



HAL
open science

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

Youhei Akimoto, Anne Auger, Tobias Glasmachers, Daiki Morinaga

► **To cite this version:**

Youhei Akimoto, Anne Auger, Tobias Glasmachers, Daiki Morinaga. Global Linear Convergence of Evolution Strategies on More Than Smooth Strongly Convex Functions. *SIAM Journal on Optimization*, 2022. hal-02941429v3

HAL Id: hal-02941429

<https://inria.hal.science/hal-02941429v3>

Submitted on 20 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **GLOBAL LINEAR CONVERGENCE OF EVOLUTION STRATEGIES**
2 **ON MORE THAN SMOOTH STRONGLY CONVEX FUNCTIONS**

3 YOUHEI AKIMOTO ^{*}, ANNE AUGER [†], TOBIAS GLASMACHERS [‡], AND DAIKI
4 MORINAGA [§]

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

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

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

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

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

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

42 perspective.

43 We focus on the arguably simplest and oldest adaptive ES, denoted (1+1)-ES. It
 44 samples a candidate solution from a Gaussian distribution whose step-size (standard
 45 deviation) is adapted. The candidate solution is accepted if and only if it is better
 46 than the current one (see pseudo-code [Algorithm 2.1](#)). The algorithm shares some
 47 similarities with simplified direct search whose complexity analysis has been presented
 48 in [38]. Yet the (1+1)-ES is comparison-based and thus invariant to strictly increasing
 49 transformations of the objective function. Simplified direct search can be thought of as
 50 a variant of mesh adaptive direct search [1, 6]. Arguably, in contrast to direct search,
 51 a sufficient decrease condition cannot be guaranteed. This causes some difficulties
 52 for the analysis. The (1+1)-ES is rotational invariant, while direct search candidate
 53 solutions are created along a predefined set of vectors. While the CMA-ES should
 54 always be preferred for practical applications over the (1+1)-ES variant analyzed here,
 55 this latter variant achieves faster linear convergence on well-conditioned problems
 56 when compared to algorithms with established complexity analysis (see [54, Table 6.3
 57 and Figure 6.1] and [8, Figure B.4] where the random pursuit algorithm and the
 58 (1+1)-ES algorithms are compared, and also Appendix A).

59 Prior theoretical studies of the (1+1)-ES with $1/5$ success rule have established
 60 the global linear convergence on differentiable positively homogeneous functions (com-
 61 posed with a strictly increasing function) with a single optimum [7, 8]. Those results
 62 establish the almost sure linear convergence from all initial states. They however
 63 do not provide the dependency of the convergence rate with respect to the dimen-
 64 sion. A more specific study on the sphere function $f(x) = \frac{1}{2}\|x\|^2$ establishes lower
 65 and upper bounds on the expected hitting time of an ϵ -ball of the optimum in
 66 $\Theta(\log(d\|m_0 - x^*\|/\epsilon))$, where x^* is the optimum of the function, m_0 is the initial
 67 solution, and d is the problem dimension [3]. Prior to that, a variant of the (1+1)-ES
 68 with one-fifth success rule had been analyzed on the sphere and certain convex qua-
 69adratic functions establishing bounds on the expected hitting time with overwhelming
 70 probability in $\Theta(\log(\kappa_f d\|m_0 - x^*\|/\epsilon))$, where κ_f is the condition number (the ra-
 71 tio between the greatest and smallest eigenvalues) of the Hessian [33–36]. Recently,
 72 the class of functions where the convergence of the (1+1)-ES was proven has been
 73 extended to continuously differentiable functions. This analysis does not address the
 74 question of linear convergence, focusing only on convergence as such, which is possibly
 75 sublinear [23].

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

87 **Notation.** Throughout the paper, we use the following notations. The set of
 88 natural numbers $\{1, 2, \dots\}$ is denoted \mathbb{N} . Open, closed, and left open intervals on \mathbb{R}
 89 are denoted by (\cdot) , $[\cdot]$, and $(\cdot]$, respectively. The set of strictly positive real numbers
 90 is denoted by $\mathbb{R}_{>}$. The Euclidean norm on \mathbb{R}^d is denoted by $\|\cdot\|$. Open and closed

91 balls with center c and radius r are denoted as $\mathcal{B}(c, r) = \{x \in \mathbb{R}^d : \|x - c\| < r\}$ and
 92 $\bar{\mathcal{B}}(c, r) = \{x \in \mathbb{R}^d : \|x - c\| \leq r\}$, respectively. Lebesgue measures on \mathbb{R} and \mathbb{R}^d are
 93 both denoted by the same symbol μ . A multivariate normal distribution with mean
 94 m and covariance matrix Σ is denoted by $\mathcal{N}(m, \Sigma)$. Its probability measure and its
 95 induced probability density under Lebesgue measure are denoted by $\Phi(\cdot; m, \Sigma)$ and
 96 $\varphi(\cdot; m, \Sigma)$. The indicator function of a set or condition C is denoted by $1\{C\}$. We
 97 use Bachmann-Landau notations: $o(\cdot)$, $O(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$, $\omega(\cdot)$.

