subseteq \{x \in \mathbb{R}^d \mid f_\mu(x) < f_\mu(m)\}$ and $\{x \in \mathbb{R}^d \mid f_\mu(x) \leqslant f_\mu(m)\} \subseteq \bar{\mathcal{B}}_u$.

We do not specify the center of those balls that may or may not be centered on an optimum of the function. We will see in Proposition 4.1 that this assumption allows to bound $p^{\text{lower}}_{(a,b]}(\bar\sigma)$ and $p^{\text{upper}}_{(a,b]}(\bar\sigma)$ by tractable functions of $\bar\sigma$ which will be essential for the analysis. The property is illustrated in Figure 2.2.

The second assumption requires that the functions are $p$-improvable for $p$ which is lower-bounded uniformly over $\mathcal{X}^b_a$.

A2 Let $f$ be a measurable function, we assume that there exists $\mathcal{X}^b_a$ and there exists $p^{\text{limit}} > p^{\text{target}}$ such that for any $m \in \mathcal{X}^b_a$ and any $\Sigma \in \mathcal{S}_\kappa$, the objective function $f$ is $p$-improvable for some $p \geqslant p^{\text{limit}}$, i.e.,

$$\liminf_{\bar\sigma \downarrow 0} p^{\text{lower}}_{(a,b]}(\bar\sigma) \geqslant p^{\text{limit}} \ . \tag{2.5}$$

The property is illustrated in Figure 2.2. This assumption implies in particular for a continuous function that $\mathcal{X}^b_a$ does not contain any local optimum. This latter assumption is required to obtain global convergence [23, Theorem 2] even without any covariance matrix adaptation (i.e. with $\kappa = 1$) and it can be intuitively understood: If we have a point which is $p$-improvable with $p < p_{\text{target}}$ and which is not a local minimum of the function, then, starting with a small step-size, the success-based step-size control may keep decreasing the step-size at such a point and the (1+1)-ES$_\kappa$ will prematurely converge to a point that is not a local optimum.

If A1 is satisfied with balls centered at the optimum $x^*$ of the function $f$, then it is easy to see that for all $x \in \mathcal{X}^b_a$

$$C_\ell f_\mu(x) \leqslant \|x - x^*\| \leqslant C_u f_\mu(x) \ . \tag{2.6}$$

If the balls are not centered at the optimum, we have the one-side inequality $\|x - x^*\| \leqslant 2C_u f_\mu(x)$. Hence, the expected first hitting time of $f_\mu(m_t)$ to $[0, \epsilon]$ translates to an upper bound for the expected first hitting time of $\|m_t - x^*\|$ to $[0, 2C_u\epsilon]$.

We remark that A1 and A2 satisfied for $a = 0$ allow to include some non-differentiable functions with non-convex sublevel sets as illustrated in Figure 2.2.

We now give two examples of functions that satisfy A1 and A2, including function classes where linear convergence of numerical optimization algorithms are typically analyzed. The first class consists of quadratically bounded functions. It includes all strongly-convex functions with Lipschitz continuous gradient. It also includes some non-convex functions. The second class consists of positively homogeneous functions. The levelsets of a positively homogeneous function are all geometrically similar around $x^*$.

A3 We assume that $f = g \circ h$ where $g$ is a strictly increasing function and $h$ is measurable, continuously differentiable with the unique critical point $x^*$, and quadratically bounded around $x^*$, i.e., for some $L_u \geqslant L_\ell > 0$,

$$(L_\ell/2)\|x - x^*\|^2 \leqslant h(x) - h(x^*) \leqslant (L_u/2)\|x - x^*\|^2 \ . \tag{2.7}$$

A4 We assume that $f = g \circ h$ where $h$ is continuously differentiable and positively homogeneous with a unique optimum $x^*$, i.e., for some $\gamma > 0$

$$h(x^* + \gamma x) = h(x^*) + \gamma\left(h(x^* + x) - h(x^*)\right) \ . \tag{2.8}$$

The following lemmas show that these assumptions imply A1 and A2. The proofs of the lemmas are presented in Appendix B.1 and Appendix B.2, respectively.

LEMMA 2.8. *Let $f$ satisfy A3. Then, $f$ satisfies A1 and A2 with $a = 0$, $b = \infty$, $C_\ell = \frac{1}{V_d}\sqrt{\frac{L_\ell}{L_u}}$ and $C_u = \frac{1}{V_d}\sqrt{\frac{L_u}{L_\ell}}$.*

LEMMA 2.9. *Let $f$ be positively homogeneous satisfying A4, then the suboptimality function $f_\mu(x)$ is proportional to $h(x) - h(x^*)$ and satisfies A1 and A2 for $a = 0$ and $b = \infty$ with $C_u = \sup\{\|x - x^*\| : f_\mu(x) = 1\}$ and $C_\ell = \inf\{\|x - x^*\| : f_\mu(x) = 1\}$.*

## 3. Methodology: Additive Drift on Unbounded Continuous Domains.

### 3.1. First Hitting Time.

We start with the generic definition of the *first hitting time* of a stochastic process $\{X_t : t \geqslant 0\}$, defined as follows.

DEFINITION 3.1 (First hitting time). *Let $\{X_t : t \geqslant 0\}$ be a sequence of real-valued random variables adapted to the natural filtration $\{\mathcal{F}_t : t \geqslant 0\}$ with initial condition $X_0 = \beta_0 \in \mathbb{R}$. For $\beta < \beta_0$, the first hitting time $T^X_\beta$ of $X_t$ to the set $(-\infty, \beta]$ is defined as $T^X_\beta = \inf\{t : X_t \leqslant \beta\}$.*

The first hitting time is the number of iterations that the stochastic process requires to reach the target level $\beta < \beta_0$ for the first time. In our situation, $X_t = \|m_t - x^*\|$ measures the distance from the current solution $m_t$ to the target point $x^*$ (typically, global or local optimal point) after $t$ iterations. Then, $\beta = \epsilon > 0$ defines the target accuracy and $T^X_\epsilon$ is the runtime of the algorithm until it finds an $\epsilon$-neighborhood $\mathcal{B}(x^*, \epsilon)$. The first hitting time $T^X_\epsilon$ is a random variable as $m_t$ is a random variable. In this paper, we focus on the *expected first hitting time* $\mathbb{E}[T^X_\epsilon]$. We want to derive lower and upper bounds on this expected hitting time that relate to the linear convergence of $X_t$ towards $x^*$. Such bounds take the following form: There exist $C_T, \tilde{C}_T \in \mathbb{R}$ and $C_R > 0$, $\tilde{C}_R > 0$ such that for any $0 < \epsilon \leqslant \beta_0$

$$\tilde{C}_T + \frac{\log\left(\|m_0 - x^*\|/\epsilon\right)}{\tilde{C}_R} \leqslant \mathbb{E}[T^X_\epsilon | \mathcal{F}_0] \leqslant C_T + \frac{\log(\|m_0 - x^*\|/\epsilon)}{C_R} \quad . \tag{3.1}$$

That is, the time to reach the target accuracy scales logarithmically with the ratio between the initial accuracy $\|m_0 - x^*\|$ and the target accuracy $\epsilon$. The first pair of constants, $C_T$ and $\tilde{C}_T$, capture the transient time, which is the time that adaptive algorithms typically spend for adaptation. The second pair of constants, $C_R$ and $\tilde{C}_R$, reflect the speed of convergence (logarithmic convergence rate). Intuitively, assuming that $C_R$ and $\tilde{C}_R$ are close, the distance to the optimum decreases in each step at a rate of approximately $\exp(-C_R) \approx \exp(-\tilde{C}_R)$. While upper-bounds inform us about the (linear) convergence, the lower-bound helps understanding whether the upper bounds are tight.

Alternatively, linear convergence can be defined as the following property: there exits $C > 0$ such that

$$\limsup_{t\to\infty} \frac{1}{t} \log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|} \leqslant -C \text{ almost surely.} \tag{3.2}$$

When we have an equality in the previous statement, we say that $\exp(-C)$ is the convergence rate.

Figure 2.1 (right plot) visualizes three different runs of the (1+1)-ES on a function with spherical level sets with different initial step-sizes. First of all, we clearly observe linear convergence. The first hitting time of $\mathcal{B}(x^*, \epsilon)$ scales linearly with $\log(1/\epsilon)$ for a sufficiently small $\epsilon > 0$. Second, its convergence speed is independent of the initial condition. Therefore, we expect to have universal constants $C_R$ and $\tilde{C}_R$ independent of the initial state. Last, depending on the initial step-size, the transient time can vary. If the initial step-size is too large or too small, it does not produce progress in terms of $\|m_t - x^*\|$ until the step-size is well adapted. Therefore, $C_T$ and $\tilde{C}_T$ depend on the initial condition, with a logarithmic dependency on the initial multiplicative mismatch.

**3.2. Bounds of the Hitting Time via Drift Conditions.** We are going to use *drift analysis* that consists in deducing properties of a sequence $\{X_t : t \geqslant 0\}$ (adapted to a natural filtration $\{\mathcal{F}_t : t \geqslant 0\}$) from its drift defined as $\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] - X_t$ [28]. Drift analysis has been widely used to analyze hitting times of evolutionary algorithms defined on discrete search spaces (mainly on binary search spaces) [10,18,19,31,32,45]. Though they were developed mainly for finite search spaces, the drift theorems can naturally be generalized to continuous domains [41,43]. Indeed, Jägersküpper's work [33,35,36] is based on the same idea, while the link to the drift analysis was implicit.

Since many drift conditions have been developed for analyzing algorithms on discrete domains, the domain of $X_t$ is often implicitly assumed to be bounded. However, this assumption is violated in our situation, where we will use $X_t = \log\big(f_\mu(m_t)\big)$ as the quality measure, which takes values in $\mathbb{R} \cup \{-\infty\}$, and is meant to approach $-\infty$. We refer to [3] for additional details. In general, translating expected progress requires bounding the tail of the progress distribution, as formalized in [28].

To control the tails of the drift distribution, we construct a stochastic process $\{Y_t : t \geqslant 0\}$ iteratively as follows: $Y_0 = X_0$ and

$$Y_{t+1} = Y_t + \max\big\{X_{t+1} - X_t, -A\big\} 1\{T_\beta^X > t\} - B 1\{T_\beta^X \leqslant t\} \tag{3.3}$$

for some $A \geqslant B > 0$ and $\beta < \beta_0$ with $X_0 = \beta_0$. It clips $X_{t+1} - X_t$ to some constant $-A$ ($A > 0$) from below. We introduce the indicator $1\{T_\beta^X > t\}$ for a technical reason. The process disregards progress larger than $A$, and it fixes the progress of the step that hits the target set to $B$. It is formalized in the following theorem, which is our main mathematical tool to derive an upper bound of the expected first hitting time of (1+1)-ES$_\kappa$ in the form of (3.1).

THEOREM 3.2. *Let $\{X_t : t \geqslant 0\}$ be a sequence of real-valued random variables adapted to a filtration $\{\mathcal{F}_t : t \geqslant 0\}$ with $X_0 = \beta_0 \in \mathbb{R}$. For $\beta < \beta_0$, let $T_\beta^X = \inf\{t : X_t \leqslant \beta\}$ be the first hitting time of the set $(-\infty, \beta]$. Define a stochastic process $\{Y_t : t \geqslant 0\}$ iteratively through (3.3) with $Y_0 = X_0$ for some $A \geqslant B > 0$, and let $T_\beta^Y = \inf\{t : Y_t \leqslant \beta\}$ be the first hitting time of the set $(-\infty, \beta]$. If $Y_t$ is integrable, i.e. $\mathbb{E}\big[|Y_t|\big] < \infty$, and*

$$\mathbb{E}\big[\max\{X_{t+1} - X_t, -A\} 1\{T_\beta^X > t\} \bigm| \mathcal{F}_t\big] \leqslant -B 1\{T_\beta^X > t\} \quad , \tag{3.4}$$

*then the expectation of $T_\beta^X$ satisfies*

$$\mathbb{E}\left[T_\beta^X\right] \leqslant \mathbb{E}\left[T_\beta^Y\right] \leqslant \frac{A + \beta_0 - \beta}{B} \quad . \tag{3.5}$$

*Proof of Theorem 3.2.* We consider the stopped process $\bar{X}_t = X_{\min\{t, T_\beta^X\}}$. We have $X_t \leqslant \bar{X}_t$ for $t \leqslant T_\beta^X$ and $\bar{X}_t \leqslant Y_{\min\{t, T_\beta^X\}}$ for all $t \geqslant 0$. Therefore, we have $T_\beta^X = T_\beta^{\bar{X}} \leqslant T_\beta^Y$. Let $\bar{Y}_t = Y_{\min\{t, T_\beta^Y\}}$. By construction it holds $Y_t \leqslant \bar{Y}_t$ for $t \leqslant T_\beta^Y$ and $T_\beta^Y = T_\beta^{\bar{Y}}$. Hence, $T_\beta^X \leqslant T_\beta^Y \leqslant T_\beta^{\bar{Y}}$.