98 2. Algorithm, Definitions and Objective Function Assumptions.

99 **2.1. Algorithm: (1+1)-ES with Success-based Step-size Control.** We
 100 analyze a generalized version of the (1+1)-ES with one-fifth success rule presented in
 101 [Algorithm 2.1](#), which implements one of the oldest approaches to adapt the step-size
 102 in randomized optimization methods [16, 50, 53]. The specific implementation was
 103 proposed in [37]. At each iteration, a candidate solution x_t is sampled. It is centered
 104 in the current incumbent m_t and follows a multivariate normal distribution with mean
 105 vector m_t and covariance matrix equal to $\sigma_t^2 I_d$ where I_d denotes the identity matrix.
 106 The candidate solution is accepted, that is m_t becomes x_t , if and only if x_t is better
 107 than m_t (i.e. $f(x_t) \leq f(m_t)$). In this case, we say that the candidate solution is
 108 successful. The step-size σ_t is adapted so as to maintain a probability of success to be
 109 approximately the target success probability denoted by $p_{\text{target}} := \frac{\log(1/\alpha_{\downarrow})}{\log(\alpha_{\uparrow}/\alpha_{\downarrow})}$. To do
 110 so, the step-size is increased by the increase factor $\alpha_{\uparrow} > 1$ in case of success (which is
 111 an indication that the step-size is likely to be too small) and decreased by the decrease
 112 factor $\alpha_{\downarrow} < 1$ otherwise. The covariance matrix Σ_t of the sampling distribution of
 113 candidate solutions is adapted in the set \mathcal{S}_{κ} of positive-definite symmetric matrices
 114 with determinant $\det(\Sigma) = 1$ and condition number $\text{Cond}(\Sigma) \leq \kappa$. We do not assume
 115 any specific update mechanism for Σ , but we assume that the update of Σ is invariant
 116 to any strictly increasing transformation of f . We call such an update comparison-
 117 based (see Lines 7 and 11 of [Algorithm 2.1](#)). Then, our algorithm behaves exact-
 118 equally on f and on $g \circ f$ for all strictly increasing functions $g : \mathbb{R} \rightarrow \mathbb{R}$ (i.e., $g(s) \leq$
 119 $g(t) \Leftrightarrow s \leq t$). This defines a class of comparison-based randomized algorithms and
 120 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  $m_0 \in \mathbb{R}^d$ ,  $\sigma_0 > 0$ ,  $\Sigma_0 = I$ ,  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , 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  $f(x_t) \leq f(m_t)$  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

```

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

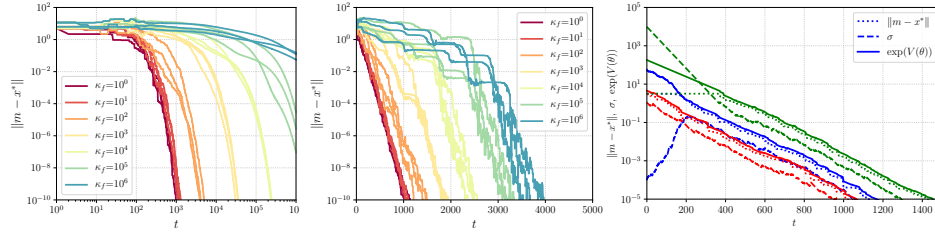


Fig. 2.1: Convergence of the (1+1)-ES (left) and the (1+1)-CMA-ES (middle) on 10 dimensional ellipsoidal function $f(x) = \frac{1}{2} \sum_{i=1}^d \kappa_f^{\frac{i-1}{d-1}} x_i^2$ with $\kappa_f = 10^0, 10^1, \dots, 10^6$. The y-axis displays the distance to the optimum (and not the function value). We employ the covariance matrix adaptation mechanism proposed by [5], where σ 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 $\alpha_{\downarrow} = e^{-0.025}$) 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 $u = \alpha_{\downarrow}^{-10}$.

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

128 Figure 2.1 shows typical runs of the (1+1)-ES and a version of (1+1)-ES $_{\kappa}$ pro-
 129 posed in [5], which is known as the (1+1)-CMA-ES, on a 10-dimensional ellipsoidal
 130 function with different condition numbers κ_f of the Hessian. It is empirically observed
 131 that Σ_t in the (1+1)-CMA-ES approaches the inverse Hessian $\nabla^2 f(m_t)$ of the objec-
 132 tive function up to the scalar factor if the objective function is convex quadratic. The
 133 runtime of (1+1)-ES scales linearly with κ_f (notice the logarithmic scale of the hori-
 134 zontal axis), while the runtime of the (1+1)-CMA-ES suffers only an additive penalty,
 135 roughly proportional to the logarithm of κ_f . Once the Hessian is well approximated
 136 by Σ (up to a scalar factor), it approaches the global optimum geometrically at the
 137 same rate for different values of κ_f .

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

142 2.2. Preliminary definitions.

143 **2.2.1. Spatial Suboptimality Function.** The algorithms studied in this paper
 144 are comparison-based and thus invariant to strictly increasing transformations of f . If
 145 the convergence of the algorithms is measured in terms of f , say by investigating the
 146 convergence or hitting time of the sequence $f(m_t)$, this will not reflect the invariance
 147 to monotonic transformations of f because the first iteration t_0 such that $f(m_{t_0}) \leq \epsilon$
 148 is not equal to the first iteration t'_0 such that $g(f(m_{t'_0})) \leq \epsilon$ for some $\epsilon > 0$. For this
 149 reason, we introduce a quality measure called *spatial suboptimality function* [23]. It
 150 is the d th root of the volume of the sub-levelset where the function value is better or

151 equal to $f(x)$:

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

$$156 \quad (2.1) \quad f_\mu(x) = \sqrt[d]{\mu(f^{-1}((-\infty, f(x)]))} = \sqrt[d]{\mu(\{y \in \mathbb{R}^d \mid f(y) \leq f(x)\})} .$$

157 We remark that for any f , the suboptimality function f_μ is greater or equal to zero.
 158 For any f and any strictly increasing function $g : \text{Im}(f) \rightarrow \mathbb{R}$, f and its composite
 159 $g \circ f$ have the same spatial suboptimality function such that hitting time of f_μ smaller
 160 than $\epsilon > 0$ will be the same for f or $g \circ f$. Moreover, there exists a strictly increasing
 161 function g such that $f_\mu(x) = g(f(x))$ holds μ -almost everywhere [23, Lemma 1].

162 We will investigate the expected first hitting time of $\|m_t - x^*\|$ to $\epsilon > 0$. For
 163 this, we will bound the first hitting time of $\|m_t - x^*\|$ to ϵ by the first hitting time
 164 of $f_\mu(m_t)$ to a constant times ϵ . To understand why, consider first a strictly convex
 165 quadratic function f with Hessian H and minimal solution x^* . We have $f_\mu(x) =$
 166 $V_d [2(f(x) - f(x^*)) / \det(H)^{1/d}]^{1/2}$ for all $x \in \mathbb{R}^d$, where $V_d = \pi^{1/2} / \Gamma^{1/d}(d/2 + 1)$ is the
 167 d th root of the volume of the d -dimensional unit hyper-sphere [2]. This implies that
 168 the first hitting time of $f_\mu(m_t)$ translates to the first hitting time of $\sqrt{f(m_t) - f(x^*)}$.
 169 We have $\sqrt{\lambda_{\min}} \|x - x^*\| \leq \sqrt{f(x) - f(x^*)} \leq \sqrt{\lambda_{\max}} \|x - x^*\|$, where λ_{\min} and λ_{\max}
 170 are the minimal and maximal eigenvalues of H . E.g., consider $f(x) = \|x - x^*\|^2 + 1$.
 171 Then the above condition also translates to the first hitting time of $\|m_t - x^*\|$. More
 172 generally, we will formalize an assumption on f later on (Assumption A1), which
 173 allow us to bound $\|x - x^*\|$ by a constant times $f_\mu(x)$ from above and below (see
 174 (2.6)), implying that the first hitting time of $\|m_t - x^*\|$ to ϵ is bounded by that of
 175 $f_\mu(m_t)$ to ϵ , times a constant.

176 **2.2.2. Success Probability.** The success probability, i.e., the probability of
 177 sampling a candidate solution x_t with an objective function better than or equal
 178 to that of the current solution m_t , plays an important role in the analysis of the
 179 (1+1)-ES $_\kappa$ with success-based step-size control mechanism. We present here several
 180 useful definitions related to the success probability.

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

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

$$186 \quad (2.2) \quad S_r(m) = \{x \in \mathbb{R}^d \mid f_\mu(x) \leq (1 - r)f_\mu(m)\} .$$

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

¹For $r = 0$, the success domain $S_0(m)$ is not necessarily equivalent to the sub-levelset $S'_0(m) := \{x \in \mathbb{R}^d \mid f(x) \leq f(m)\}$, 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 $\Phi(\cdot; 0, \Sigma)$ for $\Sigma \in \mathcal{S}_\kappa$, the success probability with rate $r = 0$ is equivalent to $\Pr_{z \sim \mathcal{N}(0, \Sigma)} [m + f_\mu(m) \cdot \bar{\sigma} z \in S'_0(m)]$, with $\bar{\sigma}$ defined in (2.3).