We will prove that

$$\mathbb{E}[\bar{Y}_{t+1} \mid \mathcal{F}_t] \leqslant \bar{Y}_t - B 1\{T_\beta^Y > t\} \quad . \tag{3.6}$$

We start from

$$\mathbb{E}[\bar{Y}_{t+1} \mid \mathcal{F}_t] = \mathbb{E}[\bar{Y}_{t+1} 1\{T_\beta^Y \leqslant t\} \mid \mathcal{F}_t] + \mathbb{E}[\bar{Y}_{t+1} 1\{T_\beta^Y > t\} \mid \mathcal{F}_t] \tag{3.7}$$

and bound the different terms:

$$\mathbb{E}[\bar{Y}_{t+1}1\{T_\beta^Y \leqslant t\} \mid \mathcal{F}_t] = \mathbb{E}[\bar{Y}_t 1\{T_\beta^Y \leqslant t\} \mid \mathcal{F}_t] = \bar{Y}_t 1\{T_\beta^Y \leqslant t\} \tag{3.8}$$

where we have used that $1_{\{T_\beta^X > t\}}$, $Y_t$, $1_{\{T_\beta^Y > t\}}$, and $\bar{Y}_t$ are all $\mathcal{F}_t$-measurable. Also

$$\begin{aligned} \mathbb{E}[\bar{Y}_{t+1}1\{T_\beta^Y > t\} \mid \mathcal{F}_t] &= \mathbb{E}[Y_{t+1} \mid \mathcal{F}_t]1\{T_\beta^Y > t\} \\ &\leqslant (Y_t - B1\{T_\beta^X > t\} - B1\{T_\beta^X \leqslant t\})1\{T_\beta^Y > t\} = (\bar{Y}_t - B)1\{T_\beta^Y > t\} \ , \end{aligned} \tag{3.9}$$

where we have used condition (3.4). Hence, by injecting (3.8) and (3.9) into (3.7), we obtain (3.6). From (3.6), by taking the expectation we deduce $\mathbb{E}[\bar{Y}_{t+1}] \leqslant \mathbb{E}[\bar{Y}_t] - B\Pr[T_\beta^Y > t]$. Following the same approach as [43, Theorem 1], since $T_\beta^Y$ is a random variable taking values in $\mathbb{N}$, it can be rewritten as $\mathbb{E}[T_\beta^Y] = \sum_{t=0}^{+\infty}\Pr[T_\beta^Y > t]$ and thus it holds

$$B\mathbb{E}\left[T_\beta^Y\right] \xleftarrow{\tilde{t}\to\infty} \sum_{t=0}^{\tilde{t}} B\Pr\left[T_\beta^Y > t\right] \leqslant \sum_{t=0}^{\tilde{t}}\Big(\mathbb{E}[\bar{Y}_t] - \mathbb{E}[\bar{Y}_{t+1}]\Big) = \mathbb{E}[\bar{Y}_0] - \mathbb{E}[\bar{Y}_{\tilde{t}}] \ .$$

Since $Y_{t+1} \geqslant Y_t - A$, we have $Y_{T_\beta^Y} \geqslant \beta - A$. Given that $\bar{Y}_{\tilde{t}} \geqslant Y_{T_\beta^Y}$, we deduce that $E[\bar{Y}_{\tilde{t}}] \geqslant \beta - A$ for all $\tilde{t}$. With $\mathbb{E}[\bar{Y}_0] = \beta_0$, we have

$$\mathbb{E}\left[T_\beta^Y\right] \leqslant (A/B) + B^{-1}(\beta_0 - \beta) \ .$$

Since $\mathbb{E}[T_\beta^X] \leqslant \mathbb{E}[T_\beta^Y]$, this completes the proof. ∎

This theorem can be intuitively understood: we assume for the sake of simplicity a process $X_t$ such that $X_{t+1} \geqslant X_t - A$. Then (3.4) states that the process progresses in expectation by at least $-B$. The theorem concludes that the expected time needed to reach a value smaller than $\beta$ when started in $\beta_0$ equals to $(\beta_0 - \beta)/B$ (what we would get for a deterministic algorithm) plus $A/B$. This last term is due to the stochastic nature of the algorithm. It is minimized if $A$ is as close as possible to $B$, which corresponds to a highly concentrated process.

Jägersküpper [35, Theorem 2] established a general lower bound of the expected first hitting time of the (1+1)-ES. We borrow the same idea to prove the following general theorem for a lower bound of the expected first hitting time, which generalizes [36, Lemma 12]. See Theorem 2.3 in [3] for its proof.

Theorem 3.3. *Let $\{X_t : t \geqslant 0\}$ be a sequence of real-valued random variables adapted to a filtration $\{\mathcal{F}_t : t \geqslant 0\}$ and integrable such that $X_0 = \beta_0$, $X_{t+1} \leqslant X_t$, and $\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] - X_t \geqslant -C$ for $C > 0$. For $\beta < \beta_0$ we define $T_\beta^X = \min\{t : X_t \leqslant \beta\}$. Then the expected hitting time is lower bounded by $\mathbb{E}\left[T_\beta^X\right] \geqslant -(1/2) + (4C)^{-1}(\beta_0 - \beta)$.*

## 4. Main Result: Expected First Hitting Time Bound.

**4.1. Mathematical Modeling of the Algorithm.** In the sequel, we will analyze the process $\{\theta_t : t \geqslant 0\}$ where $\theta_t = (m_t, \sigma_t, \Sigma_t) \in \mathbb{R}^n \times \mathbb{R}_> \times \mathcal{S}_\kappa$ generated by the (1+1)-ES$_\kappa$ algorithm. We assume from now on that the optimized objective function $f$ is measurable with respect to the Borel $\sigma$-algebra. We equip the state-space $\mathcal{X} = \mathbb{R}^n \times \mathbb{R}_> \times \mathcal{S}_\kappa$ with its Borel $\sigma$-algebra denoted $\mathcal{B}(\mathcal{X})$.

**4.2. Preliminaries.** We present two preliminary results. In Assumption A1, we assume that for $m \in \mathcal{X}_a^b$, we can include a ball of radius $C_\ell f_\mu(m)$ into the sublevel set $S_0(m)$ and embed $S_0(m)$ into a ball of radius $C_u f_\mu(m)$. This allows us to upper bound and lower bound the probability of success for all $m \in \mathcal{X}_a^b$, for all $\Sigma \in \mathcal{S}_\kappa$, by the probability to sample inside of balls of radius $C_u f_\mu(m)$ and $C_\ell f_\mu(m)$ with appropriate center. From this we can upper-bound $p_{(a,b]}^{\mathrm{upper}}(\bar\sigma)$ by a function of $\bar\sigma$. Similarly we can lower-bound $p_{(a,b]}^{\mathrm{lower}}(\bar\sigma)$ by a function of $\bar\sigma$. The corresponding proof is given in Appendix B.3.

PROPOSITION 4.1. *Suppose that $f$ satisfies A1. Consider the lower and upper success probabilities $p_{(a,b]}^{\mathrm{upper}}$ and $p_{(a,b]}^{\mathrm{lower}}$ defined in Definition 2.5, then*

$$p_{(a,b]}^{\mathrm{upper}}(\bar\sigma) \leqslant \kappa^{d/2}\Phi\left(\bar{\mathcal{B}}\left(0, \frac{C_u}{\bar\sigma\kappa^{1/2}}\right); 0, \mathrm{I}\right) \tag{4.1}$$

$$p_{(a,b]}^{\mathrm{lower}}(\bar\sigma) \geqslant \kappa^{-d/2}\Phi\left(\bar{\mathcal{B}}\left(\frac{(2C_u - C_\ell)\kappa^{1/2}}{\bar\sigma}e_1, \frac{C_\ell\kappa^{1/2}}{\bar\sigma}\right); 0, \mathrm{I}\right) \quad , \tag{4.2}$$

*where $e_1 = (1, 0, \ldots, 0)$.*

We use the previous proposition to establish the next lemma that guarantees the existence of a finite range of normalized step-size that leads to the success probability into some range $(p_u, p_\ell)$ independent of $m$ and $\Sigma$, and provides a lower bound on the success probability with rate $r$ when the normalized step-size is in the above range. Its proof is provided in Appendix B.4.

LEMMA 4.2. *We assume that $f$ satisfies A1 and A2 for some $0 \leqslant a < b \leqslant \infty$. Then, for any $p_u$ and $p_\ell$ satisfying $0 < p_u < p^{\mathrm{target}} < p_\ell < p^{\mathrm{limit}}$, the constants*

$$\bar\sigma_\ell = \sup\left\{\bar\sigma > 0 : p_{(a,b]}^{\mathrm{lower}}(\bar\sigma) \geqslant p_\ell\right\} \quad \text{and} \quad \bar\sigma_u = \inf\left\{\bar\sigma > 0 : p_{(a,b]}^{\mathrm{upper}}(\bar\sigma) \leqslant p_u\right\}$$

*exist as positive finite values. Let $\ell \leqslant \bar\sigma_\ell$ and $u \geqslant \bar\sigma_u$ such that $u/\ell \geqslant \alpha_\uparrow/\alpha_\downarrow$. Then, for $r \in [0, 1]$, $p_r^*$ defined as*

$$p_r^* := \inf_{\ell \leqslant \bar\sigma \leqslant u} \inf_{m \in \mathcal{X}_a^b} \inf_{\Sigma \in \mathcal{S}_\kappa} p_r^{\mathrm{succ}}(\bar\sigma; m, \Sigma) \tag{4.3}$$

*is lower bounded by a positive number defined by*

$$\min_{\ell \leqslant \bar\sigma \leqslant u} \kappa^{-d/2}\Phi\left(\mathcal{B}\left(\left(\frac{(2C_u - (1-r)C_\ell)\kappa^{1/2}}{\bar\sigma}\right)e_1, \frac{(1-r)C_\ell\kappa^{1/2}}{\bar\sigma}\right); 0, \mathrm{I}\right) \quad . \tag{4.4}$$

**4.3. Potential Function.** Lemma 4.2 divides the domain of the normalized step-size into three disjoint subsets: $\bar\sigma \in (0, \ell)$ is a too small normalized step-size situation where we have $p_0^{\mathrm{succ}}(\bar\sigma; m, \Sigma) \geqslant p_\ell$ for all $m \in \mathcal{X}_a^b$ and $\Sigma \in \mathcal{S}_\kappa$; $\bar\sigma \in (u, \infty)$ is a too large normalized step-size situation where we have $p_0^{\mathrm{succ}}(\bar\sigma; m, \Sigma) \leqslant p_u$ for all $m \in \mathcal{X}_a^b$ and $\Sigma \in \mathcal{S}_\kappa$; and $\bar\sigma \in [\ell, u]$ is a reasonable normalized step-size situation where the success probability with rate $r$ is lower bounded by (4.4). Since $p_{\mathrm{target}} \in [p_u, p_\ell]$, the normalized step-size is supposed to be maintained in the reasonable range.

Our potential function is defined as follows. In light of Lemma 4.2, we can take $\ell \leqslant \bar\sigma_\ell$ and $u \geqslant \bar\sigma_u$ such that $u/\ell \geqslant \alpha_\uparrow/\alpha_\downarrow$. With some constant $v > 0$, we define our potential function as

$$V(\theta) = \log(f_\mu(m)) + \max\left\{0, v\log\left[\frac{\alpha_\uparrow \ell f_\mu(m)}{\sigma}\right], v\log\left[\frac{\sigma}{\alpha_\downarrow u f_\mu(m)}\right]\right\} \quad . \tag{4.5}$$

The rationale behind the second term on the right-hand side (RHS) is as follows. The second and third terms inside max are positive only if the normalized step-size $\bar{\sigma} = \sigma/f_\mu(m)$ is smaller than $\ell\alpha_\uparrow$ and greater than $u\alpha_\downarrow$, respectively. The potential value is $\log f_\mu(m)$ if the normalized step-size is in $[\ell\alpha_\uparrow, u\alpha_\downarrow]$ and it is penalized if the normalized step-size is too small or too large. We need this penalization for the following reason. If the normalized step-size is too small, the success probability is close to 1/2 for non-critical points, assuming $f = g \circ h$ where $h$ is a continuously differentiable function, but the progress per step is very small because the step-size directly controls the progress for instance measured as $\|m_{t+1} - m_t\| = \sigma_t \|\mathcal{N}(0, \Sigma_t)\| 1_{\{f(m_{t+1}) \leqslant f(m_t)\}}$. If the normalized step-size is too large, the success probability is close to zero and produces no progress with high probability. If we would use $\log f_\mu(m)$ as a potential function instead of $V(\theta)$ then the progress is arbitrarily small in such situations, which prevents the application of drift arguments. The above potential function penalizes such situations, and guarantees a certain progress in the penalized quantity since the step-size will be increased or decreased, respectively, with high probability, leading to a certain decrease of $V(\theta)$. We illustrate in Figure 2.1 that $\log(f_\mu(m))$ cannot work alone as a potential function while $V(\theta)$ does: when we start from a too small or too large step-size, $\log(f_\mu(m))$ looks constant (doted line in green and blue). Only when the step-size is started at 1, we see progress in $\log(f_\mu(m))$. Also, the step size can always get arbitrarily worse, with a very small probability, which forces us to handle the case of badly adapted step size properly. Yet the simulation of $V(\theta)$ shows that in all three situations (small, large and well adapted step-sizes compared to the distance to the optimum), we observe a geometric decrease of $V(\theta)$.

**4.4. Upper Bound of the First Hitting Time.** We are now ready to establish that the potential function defined in (4.5) satisfies a (truncated)-drift condition from Theorem 3.2. This will in turn imply an upper bound on the expected hitting time of $f_\mu(m)$ to $[0, \epsilon]$ provided $a \leqslant \epsilon$. The proof follows the same line of argumentation as the proof of [3, Proposition 4.2], which was restricted to the case of spherical functions. It was generalized under similar assumptions as in this paper, but for a fixed covariance matrix equal to the identity, in [46, Proposition 6]. The detailed proof is given in Appendix B.5.

PROPOSITION 4.3. *Consider the (1+1)-ES$_\kappa$ described in Algorithm 2.1 with state $\theta_t = (m_t, \sigma_t, \Sigma_t)$. Assume that the minimized objective function $f$ satisfies A1 and A2 for some $0 \leqslant a < b \leqslant \infty$. Let $p_u$ and $p_\ell$ be constants satisfying $0 < p_u < p_{\text{target}} < p_\ell < p^{\text{limit}}$ and $p_\ell + p_u = 2p_{\text{target}}$. Then, there exists $\ell \leqslant \bar{\sigma}_\ell$ and $u \geqslant \bar{\sigma}_u$ such that $u/\ell \geqslant \alpha_\uparrow/\alpha_\downarrow$, where $\bar{\sigma}_\ell$ and $\bar{\sigma}_u$ are defined in Lemma 4.2. For any $A > 0$, taking $v$ satisfying $0 < v < \min\left\{1, \frac{A}{\log(1/\alpha_\downarrow)}, \frac{A}{\log(\alpha_\uparrow)}\right\}$, and the potential function (4.5), we have*

$$\mathbb{E}\left[\max\{V(\theta_{t+1}) - V(\theta_t), -A\} 1\left\{m_t \in \mathcal{X}_a^b\right\} \mid \mathcal{F}_t\right] \leqslant -B 1\left\{m_t \in \mathcal{X}_a^b\right\} \tag{4.6}$$

*where*

$$B = \min\left\{A p_r^* - v \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right), \; v\frac{p_\ell - p_u}{2} \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right)\right\} \quad , \tag{4.7}$$

*and $p_r^* = \inf_{\bar{\sigma} \in [\ell, u]} \inf_{m \in \mathcal{X}_a^b} \inf_{\Sigma \in \mathcal{S}_\kappa} p_r^{\text{succ}}(\bar{\sigma}; m, \Sigma)$ with $r = 1 - \exp\left(-\frac{A}{1-v}\right)$. Moreover, for any $A > 0$ there exists $v$ such that $B < A$ is positive.*

We apply Theorem 3.2 along with Proposition 4.3 to derive the expected first hitting time bound. To do so, we need to confirm that it satisfies the prerequisite of the theorem: integrability of the process $\{Y_t : t \geqslant 0\}$ defined in (3.3) with $X_t = V(\theta_t)$.

LEMMA 4.4. *Let $\{\theta_t : t \geqslant 0\}$ be the sequence of parameters $\theta_t = (m_t, \sigma_t, \Sigma_t)$ defined by the (1+1)-ES$_\kappa$ with the initial condition $\theta_0 = (m_0, \sigma_0, \Sigma_0)$ optimizing a measurable function $f$. Set $X_t = V(\theta_t)$ as defined in (4.5) and define the process $Y_t$ as defined in Theorem 3.2. Then, for any $A > 0$, $\{Y_t : t \geqslant 0\}$ is integrable, i.e., $\mathbb{E}[|Y_t|] < \infty$ for each $t$.*

*Proof of Lemma 4.4.* The drift $Y_{t+1} = Y_t + \max\{V(\theta_{t+1}) - V(\theta_t), -A\} 1\{T_\beta^X > t\} - B 1\{T_\beta^X \leqslant t\}$ is by construction bounded by $-A$ from below. It is also bounded by a constant from above. Indeed, from the proof of Proposition 4.3, it is easy to find the upper bound, say $C$, of the truncated one-step change, $\Delta_t$ in the proof of Proposition 4.3, without using A1 and A2. Let $D = \max\{A, C\}$. Then, by recursion, $|V(\theta_t)| \leqslant |V(\theta_0)| + |V(\theta_t) - V(\theta_0)| \leqslant |Y_0| + Dt$. Hence $\mathbb{E}[|Y_t|] \leqslant |Y_0| + Dt < \infty$ for all $t$. ◻

Finally, we derive the expected first hitting time of $\log f_\mu(m_t)$.

THEOREM 4.5. *Consider the same situation as described in Proposition 4.3. Let $T_\epsilon = \min\{t : f_\mu(m_t) \leqslant \epsilon\}$ be the first hitting time of $f_\mu(m_t)$ to $[0, \epsilon]$. Choose $a \leqslant \epsilon < f_\mu(m_t) \leqslant b$, where $a$ and $b$ appear in Definition 2.5. If $m_0 \in \mathcal{X}_a^b$, the first hitting time is upper bounded by $\mathbb{E}[T_\epsilon] \leqslant \big(V(\theta_0) - \log(\epsilon) + A\big)/B$ for $A > B > 0$ described in Proposition 4.3, where $V(\theta)$ is the potential function defined in (4.5). Equivalently, we have $\mathbb{E}[T_\epsilon] \leqslant C_T + C_R^{-1} \log(f_\mu(m_0)/\epsilon)$, where*

$$C_T = \frac{A}{B} + \frac{v}{B} \max\left\{0, \log\left(\frac{\alpha_\uparrow \ell f_\mu(m_0)}{\sigma_0}\right), \log\left(\frac{\sigma_0}{\alpha_\downarrow u f_\mu(m_0)}\right)\right\} \quad , \qquad C_R = B \ .$$

*Moreover, the above result yields an upper bound of the expected first hitting time of $\|m_t - x^*\|$ to $[0, 2C_u\epsilon]$.*

*Proof.* Theorem 3.2 with Proposition 4.3 and Lemma 4.4 together bounds the expected first hitting time of $V(\theta_t)$ to $(-\infty, \log(\epsilon)]$ by $\big(V(\theta_0) - \log(\epsilon) + A\big)/B$. Since $\log f_\mu(m_t) \leqslant V(\theta_t)$, $T_\epsilon$ is bounded by the first hitting time of $V(\theta_t)$ to $(-\infty, \log(\epsilon)]$. The inequality is preserved if we take the expectation. The last claim is trivial from the inequality $\|x - x^*\| \leqslant 2C_u f_\mu(x)$, which holds under A1. ◻

Theorem 4.5 shows an upper bound on the expected hitting time of the (1+1)-ES$_\kappa$ with success-based step-size adaptation for linear convergence towards the global optimum $x^*$ on functions satisfying A1 and A2 with $a = 0$. Moreover, for $b = \infty$, this bound holds from all initial search points $m_0$. If $a > 0$, the bound in Theorem 4.5 does not translate into linear convergence, but we still obtain an upper bound on the expected first hitting time of the target accuracy $\epsilon \geqslant a$. This is useful for understanding the behavior of (1+1)-ES$_\kappa$ on multimodal functions, and on functions with degenerated Hessian matrix at the optimum.

**4.5. Lower Bound of the First Hitting Time.** We derive a general lower bound of the expected first hitting time of $\|m_t - x^*\|$ to $[0, \epsilon]$. The following results hold for an arbitrary measurable function $f$ and for a (1+1)-ES$_\kappa$ with an arbitrary $\sigma$-control mechanism. The following lemma provides the lower bound of the expected one-step progress measured by the logarithm of the distance to the optimum.

LEMMA 4.6. *We consider the process $\{\theta_t : t \geqslant 0\}$ generated by a (1+1)-ES$_\kappa$ algorithm with an arbitrary step-size adaptation mechanism and an arbitrary covariance* *matrix update optimizing an arbitrary measurable function $f$. We assume $d \geqslant 2$ and $\kappa_t = \mathrm{Cond}(\Sigma_t) \leqslant \kappa$. We consider the natural filtration $\mathcal{F}_t$. Then, the expected single-step progress is lower-bounded by*

$$\mathbb{E}[\min(\log(\|m_{t+1} - x^*\|/\|m_t - x^*\|), 0) \mid \mathcal{F}_t] \geqslant -\kappa_t^{\frac{d}{2}}/d \ . \tag{4.8}$$

*Proof of Lemma 4.6.* Note first that $\log(\|m_{t+1} - x^*\|/\|m_t - x^*\|) = \log(\|x_t - x^*\|/\|m_t - x^*\|)1\{_{f(x_t)\leqslant f(m_t)}\}$. This value can be positive since $f(x_t) \leqslant f(m_t)$ does not imply $\|x_t - x^*\| \leqslant \|m_t - x^*\|$ in general. Clipping the positive part to zero, we obtain a lower bound, which is the RHS of the above equality times the indicator $1\{_{\|x_t-x^*\|\leqslant\|m_t-x^*\|}\}$. Since the quantity is non-positive, dropping the indicator $1\{_{f(x_t)\leqslant f(m_t)}\}$ only decreases the lower bound. Hence, we have $\min(\log(\|m_{t+1} - x^*\|/\|m_t - x^*\|), 0) \geqslant \log(\|x_t - x^*\|/\|m_t - x^*\|)1\{_{\|x_t-x^*\|\leqslant\|m_t-x^*\|}\}$. Then,

$$\begin{aligned}&\mathbb{E}[\min(\log(\|m_{t+1} - x^*\|) - \log(\|m_t - x^*\|), 0) \mid \mathcal{F}_t] \\ &\qquad\qquad \geqslant \mathbb{E}[\log(\|x_t - x^*\|/\|m_t - x^*\|)1\{_{\|x_t-x^*\|\leqslant\|m_t-x^*\|}\} \mid \mathcal{F}_t] \ .\end{aligned}$$

We rewrite the lower bound of the drift. The RHS of the above inequality is the integral of $\log(\|x-x^*\|/\|m_t-x^*\|)$ in the integral domain $\bar{\mathcal{B}}(x^*, \|m_t - x^*\|)$ under the probability measure $\Phi\left(; m_t, \sigma_t^2\Sigma_t\right)$. Performing a variable change (through rotation and scaling) so that $m_t - x^*$ becomes $e_1 = (1, 0, \cdots, 0)$ and letting $\tilde{\sigma}_t = \sigma_t/\|m_t - x^*\|$, we can further rewrite it as the integral of $\log(\|x\|)$ in $\bar{\mathcal{B}}(0, 1)$ under $\Phi\left(; e_1, \tilde{\sigma}_t^2\Sigma_t\right)$. With $\kappa_t = \mathrm{Cond}(\Sigma_t)$, we have $\varphi\left(; e_1, \tilde{\sigma}_t^2\Sigma_t\right) \leqslant \kappa_t^{d/2}\varphi\left(; e_1, \kappa_t\tilde{\sigma}_t^2\mathrm{I}\right)$, see Lemma B.1. Altogether, we obtain the lower bound $\mathbb{E}[\log(\|x_t - x^*\|/\|m_t - x^*\|)1\{_{\|x_t-x^*\|\leqslant\|m_t-x^*\|}\} \mid \mathcal{F}_t] \geqslant \kappa_t^{d/2}\int_{\bar{\mathcal{B}}(0,1)}\log(\|x\|)\varphi\left(; e_1, \kappa_t\tilde{\sigma}_t^2\mathrm{I}\right)\mathrm{d}x$. The RHS is equivalent to $-\kappa_t^{d/2}$ times the single step progress of the (1+1)-ES on the spherical function at $m_t = e_1$ and $\sigma = \sqrt{\kappa}\tilde{\sigma}_t$, which is proven in the proof of Lemma 4.4 of [3] to be lower bounded by $1/d$ for $d \geqslant 2$. This completes the proof. ∎

The following theorem proves that the expected first hitting time of (1+1)-ES$_\kappa$ is $\Omega(\log(\|m_0 - x^*\|/\epsilon))$ for any measurable function $f$, implying that it can not converge faster than linearly. In case of $\kappa = 1$ the lower runtime bound becomes $\Omega(d(\log(\|m_0 - x^*\|/\epsilon)))$, meaning that the runtime scales linearly with respect to $d$. The proof is a direct application of Lemma 4.6 to Theorem 3.3.

THEOREM 4.7. *We consider the process $\{\theta_t : t \geqslant 0\}$ generated by a (1+1)-ES$_\kappa$ described in Algorithm 2.1 and assume that $f$ is a measurable function with $d \geqslant 2$. Let $T_\epsilon = \inf\{t : \|m_t - x^*\| \leqslant \epsilon\}$ be the first hitting time of $[0, \epsilon]$ by $\|m_t - x^*\|$. Then, the expected first hitting time is lower bounded by $\mathbb{E}[T_\epsilon] \geqslant -(1/2) + \frac{d}{4\kappa^{d/2}}\log(\|m_0 - x^*\|/\epsilon)$. The bound holds for arbitrary step-size adaptation mechanisms. If A1 holds, it gives a lower bound for the expected first hitting time bound of $f_\mu(m_t)$ to $[0, 2C_\ell\epsilon]$.*

*Proof of Theorem 4.7.* Let $X_t = \log\|m_t - x^*\|$ for $t \geqslant 0$. Define $Y_t$ iteratively as $Y_0 = X_0$ and $Y_{t+1} = Y_t + \min(X_{t+1} - X_t, 0)$. Then, it is easy to see that $Y_t \leqslant X_t$ and $Y_{t+1} \leqslant Y_t$ for all $t \geqslant 0$. Note that $\mathbb{E}[Y_{t+1} - Y_t \mid \mathcal{F}_t] = \mathbb{E}[\min(X_{t+1} - X_t, 0) \mid \mathcal{F}_t] = \mathbb{E}[\min(\log(\|m_{t+1} - x^*\|/\|m_t - x^*\|), 0) \mid \mathcal{F}_t]$, where the RMS is lower bounded in light of Lemma 4.6. Then, applying Theorem 3.3, we obtain the lower bound. The last statement directly follows from $\|x - x^*\| \leqslant 2C_\ell f_\mu(x)$ under A1. ∎

**4.6. Almost Sure Linear Convergence.** Additionally to the expected first hitting time bound, we can deduce from Proposition 4.3, almost sure linear convergence as stated in the following proposition.