190 $\bar{\sigma}$ is defined as

$$191 \quad (2.3) \quad p_r^{\text{succ}}(\bar{\sigma}; m, \Sigma) = \Pr_{z \sim \mathcal{N}(0, \Sigma)} [m + f_\mu(m)\bar{\sigma}z \in S_r(m)] \ .$$

192 **Definition 2.3** introduces the notion of *normalized step-size* $\bar{\sigma}$ and the success
 193 probability is defined as a function of $\bar{\sigma}$ rather than the actual step-size $\sigma = f_\mu(m)\bar{\sigma}$.
 194 This is motivated by the fact that as m approaches the global optimum x^* of f , the
 195 step-size σ needs to shrink for the success probability to be constant. If the objective
 196 function is $f(x) = \frac{1}{2}\|x - x^*\|^2$ and the covariance matrix is the identity matrix, then
 197 the success probability is fully controlled by $\bar{\sigma}_t = \sigma_t/f_\mu(m_t) \propto \sigma_t/\|m_t - x^*\|$ and is
 198 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 $e_1 = (1, 0, \dots, 0)$, we have*

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

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

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

204 Since the success probability under a given normalized step-size depends on m
 205 and Σ , we define the upper and lower success probability as follows.

206 **DEFINITION 2.5** (Lower and Upper Success Probability). *Let $\mathcal{X}_a^b = \{x \in \mathbb{R}^d :$
 207 $a < f_\mu(x) \leq b\}$. Given the normalized step-size $\bar{\sigma} > 0$, the lower and upper success
 208 probabilities are defined as*

$$209 \quad p_{(a,b]}^{\text{lower}}(\bar{\sigma}) = \inf_{m \in \mathcal{X}_a^b} \inf_{\Sigma \in \mathcal{S}_\kappa} p_0^{\text{succ}}(\bar{\sigma}; m, \Sigma) \ , \quad p_{(a,b]}^{\text{upper}}(\bar{\sigma}) = \sup_{m \in \mathcal{X}_a^b} \sup_{\Sigma \in \mathcal{S}_\kappa} p_0^{\text{succ}}(\bar{\sigma}; m, \Sigma) \ .$$

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

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

$$223 \quad (2.4) \quad p = \liminf_{\bar{\sigma} \rightarrow +0} p_0^{\text{succ}}(\bar{\sigma}; m, \Sigma) \ .$$

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

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

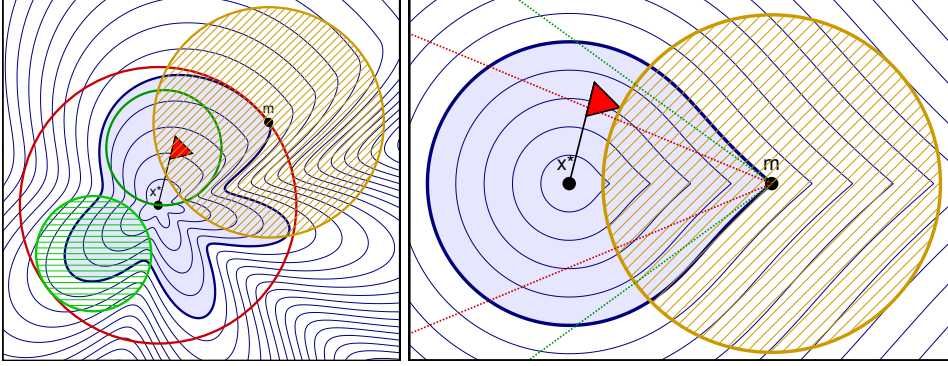


Fig. 2.2: The sampling distribution is indicated by the mean m and the shaded orange circle, indicating one standard deviation. The blue set is the sub-levelset $S_0(m)$ of points improving upon m . **Left:** Illustration of property **A1** in [Subsection 2.3](#). The blue set is enclosed in the red (outer) ball of radius $C_u f_\mu(m)$ and contains the dark green (inner) ball of radius $C_l f_\mu(m)$. 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^{limit} of improving steps (inside $S_0(m)$) for small enough step size σ (green outline). It contains a smaller cone (red outline) covering a probability mass of p^{target} .

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

236 *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 $f'(x) =$
 237 $f(m + \sqrt{\Sigma}(x - m))$. Therefore, it suffices to show that the claims hold for $\Sigma = I_d$ on
 238 f' , which is proven in Lemma 4 in [23]. \square

239 **2.3. Main Assumptions on the Objective Functions.** Given positive real
 240 numbers a and b satisfying $0 \leq a < b \leq +\infty$, and a measurable objective function,
 241 let \mathcal{X}_a^b be the set defined in [Definition 2.5](#).

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

247 **A1** We assume that f is a measurable function and that there exists \mathcal{X}_a^b such
 248 that for any $m \in \mathcal{X}_a^b$, there exist an open ball \mathcal{B}_ℓ with radius $C_\ell f_\mu(m)$ and a
 249 closed ball \mathcal{B}_u with radius $C_u f_\mu(m)$ such that it holds $\mathcal{B}_\ell \subseteq \{x \in \mathbb{R}^d \mid f_\mu(x) <$
 250 $f_\mu(m)\}$ and $\{x \in \mathbb{R}^d \mid f_\mu(x) \leq f_\mu(m)\} \subseteq \mathcal{B}_u$.

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