Proposition 4.8. *Consider the same situation as described in Proposition 4.3, where $a = 0$ and $0 < b \leqslant \infty$. Then, for any $m_0 \in \mathcal{X}_0^b$, $\sigma_0 > 0$ and $\Sigma \in \mathcal{S}_\kappa$, we have*

$$\Pr\left[\limsup_{t\to\infty} \frac{1}{t}\log f_\mu(m_t) \leqslant -B\right] = \Pr\left[\limsup_{t\to\infty} \frac{1}{t}\log \|m_t - x^*\| \leqslant -B\right] = 1 \ , \tag{4.9}$$

*where $B > 0$ is as defined in Proposition 4.3. Hence almost sure linear convergence holds at a rate $\exp(-C)$ such that $\exp(-C) \leqslant \exp(-B)$.*

*Proof of Proposition 4.8.* Let $V$ be defined in (4.5). Let $Y_0 = V(\theta_0)$ and $Y_{t+1} = Y_t + \max(-A, V(\theta_{t+1}) - V(\theta_t))$. Define $Z_t = Y_t - \mathbb{E}_{t-1}[Y_t]$ for $t \geqslant 0$. Then, $\{Z_t\}$ is a martingale difference sequence on the filtration $\{\mathcal{F}_t\}$ produced by $\{\theta_t\}$. We hence have $\frac{1}{t}\log f_\mu(m_t) \leqslant \frac{1}{t}V(\theta_t) \leqslant \frac{1}{t}Y_t$, and from Proposition 4.3 we obtain

$$Y_t = \mathbb{E}_{t-1}[Y_t] + Z_t = Y_{t-1} + \mathbb{E}_{t-1}[Y_t - Y_{t-1}] + Z_t \leqslant Y_{t-1} - B + Z_t \ .$$

By repeatedly applying the above inequality and dividing it by $t$, we obtain $\frac{1}{t}Y_t \leqslant -B + \frac{1}{t}Y_0 + \frac{1}{t}\sum_{i=1}^{t} Z_i$, where $\lim_{t\to\infty}\frac{1}{t}Y_0 = 0$ and $\sum_{i=1}^{t} Z_i$ is a martingale sequence. In light of the strong law of large numbers for martingales [14], if $\sum_{t=1}^{\infty}\mathbb{E}[Z_t^2]/t^2 < \infty$, we have $\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t} Z_i = 0$ almost surely. By the definition of $V(\theta_t)$ and the working mechanism of the (1+1)-ES$_\kappa$, we have $V(\theta_i) - V(\theta_{i-1}) \leqslant v\log(\alpha_\uparrow/\alpha_\downarrow)$. Hence, $\mathbb{E}[Z_i^2] = \mathbb{E}[(Y_i - \mathbb{E}_{i-1}[Y_i])^2] = \mathbb{E}[\max(-A, V(\theta_i) - V(\theta_{i-1}))^2] \leqslant \max(A, v\log(\alpha_\uparrow/\alpha_\downarrow))^2$. Hence, we have $\limsup_{t\to\infty}\frac{1}{t}\log f_\mu(m_t) \leqslant -B + \lim_{t\to\infty}\frac{1}{t}Y_0 + \lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t} Z_i = -B$ almost surely. Along with $\|x - x^*\| \leqslant 2C_u f_\mu(x)$, we obtain Equation (4.9). ◻

**4.7. Wrap-up of the Results: Global Linear Convergence.** As a corollary to the lower-bound from Theorem 4.7, the upper bound from Theorem 4.5, Proposition 4.8 stating the almost sure linear convergence and the fact that different assumptions discussed in Section 2.3 imply A1 and A2, we summarize our linear convergence results in the following theorem.

Theorem 4.9 (Global Linear Convergence). *We consider the (1+1)-ES$_\kappa$ optimizing an objective function $f$. Suppose either*

*(a) $f$ satisfies A1 and A2 for $a = 0$, $p^{\text{limit}} > p^{\text{target}}$, and $m_0 \in \mathcal{X}_0^b$; or*

*(b) $f$ satisfies either A3 or A4, $p^{\text{target}} < 1/2$, and $m_0 \in \mathbb{R}^d$.*

*Then, for any $\sigma_0 > 0$ and $\Sigma_0 \in \mathcal{S}_\kappa$, the expected hitting time $\mathbb{E}[T_\epsilon]$ of $\|m_t - x^*\|$ to $[0, \epsilon]$ is $\Theta\big(\log(\|m_0 - x^*\|/\epsilon)\big)$ for all $\epsilon > 0$. Moreover, both $f_\mu(m_t)$ and $\|m_t - x^*\|$ linearly converge almost surely, i.e.*

$$\Pr\left[\limsup_{t\to\infty} \frac{1}{t}\log f_\mu(m_t) \leqslant -B\right] = \Pr\left[\limsup_{t\to\infty} \frac{1}{t}\log\|m_t - x^*\| \leqslant -B\right] = 1 \ ,$$

*where $B > 0$ is as defined in Proposition 4.3. The convergence rate $\exp(-C)$ is thus upper-bounded by $\exp(-B)$.*

**4.8. Tightness in the Sphere Function Case.** Now we consider a specific convex quadratic function, namely the sphere function $f(x) = \frac{1}{2}\|x\|^2$ where the spatial suboptimality function equals $f_\mu(x) = V_d\|x\|$. In Theorem 4.9 we have formulated that the expected hitting time of a ball of radius $\epsilon$ for the (1+1)-ES$_\kappa$ equals $\Theta(\log\|m_0 - x^*\|/\epsilon)$. Yet, this statement does not give information on how the constants hidden in the $\Theta$-notation scale with the dimension. In particular the convergence rate of the algorithm is upper-bounded by $\exp(-B)$ where $B$ is given in (4.7), see Theorem 4.5. In this section, we estimate precisely the scaling of $B$ in Proposition 4.3 with respect to the dimension and compare it with the general lower bound of the expected first hitting time given in Theorem 4.7. We then conclude that the bound is tight with respect to the scaling with $d$ in the case of the sphere function.

Let us assume $\kappa = 1$, that is, we consider the (1+1)-ES without covariance matrix adaptation ($\Sigma = I$). Then, $p^{\text{lower}}_{(a,b]}(\bar{\sigma}) = p^{\text{upper}}_{(a,b]}(\bar{\sigma}) = p^{\text{succ}}_{r}(\bar{\sigma}; m, \Sigma)$, where the right-most side is independent of $m$ and $\Sigma$ as described in Lemma 2.4. This means that the success probability is solely controlled by the normalized step-size $\bar{\sigma}$.

The following proposition states that the convergence speed is $\Omega(1/d)$, hence the expected first hitting time scales as $O(1/d)$. The proof is provided in Appendix B.6.

PROPOSITION 4.10. *For $A = 1/d$, $p_{\text{target}} \in \Theta(1)$ and $\log(\alpha_{\uparrow}/\alpha_{\downarrow}) \in \omega(1/d)$, we have $B \in \Omega(1/d)$.*

Two conditions on the choice of $\alpha_{\uparrow}$ and $\alpha_{\downarrow}$: $p_{\text{target}} = \log(1/\alpha_{\downarrow})/\log(\alpha_{\uparrow}/\alpha_{\downarrow}) \in \Theta(1)$ and $\log(\alpha_{\uparrow}/\alpha_{\downarrow}) \in \omega(1/d)$, are understood as follows. The first condition implies that the target success probability $p_{\text{target}}$ must be independent of $d$. In the 1/5 success rule, $\alpha_{\uparrow}$ and $\alpha_{\downarrow}$ are set so that $p_{\text{target}} = 1/5$ independent of $d$. The second condition implies that the factors of the step-size increase and decrease must be $\log(\alpha_{\uparrow}) \in \omega(1/d)$ and $\log(1/\alpha_{\downarrow}) \in \omega(1/d)$. Note that on the sphere function the normalized step-size $\bar{\sigma} \propto \sigma/\|m - x^*\|$ is kept around a constant during the search. It implies that the convergence speed of $\|m - x^*\|$ and $\sigma$ must agree. Therefore the speed of the adaptation of the step-size must not be too small to achieve $\Theta(d)$ scaling of the expected first hitting time.

Proposition 4.10 and Theorem 4.5 imply $\mathbb{E}[T_\epsilon] \in O(d \log(\|m_0\|/\epsilon))$ and Theorem 4.7 implies $\mathbb{E}[T_\epsilon] \in \Omega(d \log(\|m_0\|/\epsilon))$. They yield $\mathbb{E}[T_\epsilon] \in \Theta(d \log(\|m_0\|/\epsilon))$. This result shows i) that the runtime of the (1+1)-ES on the sphere function is proportional to $d$ as long as $\log(\alpha_{\uparrow}/\alpha_{\downarrow}) \in \omega(1/d)$, and ii) that from our methodology one can derive a tight bound of the runtime in some cases. The result is formally stated as follows.

THEOREM 4.11. *The (1+1)-ES (Algorithm 2.1) with $\kappa = 1$ and $p^{\text{target}} < 1/2$ converges globally and linearly in terms of $\log\|m_t - x^*\|$ from any starting point $m_0 \in \mathbb{R}^d$, $\sigma_0 > 0$, and $\Sigma_0 = I$ on any function $f(x) = g(\|x - x^*\|)$, where $g$ is a strictly increasing function. Moreover, if $p^{\text{target}} \in \Theta(1)$ and $\log(\alpha_{\uparrow}/\alpha_{\downarrow}) \in \omega(1/d)$, the expected first hitting time $T_\epsilon$ of $\log\|m_t - x^*\|$ to $(-\infty, \log(\epsilon)]$ is $\Theta(d \log(\|m_0\|/\epsilon))$ and the almost sure convergence rate is upper-bounded by $\exp(-\Theta(1/d))$.*

Since the lower bound holds for an arbitrary $\sigma$-adaptation mechanism, the above result not only implies that our upper bound is tight, but it also implies that the success-based $\sigma$-control mechanism achieves the best possible convergence rate except for a constant factor on the spherical function.

**5. Discussion.** We have established the almost sure global linear convergence of the (1+1)-ES$_\kappa$ and also expressed as a bound on the expected hitting time of an $\epsilon$-neighborhood of the solution. Assumption A1 has been the key to obtaining the expected first hitting time bound of (1+1)-ES$_\kappa$ in the form of (3.1). The convergence results hold on a wide class of functions. It includes

(i) strongly convex functions with Lipschitz gradient, where linear convergence of numerical optimization algorithm is usually analyzed,

(ii) continuously differentiable positively homogenous functions, where previous linear convergence results had been introduced, and

(iii) functions with non-smooth level sets as illustrated in Figure 2.2.

Because the analyzed algorithms are invariant to strictly monotonic transformations of the objective functions, *all results that hold on $f$ also hold on $g \circ f$ where $g : \mathrm{Im}(f) \to \mathbb{R}$*

*is a strictly increasing transformation, which can thus introduce discontinuities on the objective function.* In contrast to the previous result establishing the convergence of CMA-ES [17] by adding a step to enforce a sufficient decrease (which works well for direct search methods, but which is unnatural for ESs), we did not need to modify the adaptation mechanism of the (1+1)-ES to achieve our convergence proofs. We believe that this is crucial, since it allows our analysis to reflect the main mechanism that makes the algorithm work well in practice.

Theorem 4.11 proves that we can derive a tight convergence rate with Proposition 4.3 on the sphere function in the case where $\kappa = 1$, i.e., without covariance matrix adaptation. This partially supports the utility of our methodology. However, its derivation relies on the fact that both the level sets of the objective function and the equal-density curves of the sampling distribution are isotropic, and hence does not generalize immediately. Moreover, the lower bound (Theorem 4.7) seems to be loose even for $\kappa = 1$ on convex quadratic functions, where we empirically observe that the logarithmic convergence rate scales like $\Theta(1/\operatorname{Cond}(\nabla\nabla f))$, see Figure 2.1, while its dependency on the dimension is tight.

A better lower bound of the expected first hitting time and a handy way to estimate the convergence rate are relevant directions of future work. Further directions of future work are as follows:

Proving linear convergence of (1+1)-ES$_\kappa$ does not reveal the benefits of (1+1)-ES$_\kappa$ over the (1+1)-ES without covariance matrix adaptation. The motivation of the introduction of the covariance matrix is to improve the convergence rate and to broaden the class of functions on which linear convergence is exhibited. None of them are achieved in this paper.

On convex quadratic functions, we empirically observe that the covariance matrix approaches a stable distribution that is closely concentrated around the inverse Hessian up to a scalar factor, and the convergence speed on all convex quadratic functions is equal to that on the sphere function (see Figure 2.1). This behavior is not described by our result.

Covariance matrix adaptation is also important for optimizing functions with non-smooth level sets. On continuously differentiable functions, we can always set $\alpha_\uparrow$ and $\alpha_\downarrow$ so that $p = \frac{\log(1/\alpha_\downarrow)}{\log(\alpha_\uparrow/\alpha_\downarrow)} < p^{\mathrm{limit}} = 1/2$. This is the rationale behind the 1/5 success rule, where $p = 1/5$. Indeed, $p = 1/5$ is known to approximate the optimal situation on the sphere function where the expected one-step progress is maximized [50]. Therefore, one does not need to tune these parameters in a problem-specific manner. However, if the objective is not continuously differentiable and levelsets are non-smooth, then $p^{\mathrm{limit}}$ is in general smaller than 1/2. For example, it can be as low as $p^{\mathrm{limit}} = 1/2^d$ on $f(x) = \|x\|_\infty = \max_{i=1,\dots,n}|x_i|$. Without an appropriate adaptation of the covariance matrix the success probability will be smaller than $p = 1/5$ and one must tune $\alpha_\uparrow$ and $\alpha_\downarrow$ in order to converge to the optimum, which requires information about $p^{\mathrm{limit}}$. By adapting the covariance matrix appropriately, the success probability can be increased arbitrary close to 1/2 (by elongating steps in the direction of the success domain) and $\alpha_\uparrow$ and $\alpha_\downarrow$ do not require tuning.

To achieve a reasonable convergence rate bound and broaden the class of functions on which linear convergence is exhibited, one needs to find another potential function $V$ that may penalize a high condition number $\operatorname{Cond}(\nabla\nabla f(m_t)\Sigma_t)$ and replace the definitions of $p^{\mathrm{upper}}$ and $p^{\mathrm{lower}}$ accordingly. This point is left for future work.

**Acknowledgement.** We gratefully acknowledge support by Dagstuhl seminar 17191 “Theory of Randomized Search Heuristics”. We would like to thank Per Kristian Lehre, Carsten Witt, and Johannes Lengler for valuable discussions and advice on drift theory. Y. A. is supported by JSPS KAKENHI Grant Number 19H04179.

## REFERENCES

[1] M.A. Abramson, C. Audet, J.E. Dennis Jr, and S. Le Digabel, *OrthoMADS: A deterministic MADS instance with orthogonal directions*, SIAM Journal on Optimization, 20 (2009), pp. 948–966.

[2] Y. Akimoto, *Analysis of a natural gradient algorithm on monotonic convex-quadratic-composite functions*, in GECCO, 2012, pp. 1293–1300.

[3] Y. Akimoto, A. Auger, and T. Glasmachers, *Drift theory in continuous search spaces: expected hitting time of the (1+ 1)-ES with 1/5 success rule*, in GECCO, 2018, pp. 801–808.

[4] S. Alvernaz and J. Togelius, *Autoencoder-augmented neuroevolution for visual doom playing*, in IEEE CIG, 2017, pp. 1–8.

[5] D. V. Arnold and N. Hansen, *Active covariance matrix adaptation for the (1+1)-CMA-ES*, in GECCO, 2010, pp. 385–392.

[6] C. Audet and J.E. Dennis Jr, *Mesh adaptive direct search algorithms for constrained optimization*, SIAM Journal on Optimization, 17 (2006), pp. 188–217.

[7] A. Auger and N. Hansen, *Linear convergence on positively homogeneous functions of a comparison based step-size adaptive randomized search: the (1+1)-ES with generalized one-fifth success rule*, 2013, https://arxiv.org/abs/1310.8397.

[8] A. Auger and N. Hansen, *Linear convergence of comparison-based step-size adaptive randomized search via stability of Markov chains*, SIAM Journal on Optimization, 26 (2016), pp. 1589–1624.

[9] 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.

[10] B. Baritompa and M. Steel, *Bounds on absorption times of directionally biased random sequences*, Random Structures & Algorithms, 9 (1996), pp. 279–293.

[11] P. Bontrager, A. Roy, J. Togelius, N. Memon, and A. Ross, *DeepMasterPrints: Generating MasterPrints for Dictionary Attacks via Latent Variable Evolution*, in IEEE BTAS, 2018, pp. 1–9.

[12] S. Bubeck, *Convex optimization: Algorithms and complexity*, 2014, https://arxiv.org/abs/1405.4980.

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

[14] Y. S. Chow, *On a strong law of large numbers for martingales*, Ann. Math. Statist., 38 (1967), p. 610.

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

[16] L. Devroye, *The compound random search*, in International Symposium on Systems Engineering and Analysis, 1972, pp. 195–110.

[17] Y. Diouane, S. Gratton, and L. N. Vicente, *Globally convergent evolution strategies*, Mathematical Programming, 152 (2015), pp. 467–490.

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

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

[20] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, *Efficient decision-based black-box adversarial attacks on face recognition*, in CVPR, 2019.

[21] G. Fujii, M. Takahashi, and Y. Akimoto, *CMA-ES-based structural topology optimization using a level set boundary expression—application to optical and carpet cloaks*, Computer Methods in Applied Mechanics and Engineering, 332 (2018), pp. 624 – 643.

[22] T. Geijtenbeek, M. Van De Panne, and A. F. Van Der Stappen, *Flexible muscle-based locomotion for bipedal creatures*, ACM Transactions on Graphics (TOG), 32 (2013), pp. 1–11.

[23] T. Glasmachers, *Global convergence of the (1+1) Evolution Strategy to a critical point*, Evolutionary Computation, 28 (2020), pp. 27–53.

[24] D. Golovin, J. Karro, G. Kochanski, C. Lee, X. Song, and Q. Zhang, *Gradientless descent: High-dimensional zeroth-order optimization*, in ICLR, 2020.

[25] 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.
[26] S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang, *Complexity and global rates of trust-region methods based on probabilistic models*, IMA Journal of Numerical Analysis, 38 (2017), pp. 1579–1597.
[27] D. Ha and J. Schmidhuber, *Recurrent world models facilitate policy evolution*, in NeurIPS, 2018, pp. 2450–2462.
[28] B. Hajek, *Hitting-time and occupation-time bounds implied by drift analysis with applications*, Advances in Applied probability, 14 (1982), pp. 502–525.
[29] N. Hansen, A. Auger, R. Ros, S. Finck, and P. Pošík, *Comparing results of 31 algorithms from the black-box optimization benchmarking bbob-2009*, in GECCO, 2010, pp. 1689–1696.
[30] N. Hansen and A. Ostermeier, *Completely derandomized self-adaptation in evolution strategies*, Evolutionary Computation, 9 (2001), pp. 159–195.
[31] J. He and X. Yao, *Drift analysis and average time complexity of evolutionary algorithms*, Artificial intelligence, 127 (2001), pp. 57–85.
[32] J. He and X. Yao, *A study of drift analysis for estimating computation time of evolutionary algorithms*, Natural Computing, 3 (2004), pp. 21–35.
[33] J. Jägersküpper, *Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces*, Automata, Languages and Programming, 2003, pp. 188–188.
[34] J. Jägersküpper, *Rigorous runtime analysis of the (1+1)-ES: 1/5-rule and ellipsoidal fitness landscapes*, in FOGA, 2005, pp. 260–281.
[35] J. Jägersküpper, *How the (1+1)-ES using isotropic mutations minimizes positive definite quadratic forms*, Theoretical Computer Science, 361 (2006), pp. 38–56.
[36] J. Jägersküpper, *Algorithmic analysis of a basic evolutionary algorithm for continuous optimization*, Theoretical Computer Science, 379 (2007), pp. 329–347.
[37] S. Kern, S. D. Müller, N. Hansen, D. Büche, J. Ocenasek, and P. Koumoutsakos, *Learning probability distributions in continuous evolutionary algorithms–a comparative review*, Natural Computing, 3 (2004), pp. 77–112.
[38] J. Konečnỳ and P. Richtárik, *Simple complexity analysis of simplified direct search*, 2014, https://arxiv.org/abs/1410.0390.
[39] I. Kriest, V. Sauerland, S. Khatiwala, A. Srivastav, and A. Oschlies, *Calibrating a global three-dimensional biogeochemical ocean model (mops-1.0)*, Geoscientific Model Development, 10 (2017), p. 127.
[40] J. Larson, M. Menickelly, and S. M. Wild, *Derivative-free optimization methods*, Acta Numerica, 28 (2019), pp. 287–404.
[41] P. K. Lehre and C. Witt, *General drift analysis with tail bounds*, 2013, https://arxiv.org/abs/1307.2559.
[42] J. Lengler, *Drift analysis*, in Theory of Evolutionary Computation, Springer, 2020, pp. 89–131.
[43] J. Lengler and A. Steger, *Drift analysis and evolutionary algorithms revisited*, 2016, https://arxiv.org/abs/1608.03226.
[44] P. MacAlpine, S. Barrett, D. Urieli, V. Vu, and P. Stone, *Design and optimization of an omnidirectional humanoid walk: A winning approach at the RoboCup 2011 3D simulation competition*, in AAAI, 2012.
[45] B. Mitavskiy, J. Rowe, and C. Cannings, *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 (2009), pp. 243–284.
[46] D. Morinaga and Y. Akimoto, *Generalized drift analysis in continuous domain: linear convergence of (1+ 1)-ES on strongly convex functions with lipschitz continuous gradients*, in FOGA, 2019, pp. 13–24.
[47] A. Nemirovski, *Information-based complexity of convex programming*, Lecture Notes, (1995).
[48] Y. Nesterov, *Lectures on convex optimization*, vol. 137, Springer, 2018.
[49] C. Paquette and K. Scheinberg, *A stochastic line search method with convergence rate analysis*, 2018, https://arxiv.org/abs/1807.07994.
[50] I. Rechenberg, *Evolutionsstrategie: Optimierung technisher Systeme nach Prinzipien der biologischen Evolution*, Frommann-Holzboog, 1973.
[51] I. Rechenberg, *Evolutionsstrategie'94*, frommann-holzboog, 1994.
[52] L. M. Rios and N. V. Sahinidis, *Derivative-free optimization: a review of algorithms and comparison of software implementations*, Journal of Global Optimization, 56 (2013), pp. 1247–1293.
[53] M. Schumer and K. Steiglitz, *Adaptive step size random search*, Automatic Control, IEEE Transactions on, 13 (1968), pp. 270–276.
[54] S. U. Stich, C. L. Muller, and B. Gartner, *Optimization of convex functions with random*

*pursuit*, SIAM Journal on Optimization, 23 (2013), pp. 1284–1309.
[55] S. U. Stich, C. L. Müller, and B. Gärtner, *Variable metric random pursuit*, Mathematical Programming, 156 (2016), pp. 549–579.
[56] J. Uhlendorf, A. Miermont, T. Delaveau, G. Charvin, F. Fages, S. Bottani, G. Batt, and P. Hersen, *Long-term model predictive control of gene expression at the population and single-cell levels*, Proceedings of the National Academy of Sciences, 109 (2012), pp. 14271–14276.
[57] V. Volz, J. Schrum, J. Liu, S. M. Lucas, A. Smith, and S. Risi, *Evolving Mario levels in the latent space of a deep convolutional generative adversarial network*, in GECCO, 2018, pp. 221–228.

## Appendix A. Some Numerical Results.

We present experiments with five algorithms on two convex quadratic functions. We compare (1+1)-ES, (1+1)-CMA-ES, simplified direction search [38], random pursuit [54], and gradientless descent [24].

All algorithms were started at the initial search point $x_0 = \frac{1}{\sqrt{d}}(1, \dots, 1) \in \mathbb{R}^d$. We implemented the algorithms as follows, with their parameters tuned where necessary: The ES always uses the setting $\alpha_\uparrow = \exp(4/d)$ and $\alpha_\downarrow = \alpha_\uparrow^{-1/4}$ for step size adaptation. We set the constant $c$ in the sufficient decrease condition of Simplified Direction Search to $\frac{1}{10}$, and we employed the standard basis as well as the negatives of these vectors as candidate directions. In each iteration we looped over the set of directions in random order. Randomizing the order greatly boosted performance over a fixed order. Random Pursuit was implemented with a golden section line search in the range $[-2\sigma, 2\sigma]$ with a rather loose target precision of $\sigma/2$, where $\sigma$ is either the initial step size or the length of the previous step. For Gradientless Descent we used the initial step size as the maximal step size and defined a target precision of $10^{-10}$. This target is reached by the ES in all cases. The experiments are designed to demonstrate several different effects: (a) We perform all experiments in $d = 10$ and $d = 50$ dimensions to investigate dimension-dependent effects. (b) We investigate best-case performance by running the algorithms on the spherical function $\|x\|^2$, i.e., on the separable convex quadratic function with minimal condition number. The initial step size is set to $\sigma_0 = 1$. All algorithms have a budget of $100d$ function evaluations. (c) We investigate the dependency of the performance on initial parameter settings by repeating the same experiment as above, but with an initial step size of $\sigma_0 = \frac{1}{1000}$. All algorithms have a budget of $700d$ function evaluations. (d) We investigate the dependence on problem difficulty by running the algorithms on an ellipsoid problem with a moderate condition number of $\kappa_f = 100$. The eigenvalues of the Hessian are evenly distributed on a log-scale. We use $\sigma_0 = 1$ like in the first experiment. All algorithms have a budget of $500d$ function evaluations. The experimental results are presented in Figure A.1.