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

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

$$260 \quad (2.5) \quad \liminf_{\bar{\sigma} \downarrow 0} p_{(a,b]}^{\text{lower}}(\bar{\sigma}) \geq p^{\text{limit}} .$$

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

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

$$271 \quad (2.6) \quad C_\ell f_\mu(x) \leq \|x - x^*\| \leq C_u f_\mu(x) .$$

272 If the balls are not centered at the optimum, we have the one-side inequality $\|x - x^*\| \leq$
 273 $2C_u f_\mu(x)$. Hence, the expected first hitting time of $f_\mu(m_t)$ to $[0, \epsilon]$ translates to an
 274 upper bound for the expected first hitting time of $\|m_t - x^*\|$ to $[0, 2C_u \epsilon]$.

275 We remark that **A1** and **A2** satisfied for $a = 0$ allow to include some non-
 276 differentiable functions with non-convex sublevel sets as illustrated in [Figure 2.2](#).

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

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

$$287 \quad (2.7) \quad (L_\ell/2)\|x - x^*\|^2 \leq h(x) - h(x^*) \leq (L_u/2)\|x - x^*\|^2 .$$

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

$$291 \quad (2.8) \quad h(x^* + \gamma x) = h(x^*) + \gamma (h(x^* + x) - h(x^*)) .$$

292 The following lemmas show that these assumptions imply **A1** and **A2**. The proofs
 293 of the lemmas are presented in [Appendix B.1](#) and [Appendix B.2](#), respectively.

294 **LEMMA 2.8.** *Let f satisfy **A3**. Then, f satisfies **A1** and **A2** with $a = 0$, $b = \infty$,*

$$295 \quad C_\ell = \frac{1}{\sqrt{d}} \sqrt{\frac{L_\ell}{L_u}} \text{ and } C_u = \frac{1}{\sqrt{d}} \sqrt{\frac{L_u}{L_\ell}} .$$

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

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

300 **3.1. First Hitting Time.** We start with the generic definition of the *first hitting*
 301 *time* of a stochastic process $\{X_t : t \geq 0\}$, defined as follows.

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

306 The first hitting time is the number of iterations that the stochastic process
 307 requires to reach the target level $\beta < \beta_0$ for the first time. In our situation, $X_t =$
 308 $\|m_t - x^*\|$ measures the distance from the current solution m_t to the target point
 309 x^* (typically, global or local optimal point) after t iterations. Then, $\beta = \epsilon > 0$
 310 defines the target accuracy and T_ϵ^X is the runtime of the algorithm until it finds an
 311 ϵ -neighborhood $\mathcal{B}(x^*, \epsilon)$. The first hitting time T_ϵ^X is a random variable as m_t is a
 312 random variable. In this paper, we focus on the *expected first hitting time* $\mathbb{E}[T_\epsilon^X]$. We
 313 want to derive lower and upper bounds on this expected hitting time that relate to
 314 the linear convergence of X_t towards x^* . Such bounds take the following form: There
 315 exist $C_T, \tilde{C}_T \in \mathbb{R}$ and $C_R > 0, \tilde{C}_R > 0$ such that for any $0 < \epsilon \leq \beta_0$

$$316 \quad (3.1) \quad \tilde{C}_T + \frac{\log(\|m_0 - x^*\|/\epsilon)}{\tilde{C}_R} \leq \mathbb{E}[T_\epsilon^X | \mathcal{F}_0] \leq C_T + \frac{\log(\|m_0 - x^*\|/\epsilon)}{C_R} .$$

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

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

$$328 \quad (3.2) \quad \limsup_{t \rightarrow \infty} \frac{1}{t} \log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|} \leq -C \text{ almost surely.}$$

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

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

341 **3.2. Bounds of the Hitting Time via Drift Conditions.** We are going to use
 342 *drift analysis* that consists in deducing properties of a sequence $\{X_t : t \geq 0\}$ (adapted
 343 to a natural filtration $\{\mathcal{F}_t : t \geq 0\}$) from its drift defined as $\mathbb{E}[X_{t+1} | \mathcal{F}_t] - X_t$ [28].
 344 Drift analysis has been widely used to analyze hitting times of evolutionary algorithms