**Interpretation.** We observe only moderate dimension-dependent effects, besides the expected linear increase of the runtime. We see robust performance of the ES, in particular with covariance matrix adaptation. The second experiment demonstrates the practical importance of the ability to grow the step size: the ES is essentially unaffected by wrong initial parameter settings while the gradientless descent and the simplified direct search are (which can be understood directly from the algorithms themselves). This property does not show up in convergence rates and is therefore often (but not always) neglected in algorithm design. The last experiment clearly demonstrates the benefit of variable-metric methods like CMA-ES. It should be noted that variable metric techniques can be implemented into most existing algorithms. This is rarely done though, with random pursuit being a notable exception [55].

## Appendix B. Proofs.

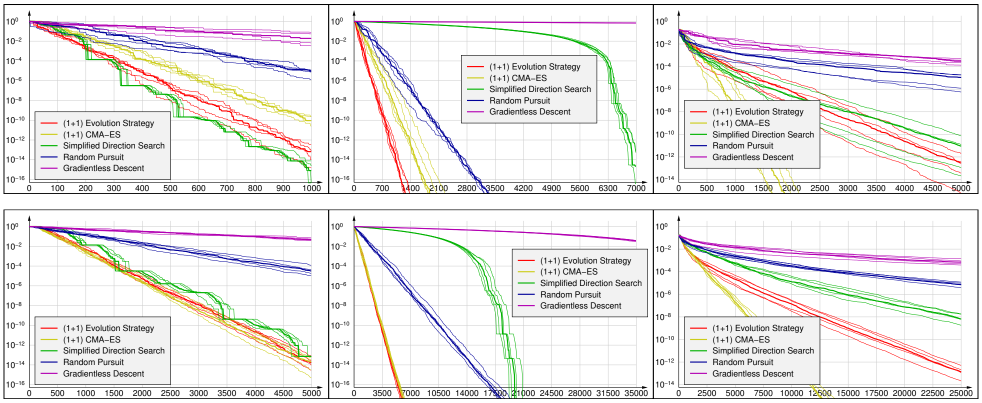


Fig. A.1: Comparison of (1+1)-ES with and without covariance matrix adaptation with three well-analyzed derivative-free optimization algorithms on two convex quadratic functions. The left column of plots shows the performance on the sphere function $\|x\|^2$ in dimensions 10 (top) and 50 (bottom). The middle column shows the same problem, but the initial step size is smaller by a factor of 1000 (and the horizontal axis differs), simulating that the distance to the optimum was under-estimated. The right column shows the performance on the ellipsoid function (defined in Figure 2.1). The plots show the evolution of the best-so-far function value (on a logarithmic scale), with five individual runs (thin curves) as well as median performance (bold curves).

**B.1. Proof of Lemma 2.8.** Since $f_\mu$ is invariant to $g$, without loss of generality we assume $f(x) = h(x) - h(x^*)$ in this proof. Inequality (2.7) implies that $f(y) \leqslant f(x) \Rightarrow (L_\ell/2)\|y - x^*\|^2 \leqslant f(x)$, meaning that $\{y : f(y) \leqslant f(x)\} \subseteq \bar{\mathcal{B}}\Big(x^*, \sqrt{\frac{f(x)}{L_\ell/2}}\Big)$. Since $f_\mu(x)$ is the $d$th root of the volume of the left-hand side of the above relation, we find $f_\mu(x) \leqslant \mu^{\frac{1}{d}}\Big(\bar{\mathcal{B}}\Big(x^*, \sqrt{\frac{f(x)}{L_\ell/2}}\Big)\Big) = V_d\sqrt{\frac{f(x)}{L_\ell/2}}$. Analogously, we obtain $\mathcal{B}\Big(x^*, \sqrt{\frac{f(x)}{L_u/2}}\Big) \subseteq \{y : f(y) < f(x)\}$ and $f_\mu(x) \geqslant V_d\sqrt{\frac{f(x)}{L_u/2}}$. From these inequalities, we obtain $\{y : f(y) \leqslant f(x)\} \subseteq \bar{\mathcal{B}}\Big(x^*, \sqrt{\frac{L_u}{L_\ell}}\frac{f_\mu(x)}{V_d}\Big)$ and $\mathcal{B}\Big(x^*, \sqrt{\frac{L_\ell}{L_u}}\frac{f_\mu(x)}{V_d}\Big) \subseteq \{y : f(y) < f(x)\}$. This implies A1 for $\mathcal{X}_0^\infty$. A2 is immediately implied by Proposition 2.7. This completes the proof.

**B.2. Proof of Lemma 2.9.** We first prove that A1 holds for $a = 0$ and $b = \infty$ with $C_u = \sup\{\|x - x^*\| : f_\mu(x) = 1\}$ and $C_\ell = \inf\{\|x - x^*\| : f_\mu(x) = 1\}$ and they are finite.

It is easy to see that the spatial suboptimality function $f_\mu(x)$ is proportional to $h(x) - h(x^*)$. Let $f_\mu(x) = c(h(x) - h(x^*))$ for some $c > 0$. Then, $f_\mu$ is also a homogeneous function. Since it is homogeneous, A1 reduces to that there are open and closed balls with radius $C_\ell$ and $C_u$ satisfying the conditions described in the assumption with $f_\mu(m) = 1$. Such constants are obtained by $C_u = \sup\{\|x - x^*\| : f_\mu(x) = 1\}$ and $C_\ell = \inf\{\|x - x^*\| : f_\mu(x) = 1\}$.

Due to the continuity of $f$ there exists an open ball $B$ around $x^*$ such that $h(x) < h(x^*) + 1/c$ for all $x \in B$. Then, it holds that $f_\mu(x) < 1$ for all $x \in B$. It implies that $C_\ell$ is no smaller than the radius of $B$, which is positive. Hence, $C_\ell > 0$.

We show the finiteness of $C_u$ by a contradiction argument. Suppose $C_u = \infty$. Then, there is a direction $v$ such that $f_\mu(x^* + Mv) \leqslant 1$ with an arbitrarily large $M > 0$. Since $f_\mu$ is homogeneous, we have $f_\mu(x^* + v) \leqslant 1/M$ and this must hold for any $M > 0$. This implies $f_\mu(x^* + v) = c(h(x) - h(x^*)) = 0$, which contradicts the assumption that $x^*$ is the unique global optimum. Hence, $C_u < \infty$.

The above argument proves that A1 holds with the above constants for $a = 0$ and $b = \infty$. Proposition 2.7 proves A2.

**B.3. Proof of Proposition 4.1.** For a given $m \in \mathcal{X}_a^b$, there is a closed ball $\bar{\mathcal{B}}_u$ such that $S_0(m) \subseteq \bar{\mathcal{B}}_u$, see Figure 2.2. We have

$$\begin{aligned} p_{(a,b]}^{\text{upper}}(\bar\sigma) &= \sup_{m\in\mathcal{X}_a^b} \sup_{\Sigma\in\mathcal{S}_\kappa} \int_{S_0(m)} \varphi\big(x; m, (f_\mu(m)\bar\sigma)^2\,\Sigma\big) dx \\ &\leqslant \sup_{m\in\mathcal{X}_a^b} \sup_{\Sigma\in\mathcal{S}_\kappa} \underbrace{\int_{\bar{\mathcal{B}}_u} \varphi\big(x; m, (f_\mu(m)\bar\sigma)^2\,\Sigma\big) dx}_{(*1)} \enspace . \end{aligned}$$

The integral is maximized if the ball is centered at $m$. By a variable change ($x \leftarrow x - m$),

$$\begin{aligned} (*1) &\leqslant \int_{\|x\|\leqslant C_u f_\mu(m)} \varphi\big(x; 0, (f_\mu(m)\bar\sigma)^2\,\Sigma\big) dx = \int_{\|x\|\leqslant C_u/\bar\sigma} \varphi(x; 0, \Sigma) dx \\ &\leqslant \kappa^{d/2}\Phi\left(\bar{\mathcal{B}}\left(0, \tfrac{C_u}{\bar\sigma\kappa^{1/2}}\right); 0, \mathrm{I}\right) \enspace . \end{aligned}$$

Here we used $\Phi\big(\bar{\mathcal{B}}(0, r)\big); 0, \Sigma) \leqslant \kappa^{d/2}\Phi\left(\bar{\mathcal{B}}\left(0, \kappa^{-1/2}r\right); 0, \mathrm{I}\right)$ for any $r > 0$, which is proven in Lemma B.1 below. The right-most side (RMS) of the above inequality is independent of $m$. It proves (4.1).

Similarly, there are balls $\mathcal{B}_\ell$ and $\bar{\mathcal{B}}_u$ such that $\mathcal{B}_\ell \subseteq S_0(m) \subseteq \bar{\mathcal{B}}_u$. We have

$$\begin{aligned} p_{(a,b]}^{\text{lower}}(\bar\sigma) &= \inf_{m\in\mathcal{X}_a^b} \inf_{\Sigma\in\mathcal{S}_\kappa} \int_{S_0(m)} \varphi\big(x; m, (f_\mu(m)\bar\sigma)^2\,\Sigma\big) dx \\ &\geqslant \inf_{m\in\mathcal{X}_a^b} \inf_{\Sigma\in\mathcal{S}_\kappa} \underbrace{\int_{\mathcal{B}_\ell} \varphi\big(x; m, (f_\mu(m)\bar\sigma)^2\,\Sigma\big) dx}_{(*2)} \enspace . \end{aligned}$$

The integral is minimized if the ball is at the opposite side of $m$ on the ball $\bar{\mathcal{B}}_u$, see Figure 2.2. By a variable change (moving $m$ to the origin) and letting $e_m = m/\|m\|$,

$$\begin{aligned} (*2) &\geqslant \int_{\|x - ((2C_u - C_\ell) f_\mu(m)) e_m\| \leqslant C_\ell f_\mu(m)} \varphi\big(x; 0, (f_\mu(m)\bar\sigma)^2\,\Sigma\big) dx \\ &= \int_{\|x - ((2C_u - C_\ell)/\bar\sigma) e_m\| \leqslant C_\ell/\bar\sigma} \varphi(x; 0, \Sigma) dx \\ &\geqslant \kappa^{-d/2}\Phi\left(\bar{\mathcal{B}}\left(\left(\tfrac{(2C_u - C_\ell)\kappa^{1/2}}{\bar\sigma}\right) e_m, \tfrac{C_\ell\kappa^{1/2}}{\bar\sigma}\right); 0, \mathrm{I}\right) \enspace . \end{aligned}$$

Here we used $\Phi\big(\bar{\mathcal{B}}(c, r); 0, \Sigma\big) \geqslant \kappa^{-d/2}\Phi\big(\bar{\mathcal{B}}(\kappa^{1/2}c, \kappa^{1/2}r); 0, \mathrm{I}\big)$ for any $c \in \mathbb{R}^d$ and $r > 0$ (Lemma B.1). The RMS of the above inequality is independent of $m$ as its value is constant over all unit vectors $e_m$. Replacing $e_m$ with $e_1$, we have (4.2).

LEMMA B.1. *For all* $\Sigma \in \mathcal{S}_\kappa$, $\kappa^{-d/2}\varphi\left(x; 0, \kappa^{-1}\mathrm{I}\right) \leqslant \varphi\big(x; 0, \Sigma\big) \leqslant \kappa^{d/2}\varphi\left(x; 0, \kappa\mathrm{I}\right)$ *and* $\kappa^{-d/2}\Phi\left(\mathcal{B}(\sqrt{\kappa}c, \sqrt{\kappa}r); 0, \mathrm{I}\right) \leqslant \Phi\big(\mathcal{B}(c, r); 0, \Sigma\big) \leqslant \kappa^{d/2}\Phi\left(\mathcal{B}(c/\sqrt{\kappa}, r/\sqrt{\kappa}); 0, \mathrm{I}\right)$.

*Proof.* For $\Sigma \in \mathcal{S}_\kappa$, we have $\det(\Sigma) = 1$ and $\mathrm{Cond}(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma) \leqslant \kappa$. Since $\det(\Sigma) = 1$ and $\det(\Sigma) = \prod_{i=1}^d \lambda_i(\Sigma)$, we have $\lambda_{\max}(\Sigma) \geqslant 1 \geqslant \lambda_{\min}(\Sigma)$. Therefore, we have $\lambda_{\min}(\Sigma) \geqslant \lambda_{\max}/\kappa \geqslant \kappa^{-1}$ and $\lambda_{\max}(\Sigma) \leqslant \kappa\lambda_{\min}(\lambda) \leqslant \kappa$. Then we obtain $\kappa^{-1}x^{\mathrm{T}}\mathrm{I}x \leqslant x^{\mathrm{T}}\Sigma^{-1}x \leqslant \kappa x^{\mathrm{T}}\mathrm{I}x$. With this inequality we have

$$\begin{aligned} \varphi\big(x; 0, \Sigma\big) = (2\pi)^{-d/2}\exp(-x^{\mathrm{T}}\Sigma^{-1}x/2) &\leqslant (2\pi)^{-d/2}\exp(-x^{\mathrm{T}}\mathrm{I}x/(2\kappa))) \\ &= \kappa^{d/2}(2\pi\kappa)^{-d/2}\exp(-x^{\mathrm{T}}\mathrm{I}x/(2\kappa))) = \kappa^{d/2}\varphi\big(x; 0, \kappa\mathrm{I}\big) \enspace . \end{aligned}$$

Analogously, we obtain $\varphi(x;0,\Sigma) \geqslant \kappa^{-d/2}\varphi(x;0,\kappa^{-1}\mathrm{I})$. Taking the integral over $\mathcal{B}(c,r)$, we obtain the second statement. ∎

**B.4. Proof of Lemma 4.2.** The upper bound of $p^{\mathrm{upper}}_{(a,b]}$ given in (4.1) is strictly decreasing in $\bar\sigma$ and converges to zero when $\bar\sigma$ goes to infinity. This guarantees the existence of $\bar\sigma_u$ as a finite value. The existence of $\bar\sigma_\ell > 0$ is obvious under A2. A1 guarantees that there exists an open ball $B_\ell$ with radius $C_\ell(1-r)f_\mu(m)$ such that $\mathcal{B}_\ell \subseteq \{x \in \mathbb{R}^d \mid f_\mu(x) < (1-r)f_\mu(m)\}$. Then, analogously to the proof of Proposition 4.1, the success probability with rate $r$ is lower bounded by

$$p^{\mathrm{succ}}_r(\bar\sigma; m, \Sigma) \geqslant \kappa^{-d/2}\Phi\left(\mathcal{B}\left(\left(\tfrac{(2C_u-(1-r)C_\ell)\kappa^{1/2}}{\bar\sigma}\right)e_1, \tfrac{(1-r)C_\ell\kappa^{1/2}}{\bar\sigma}\right); 0, \mathrm{I}\right). \tag{B.1}$$

The probability is independent of $m$, positive, and continuous in $\bar\sigma \in [\ell, u]$. Therefore the minimum is attained. This completes the proof.

**B.5. Proof of Proposition 4.3.** First, we remark that $m_t \in \mathcal{X}_{a,b}$ is equivalent to the condition $a < f_\mu(m_t) \leqslant b$. If $f_\mu(m_t) \leqslant a$ or $f_\mu(m_t) > b$, both sides of (4.6) are zero, hence the inequality is trivial. In the following we assume that $m_t \in \mathcal{X}^b_a$.

For the sake of simplicity we introduce $\log^+(x) = \log(x)1\{x\geqslant 1\}$. We rewrite the potential function as

$$V(\theta_t) = \log\left(f_\mu(m_t)\right) + v\log^+\left(\tfrac{\alpha_\uparrow \ell f_\mu(m_t)}{\sigma_t}\right) + v\log^+\left(\tfrac{\sigma_t}{\alpha_\downarrow u f_\mu(m_t)}\right) \ . \tag{B.2}$$

The potential function at time $t+1$ can be written as

$$\begin{aligned} V(\theta_{t+1}) = \log f_\mu(m_{t+1}) &+ \underbrace{v\log^+\tfrac{\ell f_\mu(m_{t+1})}{\sigma_t}1\{\sigma_{t+1}>\sigma_t\}}_{P_2} + \underbrace{v\log^+\tfrac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t}1\{\sigma_{t+1}<\sigma_t\}}_{P_3} \\ &+ \underbrace{v\log^+\tfrac{\alpha_\uparrow\sigma_t}{\alpha_\downarrow u f_\mu(m_{t+1})}1\{\sigma_{t+1}>\sigma_t\}}_{P_4} + \underbrace{v\log^+\tfrac{\sigma_t}{u f_\mu(m_t)}1\{\sigma_{t+1}<\sigma_t\}}_{P_5} \ . \end{aligned}$$

We want to estimate the conditional expectation

$$\mathbb{E}\left[\max\{V(\theta_{t+1}) - V(\theta_t)\,,\, -A\} \mid \theta_t\right]. \tag{B.3}$$

We partition the possible values of $\theta_t$ into three sets: first the set of $\theta_t$ such that $\sigma_t < \ell f_\mu(m_t)$ ($\sigma_t$ is small), second the set of $\theta_t$ such that $\sigma_t > u f_\mu(m_t)$ ($\sigma_t$ is large), and last the set of $\theta_t$ such that $\ell f_\mu(m_t) \leqslant \sigma_t \leqslant u f_\mu(m_t)$ (reasonable $\sigma_t$). In the following, we bound (B.3) for each of the three cases and in the end our bound $B$ will equal the minimum of the three bounds obtained for each case.

*Reasonable $\sigma_t$ case:* $\frac{f_\mu(m_t)}{\sigma_t} \in \left[\frac{1}{u}, \frac{1}{\ell}\right]$. In case of success, where $1\{\sigma_{t+1}>\sigma_t\} = 1$, we have $f_\mu(m_{t+1})/\sigma_{t+1} \leqslant f_\mu(m_t)/(\alpha_\uparrow\sigma_t) \leqslant 1/(\alpha_\uparrow\ell)$, implying that $P_2$ is always 0. Similarly, in case of failure, $f_\mu(m_{t+1})/\sigma_{t+1} = f_\mu(m_t)/(\alpha_\downarrow\sigma_t) \geqslant 1/(\alpha_\downarrow u)$ and we find that $P_5$ is always zero. We rearrange $P_3$ and $P_4$ into

$$\begin{aligned} P_3 &= v\log^+\left(\tfrac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow\sigma_t}\right)1\{\sigma_{t+1}<\sigma_t\} \ , \\ P_4 &= v\left[\log\left(\tfrac{\alpha_\uparrow\sigma_t}{\alpha_\downarrow u f_\mu(m_t)}\right) - \log\left(\tfrac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)\right]1\left\{\tfrac{\alpha_\downarrow u f_\mu(m_{t+1})}{\alpha_\uparrow\sigma_t}<1\right\}1\{\sigma_{t+1}>\sigma_t\} \ . \end{aligned}$$

Then, the one-step change $\Delta_t = V(\theta_{t+1}) - V(\theta_t)$ is upper bounded by

$$\text{(B.4)}\quad \Delta_t \leqslant \left(1 - v1\left\{\tfrac{\alpha_\downarrow u f_\mu(m_t)}{\alpha_\uparrow \sigma_t} < 1\right\} 1\left\{\sigma_{t+1} > \sigma_t\right\}\right) \log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)$$
$$+ v \log^+\left(\frac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t}\right) 1\left\{\sigma_{t+1} < \sigma_t\right\} + v \log^+\left(\frac{\alpha_\uparrow \sigma_t}{\alpha_\downarrow u f_\mu(m_t)}\right) 1\left\{\sigma_{t+1} > \sigma_t\right\}$$
$$\leqslant (1-v)\log\frac{f_\mu(m_{t+1})}{f_\mu(m_t)} + v\log^+\frac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} 1\left\{\sigma_{t+1} < \sigma_t\right\} + v\log^+\frac{\alpha_\uparrow \sigma_t}{\alpha_\downarrow u f_\mu(m_t)} 1\left\{\sigma_{t+1} > \sigma_t\right\} \quad .$$

The truncated one-step change $\max\{\Delta_t\,, -A\}$ is upper bounded by

$$\text{(B.5)}\quad \max\{\Delta_t\,, -A\} \leqslant (1-v)\max\left\{\log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)\,, -\frac{A}{1-v}\right\}$$
$$+ v\log^+\left(\frac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t}\right) 1\left\{\sigma_{t+1} < \sigma_t\right\} + v\log^+\left(\frac{\alpha_\uparrow \sigma_t}{\alpha_\downarrow u f_\mu(m_t)}\right) 1\left\{\sigma_{t+1} > \sigma_t\right\} \quad .$$

To consider the expectation of the above upper bound, we need to compute the expectation of the maximum of $\log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)$ and $-\frac{A}{1-v}$. Let $a \leqslant 0$ and $b \in \mathbb{R}$ then $\max(a,b) = a1\{a>b\} + b1\{a \leqslant b\} \leqslant b1\{a \leqslant b\}$. Applying this and taking the conditional expectation, a trivial upper bound for the conditional expectation of $\max\left\{\log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)\,, -\frac{A}{1-v}\right\}$ is $-\frac{A}{1-v}$ times the probability of $\log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)$ being no greater than $-\frac{A}{1-v}$. The latter condition is equivalent to $f_\mu(m_{t+1}) \leqslant (1-r) f_\mu(m_t)$ corresponding to successes with rate $r = 1 - \exp\left(-\frac{A}{1-v}\right)$ or better. That is,

$$\text{(B.6)}\qquad (1-v)\mathbb{E}\left[\max\left\{\log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)\,, -\frac{A}{1-v}\right\}\right] \leqslant -A p_r^{\text{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right) \quad .$$

Note also that the expected value of $1\{\sigma_{t+1} > \sigma_t\}$ is the success probability, namely, $p_0^{\text{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right)$. We obtain an upper bound for the conditional expectation of $\max\{\Delta_t\,, -A\}$ in the case of reasonable $\sigma_t$ as

$$\text{(B.7)}\quad \mathbb{E}\left[\max\{\Delta_t\,, -A\} | \theta_t\right] \leqslant -A p_r^{\text{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right)$$
$$+ \left(\log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) + \underbrace{\log\left(\frac{\ell f_\mu(m_t)}{\sigma_t}\right)}_{\leqslant 0}\right) v \left(1 - p_0^{\text{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right)\right)$$
$$+ \left(\log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) + \underbrace{\log\left(\frac{\sigma_t}{u f_\mu(m_t)}\right)}_{\leqslant 0}\right) v p_0^{\text{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right) \leqslant -A p_r^* + v \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) \quad .$$

*Small $\sigma_t$ case:* $\frac{f_\mu(m_t)}{\sigma_t} > \frac{1}{\ell}$. If $\ell f_\mu(m_t) > \sigma_t$, the 2nd summand in (B.2) is positive. Moreover, if $\sigma_{t+1} < \sigma_t$, we have $\ell f_\mu(m_{t+1}) = \ell f_\mu(m_t) > \sigma_t > \sigma_{t+1}$ and hence the 2nd summand in (B.2) is positive for $V(\theta_{t+1})$ as well. If $\sigma_{t+1} > \sigma_t$, any regime can happen. Then, $V(\theta_{t+1}) - V(\theta_t) =$

$$
\begin{aligned}
&= \log \tfrac{f_\mu(m_{t+1})}{f_\mu(m_t)} - v \log \tfrac{\alpha_\uparrow \ell f_\mu(m_t)}{\sigma_t} + v \log \tfrac{\ell f_\mu(m_{t+1})}{\sigma_t} 1\left\{\tfrac{\ell f_\mu(m_{t+1})}{\sigma_t} > 1\right\} 1\left\{\sigma_{t+1} > \sigma_t\right\} \\
&\quad + v \log \tfrac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} 1\left\{\tfrac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} > 1\right\} 1\left\{\sigma_{t+1} < \sigma_t\right\} \\
&\quad + v \log \tfrac{\alpha_\uparrow \sigma_t}{\alpha_\downarrow u f_\mu(m_{t+1})} 1\left\{\tfrac{\alpha_\downarrow u f_\mu(m_{t+1})}{\alpha_\uparrow \sigma_t} < 1\right\} 1\left\{\sigma_{t+1} > \sigma_t\right\} \\
&= \log\left(\tfrac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right) \left[1 + v\left(1\left\{\tfrac{\ell f_\mu(m_{t+1})}{\sigma_t} > 1\right\} - 1\left\{\tfrac{\alpha_\downarrow u f_\mu(m_{t+1})}{\alpha_\uparrow \sigma_t} < 1\right\}\right) 1\left\{\sigma_{t+1} > \sigma_t\right\}\right] \\
&\quad - v \log\left(\tfrac{\alpha_\downarrow u f_\mu(m_t)}{\alpha_\uparrow \sigma_t}\right) 1\left\{\tfrac{\alpha_\downarrow u f_\mu(m_{t+1})}{\alpha_\uparrow \sigma_t} < 1\right\} 1\left\{\sigma_{t+1} > \sigma_t\right\} \\
&\quad - v \log\left(\tfrac{\ell f_\mu(m_t)}{\sigma_t}\right) \left[1 - 1\left\{\tfrac{\ell f_\mu(m_{t+1})}{\sigma_t} > 1\right\} 1\left\{\sigma_{t+1} > \sigma_t\right\} - 1\left\{\tfrac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} > 1\right\} 1\left\{\sigma_{t+1} < \sigma_t\right\}\right] \\
&\quad - v\left(\log(\alpha_\uparrow) - \log\left(\tfrac{\alpha_\uparrow}{\alpha_\downarrow}\right) 1\left\{\tfrac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} > 1\right\} 1\left\{\sigma_{t+1} < \sigma_t\right\}\right) \quad .
\end{aligned}
$$

On the RMS of the above equality, the first term is guaranteed to be non-positive since $v \in (0,1)$. The second and third terms are non-positive as well since $\frac{\alpha_\downarrow u f_\mu(m_t)}{\alpha_\uparrow \sigma_t} > \frac{\alpha_\downarrow u}{\alpha_\uparrow \ell} \geqslant 1$ and $\frac{\ell f_\mu(m_t)}{\sigma_t} > 1$. Replacing the indicator $1\left\{\frac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} > 1\right\}$ with 1 in the last term provides an upper bound. Altogether, we obtain

$$
\Delta_t = V(\theta_{t+1}) - V(\theta_t) \leqslant -v\left(\log(\alpha_\uparrow) - \log(\alpha_\uparrow/\alpha_\downarrow) 1\left\{\sigma_{t+1} < \sigma_t\right\}\right) \quad .
$$

Note that the RHS is larger than $-A$ since it is lower bounded by $-v\log(\alpha_\uparrow)$ and $v \leqslant A/\log(\alpha_\uparrow)$. Then, the conditional expectation of $\max\{\Delta_t\,, -A\}$ is

$$
\begin{aligned}
\text{(B.8)} \quad & \mathbb{E}\left[\max\{\Delta_t\,, -A\} | \mathcal{F}_t\right] \leqslant -v\left(\log\left(\tfrac{\alpha_\uparrow}{\alpha_\downarrow}\right) p_0^{\text{succ}}\left(\tfrac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right) + \log(\alpha_\downarrow)\right) \\
& \leqslant -v\left(\log\left(\tfrac{\alpha_\uparrow}{\alpha_\downarrow}\right) p_\ell + \log(\alpha_\downarrow)\right) = -v \log\left(\tfrac{\alpha_\uparrow}{\alpha_\downarrow}\right)(p_\ell - p_{\text{target}}) = -v \tfrac{p_\ell - p_u}{2} \log\left(\tfrac{\alpha_\uparrow}{\alpha_\downarrow}\right) \quad .
\end{aligned}
$$

Here we used $\mathbb{E}[1\{\sigma_{t+1} < \sigma_t\} \mid \mathcal{F}_t] = 1 - p_0^{\text{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right)$ for the first inequality, $p_0^{\text{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right) > p_\ell$ for the second inequality, and $p_{\text{target}} = \log\left(\frac{1}{\alpha_\downarrow}\right) / \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) = (p_u + p_\ell)/2$ for the last equality.