 345 defined on discrete search spaces (mainly on binary search spaces) [10,18,19,31,32,45].
 346 Though they were developed mainly for finite search spaces, the drift theorems can
 347 naturally be generalized to continuous domains [41,43]. Indeed, Jägersküpfer's work
 348 [33,35,36] is based on the same idea, while the link to the drift analysis was implicit.

349 Since many drift conditions have been developed for analyzing algorithms on dis-
 350 crete domains, the domain of X_t is often implicitly assumed to be bounded. However,
 351 this assumption is violated in our situation, where we will use $X_t = \log(f_\mu(m_t))$
 352 as the quality measure, which takes values in $\mathbb{R} \cup \{-\infty\}$, and is meant to approach
 353 $-\infty$. We refer to [3] for additional details. In general, translating expected progress
 354 requires bounding the tail of the progress distribution, as formalized in [28].

355 To control the tails of the drift distribution, we construct a stochastic process
 356 $\{Y_t : t \geq 0\}$ iteratively as follows: $Y_0 = X_0$ and

$$357 \quad (3.3) \quad Y_{t+1} = Y_t + \max \{X_{t+1} - X_t, -A\} 1_{\{T_\beta^X > t\}} - B 1_{\{T_\beta^X \leq t\}}$$

358 for some $A \geq B > 0$ and $\beta < \beta_0$ with $X_0 = \beta_0$. It clips $X_{t+1} - X_t$ to some constant
 359 $-A$ ($A > 0$) from below. We introduce the indicator $1_{\{T_\beta^X > t\}}$ for a technical reason.
 360 The process disregards progress larger than A , and it fixes the progress of the step
 361 that hits the target set to B . It is formalized in the following theorem, which is our
 362 main mathematical tool to derive an upper bound of the expected first hitting time
 363 of (1+1)-ES $_\kappa$ in the form of (3.1).

364 **THEOREM 3.2.** *Let $\{X_t : t \geq 0\}$ be a sequence of real-valued random variables
 365 adapted to a filtration $\{\mathcal{F}_t : t \geq 0\}$ with $X_0 = \beta_0 \in \mathbb{R}$. For $\beta < \beta_0$, let $T_\beta^X =$
 366 $\inf \{t : X_t \leq \beta\}$ be the first hitting time of the set $(-\infty, \beta]$. Define a stochastic process
 367 $\{Y_t : t \geq 0\}$ iteratively through (3.3) with $Y_0 = X_0$ for some $A \geq B > 0$, and let
 368 $T_\beta^Y = \inf \{t : Y_t \leq \beta\}$ be the first hitting time of the set $(-\infty, \beta]$. If Y_t is integrable,
 369 i.e. $\mathbb{E}[|Y_t|] < \infty$, and*

$$370 \quad (3.4) \quad \mathbb{E}[\max \{X_{t+1} - X_t, -A\} 1_{\{T_\beta^X > t\}} | \mathcal{F}_t] \leq -B 1_{\{T_\beta^X > t\}} \quad ,$$

371 *then the expectation of T_β^X satisfies*

$$372 \quad (3.5) \quad \mathbb{E}[T_\beta^X] \leq \mathbb{E}[T_\beta^Y] \leq \frac{A + \beta_0 - \beta}{B} \quad .$$

373 *Proof of Theorem 3.2.* We consider the stopped process $\bar{X}_t = X_{\min\{t, T_\beta^X\}}$. We
 374 have $X_t \leq \bar{X}_t$ for $t \leq T_\beta^X$ and $\bar{X}_t \leq Y_{\min\{t, T_\beta^X\}}$ for all $t \geq 0$. Therefore, we have
 375 $T_\beta^X = T_\beta^{\bar{X}} \leq T_\beta^Y$. Let $\bar{Y}_t = Y_{\min\{t, T_\beta^Y\}}$. By construction it holds $Y_t \leq \bar{Y}_t$ for $t \leq T_\beta^Y$
 376 and $T_\beta^Y = T_\beta^{\bar{Y}}$. Hence, $T_\beta^X \leq T_\beta^Y \leq T_\beta^{\bar{Y}}$.

377 We will prove that

$$378 \quad (3.6) \quad \mathbb{E}[\bar{Y}_{t+1} | \mathcal{F}_t] \leq \bar{Y}_t - B 1_{\{T_\beta^Y > t\}} \quad .$$

379 We start from

$$380 \quad (3.7) \quad \mathbb{E}[\bar{Y}_{t+1} | \mathcal{F}_t] = \mathbb{E}[\bar{Y}_{t+1} 1_{\{T_\beta^Y \leq t\}} | \mathcal{F}_t] + \mathbb{E}[\bar{Y}_{t+1} 1_{\{T_\beta^Y > t\}} | \mathcal{F}_t]$$

381 and bound the different terms:

382 (3.8)
$$\mathbb{E}[\bar{Y}_{t+1} 1_{\{T_\beta^Y \leq t\}} | \mathcal{F}_t] = \mathbb{E}[\bar{Y}_t 1_{\{T_\beta^Y \leq t\}} | \mathcal{F}_t] = \bar{Y}_t 1_{\{T_\beta^Y \leq t\}}$$

383 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

384
 385 (3.9)
$$\mathbb{E}[\bar{Y}_{t+1} 1_{\{T_\beta^Y > t\}} | \mathcal{F}_t] = \mathbb{E}[Y_{t+1} | \mathcal{F}_t] 1_{\{T_\beta^Y > t\}}$$

 386
$$\leq (Y_t - B 1_{\{T_\beta^X > t\}} - B 1_{\{T_\beta^X \leq t\}}) 1_{\{T_\beta^Y > t\}} = (\bar{Y}_t - B) 1_{\{T_\beta^Y > t\}},$$

 387

388 where we have used condition (3.4). Hence, by injecting (3.8) and (3.9) into (3.7),
 389 we obtain (3.6). From (3.6), by taking the expectation we deduce $\mathbb{E}[\bar{Y}_{t+1}] \leq \mathbb{E}[\bar{Y}_t] -$
 390 $B \Pr[T_\beta^Y > t]$. Following the same approach as [43, Theorem 1], since T_β^Y is a random
 391 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
 392 it holds

393
$$B \mathbb{E}[T_\beta^Y] \stackrel{\tilde{t} \rightarrow \infty}{\longleftarrow} \sum_{t=0}^{\tilde{t}} B \Pr[T_\beta^Y > t] \leq \sum_{t=0}^{\tilde{t}} (\mathbb{E}[\bar{Y}_t] - \mathbb{E}[\bar{Y}_{t+1}]) = \mathbb{E}[\bar{Y}_0] - \mathbb{E}[\bar{Y}_{\tilde{t}}].$$

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

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

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

395 This theorem can be intuitively understood: we assume for the sake of simplicity a
 396 process X_t such that $X_{t+1} \geq X_t - A$. Then (3.4) states that the process progresses in
 397 expectation by at least $-B$. The theorem concludes that the expected time needed to
 398 reach a value smaller than β when started in β_0 equals to $(\beta_0 - \beta)/B$ (what we would
 399 get for a deterministic algorithm) plus A/B . This last term is due to the stochastic
 400 nature of the algorithm. It is minimized if A is as close as possible to B , which
 401 corresponds to a highly concentrated process.

402 Jägersküpper [35, Theorem 2] established a general lower bound of the expected
 403 first hitting time of the (1+1)-ES. We borrow the same idea to prove the following
 404 general theorem for a lower bound of the expected first hitting time, which generalizes
 405 [36, Lemma 12]. See Theorem 2.3 in [3] for its proof.

406 **THEOREM 3.3.** *Let $\{X_t : t \geq 0\}$ be a sequence of real-valued random variables*
 407 *adapted to a filtration $\{\mathcal{F}_t : t \geq 0\}$ and integrable such that $X_0 = \beta_0$, $X_{t+1} \leq X_t$, and*
 408 *$\mathbb{E}[X_{t+1} | \mathcal{F}_t] - X_t \geq -C$ for $C > 0$. For $\beta < \beta_0$ we define $T_\beta^X = \min\{t : X_t \leq \beta\}$.*

409 *Then the expected hitting time is lower bounded by $\mathbb{E}[T_\beta^X] \geq -(1/2) + (4C)^{-1}(\beta_0 - \beta)$.*

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

411 **4.1. Mathematical Modeling of the Algorithm.** In the sequel, we will an-
 412alyze the process $\{\theta_t : t \geq 0\}$ where $\theta_t = (m_t, \sigma_t, \Sigma_t) \in \mathbb{R}^n \times \mathbb{R}_> \times \mathcal{S}_\kappa$ generated by
 413the (1+1)-ES $_\kappa$ algorithm. We assume from now on that the optimized objective func-
 414tion f is measurable with respect to the Borel σ -algebra. We equip the state-space
 415 $\mathcal{X} = \mathbb{R}^n \times \mathbb{R}_> \times \mathcal{S}_\kappa$ with its Borel σ -algebra denoted $\mathcal{B}(\mathcal{X})$.

416 **4.2. Preliminaries.** We present two preliminary results. In Assumption [A1](#), we
 417 assume that for $m \in \mathcal{X}_a^b$, we can include a ball of radius $C_\ell f_\mu(m)$ into the sublevel
 418 set $S_0(m)$ and embed $S_0(m)$ into a ball of radius $C_u f_\mu(m)$. This allows us to upper
 419 bound and lower bound the probability of success for all $m \in \mathcal{X}_a^b$, for all $\Sigma \in \mathcal{S}_\kappa$,
 420 by the probability to sample inside of balls of radius $C_u f_\mu(m)$ and $C_\ell f_\mu(m)$ with
 421 appropriate center. From this we can upper-bound $p_{(a,b)}^{\text{upper}}(\bar{\sigma})$ by a function of $\bar{\sigma}$.
 422 Similarly we can lower-bound $p_{(a,b)}^{\text{lower}}(\bar{\sigma})$ by a function of $\bar{\sigma}$. The corresponding proof
 423 is given in [Appendix B.3](#).

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

$$426 \quad (4.1) \quad p_{(a,b)}^{\text{upper}}(\bar{\sigma}) \leq \kappa^{d/2} \Phi \left(\bar{\mathcal{B}} \left(0, \frac{C_u}{\bar{\sigma} \kappa^{1/2}} \right); 0, \mathbf{I} \right)$$