*Large $\sigma_t$ case:* $\frac{f_\mu(m_t)}{\sigma_t} < \frac{1}{u}$. Since $\frac{f_\mu(m_{t+1})}{\sigma_{t+1}} \leqslant \frac{f_\mu(m_t)}{\alpha_\downarrow \sigma_t} < \frac{1}{\alpha_\downarrow u}$, the 3rd summand in (B.2) is positive in both $V(\theta_t)$ and $V(\theta_{t+1})$. For the 2nd summand in (B.2), recall that $\alpha_\uparrow \ell f_\mu(m_t)/\sigma_t < \alpha_\uparrow \ell/u \leqslant \alpha_\downarrow < 1$ since we have assumed that $u/\ell \geqslant \alpha_\uparrow/\alpha_\downarrow$. Hence, for $V(\theta_t)$ the 2nd summand in (B.2) is zero. Also, $\alpha_\uparrow \ell \|m_{t+1}\|/\sigma_{t+1} \leqslant \alpha_\uparrow \ell/(\alpha_\downarrow u) = (\alpha_\uparrow/\alpha_\downarrow)\ell/u \geqslant 1$ and thus for $V(\theta_{t+1})$ the 2nd summand in (B.2) also equals 0. We obtain

$$
V(\theta_{t+1}) - V(\theta_t) = (1 - v)\big(\log\left(f_\mu(m_{t+1})\right) - \log\left(f_\mu(m_t)\right)\big) + v \log\left(\sigma_{t+1}/\sigma_t\right).
$$

The first term on the RHS is guaranteed to be non-positive since $v < 1$, yielding $\Delta_t \leqslant v\log(\sigma_{t+1}/\sigma_t)$. On the other hand,

$$
\begin{aligned}
v \log(\sigma_{t+1}/\sigma_t) &= v\left(\log(\alpha_\uparrow) 1\left\{\sigma_{t+1} > \sigma_t\right\} + \log(\alpha_\downarrow) 1\left\{\sigma_{t+1} < \sigma_t\right\}\right) \\
&= v\left(\log(\alpha_\uparrow/\alpha_\downarrow) 1\left\{\sigma_{t+1} > \sigma_t\right\} - \log(1/\alpha_\downarrow)\right) \\
&\geqslant -v\log(1/\alpha_\downarrow) \geqslant -A \ ,
\end{aligned}
$$

where the last inequality comes from the prerequisite $v \leqslant A/\log(1/\alpha_{\downarrow})$. Hence,

$$\max\{\Delta_t\,,\,-A\} \leqslant \max\{v\log(\sigma_{t+1}/\sigma_t), -A\} = v\log(\sigma_{t+1}/\sigma_t)\ .$$

Then, the conditional expectation of $\max\{\Delta_t\,,\,-A\}$ is

$$\text{(B.9)}\quad \mathbb{E}\left[\max\{\Delta_t\,,\,-A\}|\theta_t\right] \leqslant v\left(\log(\alpha_{\downarrow}) + \log\left(\frac{\alpha_{\uparrow}}{\alpha_{\downarrow}}\right)p_0^{\mathrm{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right)\right)$$

$$\leqslant v\left(\log(\alpha_{\downarrow}) + \log\left(\frac{\alpha_{\uparrow}}{\alpha_{\downarrow}}\right)p_u\right) = v\log\left(\frac{\alpha_{\uparrow}}{\alpha_{\downarrow}}\right)(-p_{\mathrm{target}} + p_u) = -v\frac{p_\ell - p_u}{2}\log\left(\frac{\alpha_{\uparrow}}{\alpha_{\downarrow}}\right)\ .$$

Here we used $p_0^{\mathrm{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right) \leqslant p_u$.

*Conclusion.* Inequalities (B.7)–(B.9) together cover all possible cases and we hence obtain (4.7).

Finally, we prove the positivity of $B$ for an arbitrary $A > 0$. Lemma 4.2 guarantees the positivity of $p_r^*$ for any choice of $A$ since $r = 1 - \exp(-A/(1 - v)) \in (0, 1)$ for any $A > 0$ and $v < 1$. Therefore, $Ap_r^* > 0$ for any $A$ and $v \leqslant \min(1,\ A/\log(1/\alpha_{\downarrow}),\ A/\log(\alpha_{\uparrow}))$. Moreover, for a sufficiently small $v$, $p_r^*$ is strictly positive for any $A > 0$. Therefore, one can take a sufficiently small $v$ that satisfies $Ap_r^* > v\log(\alpha_{\uparrow}/\alpha_{\downarrow})$. The first term in the minimum in (4.7) is positive. The second term therein is clearly positive for $v > 0$. This completes the proof.

**B.6. Proof of Proposition 4.10.** Consider $d \geqslant 2$. We set $A = 1/d$. We bound $B$ from below by taking a specific value for $v \in (0,\ \min(1, A/\log(1/\alpha_{\downarrow}),\ A/\log(\alpha_{\uparrow}))$ instead of considering sup for $v$. Our candidate is $v = \frac{Ap'}{\log(\alpha_{\uparrow}/\alpha_{\downarrow})}\frac{2}{(2+p_\ell - p_u)}$, where $p' = \inf_{\bar\sigma\in[\ell,u]} p_{r'}(\bar\sigma)$ and $r' = 1 - \exp\left(-A\left(1 - \frac{1}{d\log(\alpha_{\uparrow}/\alpha_{\downarrow})}\right)^{-1}\right)$. It holds $v < \frac{1}{d\log(\alpha_{\uparrow}/\alpha_{\downarrow})}$ and hence $r' > r$, from which we obtain $p' < p^*$.

We bound the terms in (4.7) as: $Ap^* - v\log(\alpha_{\uparrow}/\alpha_{\downarrow}) = \frac{p'}{d}\left(\frac{p^*}{p'} - \frac{2}{2+p_\ell-p_u}\right) \geqslant \frac{p'}{d}\left(\frac{p_\ell - p_u}{2+p_\ell-p_u}\right)$ and $v\frac{p_\ell-p_u}{2}\log\left(\frac{\alpha_{\uparrow}}{\alpha_{\downarrow}}\right) = \frac{p'}{d}\frac{p_\ell-p_u}{2+p_\ell-p_u}$. Therefore, we have $B \geqslant \frac{p'}{d}\frac{p_\ell-p_u}{2+p_\ell-p_u}$. Note that one can take $p_\ell - p_u \in \Theta(1)$ since the only condition is $p_{\mathrm{target}} = (p_\ell + p_u)/2 \in \Theta(1)$. To obtain $B \in \Omega(1/d)$, it is sufficient to show $p' \in \Theta(1)$ for $d \to \infty$.

Fix $p_\ell$ and $p_u$ independently of $d$. In the light of Lemma 3.1 in [3], we have that $p_0 : \mathbb{R}_> \to (0, 1/2)$ is continuous and strictly decreasing from $1/2$ to $0$ for all $d \in \mathbb{N}$. Therefore, for each $d \in \mathbb{N}$ there exists an inverse map $p_0^{-1} : (0, 1/2) \to \mathbb{R}_>$. Define $\hat\sigma_\ell^d = dV_d p_0^{-1}(p_\ell)$ and $\hat\sigma_u^d = dV_d p_0^{-1}(p_u)$ for each $d \in \mathbb{N}$. It follows from Lemma 3.2 in [3] that $p_0^{\lim} : \bar\sigma \mapsto \lim_{d\to\infty} p_0(\bar\sigma)$ is also strictly decreasing, hence invertible. The existence of $\lim_{d\to\infty} p_0(\cdot)$ is also proved in [3]. We let $\hat\sigma_\ell^\infty = (p_0^{\lim})^{-1}(p_\ell)$ and $\hat\sigma_u^\infty = (p_0^{\lim})^{-1}(p_u)$. Because of the pointwise convergence of $p_0(\bar\sigma = \hat\sigma/(dV_d))$ to $p_0^{\lim}(\hat\sigma)$, we have $\hat\sigma_\ell^d \to \hat\sigma_\ell^\infty$ and $\hat\sigma_u^d \to \hat\sigma_u^\infty$ for $d \to \infty$. Hence, for any $\hat u > \hat\sigma_u^\infty$ and $\hat\ell < \hat\sigma_\ell^\infty$ with $u/\ell \geqslant \alpha_{\uparrow}/\alpha_{\downarrow}$, there exists $D \in \mathbb{N}$ such that for all $d \geqslant D$ we have $\hat u > \hat\sigma_u^d$ and $\hat\ell < \hat\sigma_\ell^d$. Now we fix $\hat u$ and $\hat\ell$ in this way. This amounts to selecting $u = dV_d\hat u$ and $\ell = dV_d\hat\ell$.

We have $\lim_{d\to\infty} dr' = 1$ since $\lim_{d\to\infty} d\log(\alpha_{\uparrow}/\alpha_{\downarrow}) = \infty$ and hence according to Lemma 3.2 in [3] we have

$$\liminf_{d\to\infty} p' = \liminf_{d\to\infty} \min_{\bar\sigma\in[\ell,u]}\{p_{r'}(\bar\sigma)\} = \liminf_{d\to\infty}\min_{\hat\sigma\in[\hat\ell,\hat u]} p_{r'}\left(\frac{\hat\sigma}{dV_d}\right)$$

$$\overset{(\star)}{=} \min_{\hat\sigma\in[\hat\ell,\hat u]}\lim_{d\to\infty}\left(p_{r'}\left(\frac{\hat\sigma}{dV_d}\right)\right) = \min_{\hat\sigma\in[\hat\ell,\hat u]}\Psi\left(-\frac{1}{\hat\sigma} - \frac{\hat\sigma}{2}\right)\ ,$$

where the equality $(\star)$ follows from the pointwise convergence of $p_{r'}$ to $\lim_{d\to\infty} p_{r'}$ and the continuity of $p_{r'}$ and $\lim_{d\to\infty} p_{r'}$.[2] This completes the proof.

[2]Let $\{f_n : n \geqslant 1\}$ be a sequence of continuous functions on $\mathbb{R}$ and $f$ be a continuous function such that $f$ is the pointwise limit $\lim_n f_n(x) = f(x)$ of the sequence. Since they are continuous, there exist the minimizers of $f_n$ and $f$ in a compact set $[\ell, u]$. Let $x_n = \operatorname{argmin} f_n(x)$ and $x^* = \operatorname{argmin} f(x)$, where argmin is taken over $x \in [\ell, u]$ and we pick one if there exist more than one minimizers. It is easy to see that $f_n(x_n) \leqslant f_n(x^*)$, hence $\liminf_n f_n(x_n) \leqslant \liminf_n f_n(x^*) = f(x^*)$. Let $\{n_i : i \geqslant 1\}$ be the sub-sequence of the indices such that $\liminf_n f_n(x_n) = \lim_i f_{n_i}(x_{n_i})$. Since $\{x_{n_i} : i \geqslant 1\}$ is a bounded sequence, Bolzano-Weirstraß theorem provides a convergent sub-sequence $\{x_{n_{i_k}} : k \geqslant 1\}$ and we denote its limit as $x_*$. Of course we have $\liminf_n f_n(x_n) = \lim_k f_{n_{i_k}}(x_{n_{i_k}})$. Due to the continuity of $\{f_n : n \geqslant 1\}$ and the pointwise convergence to $f$, we have $\lim_k f_{n_{i_k}}(x_{n_{i_k}}) = \lim_k f_{n_{i_k}}(x_*) = f(x_*)$. Therefore, $\liminf_n f_n(x_n) = f(x_*) \leqslant f(x^*)$. Since $x^*$ is the minimizer of $f$ in $[\ell, u]$ and $x_* \in [\ell, u]$, it must hold $f(x_*) \geqslant f(x^*)$. Hence, $\liminf_n f_n(x_n) = f(x^*)$.