$$427 \quad (4.2) \quad p_{(a,b)}^{\text{lower}}(\bar{\sigma}) \geq \kappa^{-d/2} \Phi \left(\bar{\mathcal{B}} \left(\frac{(2C_u - C_\ell) \kappa^{1/2}}{\bar{\sigma}} e_1, \frac{C_\ell \kappa^{1/2}}{\bar{\sigma}} \right); 0, \mathbf{I} \right),$$

428 where $e_1 = (1, 0, \dots, 0)$.

430 We use the previous proposition to establish the next lemma that guarantees the
 431 existence of a finite range of normalized step-size that leads to the success probability
 432 into some range (p_u, p_ℓ) independent of m and Σ , and provides a lower bound on the
 433 success probability with rate r when the normalized step-size is in the above range.
 434 Its proof is provided in [Appendix B.4](#).

435 **LEMMA 4.2.** *We assume that f satisfies [A1](#) and [A2](#) for some $0 \leq a < b \leq \infty$.*
 436 *Then, for any p_u and p_ℓ satisfying $0 < p_u < p^{\text{target}} < p_\ell < p^{\text{limit}}$, the constants*

$$437 \quad \bar{\sigma}_\ell = \sup \left\{ \bar{\sigma} > 0 : p_{(a,b)}^{\text{lower}}(\bar{\sigma}) \geq p_\ell \right\} \quad \text{and} \quad \bar{\sigma}_u = \inf \left\{ \bar{\sigma} > 0 : p_{(a,b)}^{\text{upper}}(\bar{\sigma}) \leq p_u \right\}$$

438 exist as positive finite values. Let $\ell \leq \bar{\sigma}_\ell$ and $u \geq \bar{\sigma}_u$ such that $u/\ell \geq \alpha_\uparrow/\alpha_\downarrow$. Then,
 439 for $r \in [0, 1]$, p_r^* defined as

$$441 \quad (4.3) \quad p_r^* := \inf_{\ell \leq \bar{\sigma} \leq u} \inf_{m \in \mathcal{X}_a^b} \inf_{\Sigma \in \mathcal{S}_\kappa} p_r^{\text{succ}}(\bar{\sigma}; m, \Sigma)$$

442 is lower bounded by a positive number defined by

$$443 \quad (4.4) \quad \min_{\ell \leq \bar{\sigma} \leq u} \kappa^{-d/2} \Phi \left(\bar{\mathcal{B}} \left(\left(\frac{(2C_u - (1-r)C_\ell) \kappa^{1/2}}{\bar{\sigma}} \right) e_1, \frac{(1-r)C_\ell \kappa^{1/2}}{\bar{\sigma}} \right); 0, \mathbf{I} \right).$$

444 **4.3. Potential Function.** [Lemma 4.2](#) divides the domain of the normalized
 445 step-size into three disjoint subsets: $\bar{\sigma} \in (0, \ell)$ is a too small normalized step-size
 446 situation where we have $p_0^{\text{succ}}(\bar{\sigma}; m, \Sigma) \geq p_\ell$ for all $m \in \mathcal{X}_a^b$ and $\Sigma \in \mathcal{S}_\kappa$; $\bar{\sigma} \in (u, \infty)$
 447 is a too large normalized step-size situation where we have $p_0^{\text{succ}}(\bar{\sigma}; m, \Sigma) \leq p_u$ for all
 448 $m \in \mathcal{X}_a^b$ and $\Sigma \in \mathcal{S}_\kappa$; and $\bar{\sigma} \in [\ell, u]$ is a reasonable normalized step-size situation where
 449 the success probability with rate r is lower bounded by [\(4.4\)](#). Since $p^{\text{target}} \in [p_u, p_\ell]$,
 450 the normalized step-size is supposed to be maintained in the reasonable range.

451 Our potential function is defined as follows. In light of [Lemma 4.2](#), we can take
 452 $\ell \leq \bar{\sigma}_\ell$ and $u \geq \bar{\sigma}_u$ such that $u/\ell \geq \alpha_\uparrow/\alpha_\downarrow$. With some constant $v > 0$, we define our
 453 potential function as

$$454 \quad (4.5) \quad V(\theta) = \log(f_\mu(m)) + \max \left\{ 0, v \log \left[\frac{\alpha_\uparrow \ell f_\mu(m)}{\sigma} \right], v \log \left[\frac{\sigma}{\alpha_\downarrow u f_\mu(m)} \right] \right\}.$$

455 The rationale behind the second term on the right-hand side (RHS) is as follows.
 456 The second and third terms inside max are positive only if the normalized step-size
 457 $\bar{\sigma} = \sigma/f_\mu(m)$ is smaller than $\ell\alpha_\uparrow$ and greater than $u\alpha_\downarrow$, respectively. The potential
 458 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
 459 normalized step-size is too small or too large. We need this penalization for the follow-
 460 ing reason. If the normalized step-size is too small, the success probability is close to
 461 $1/2$ for non-critical points, assuming $f = g \circ h$ where h is a continuously differentiable
 462 function, but the progress per step is very small because the step-size directly controls
 463 the progress for instance measured as $\|m_{t+1} - m_t\| = \sigma_t \|\mathcal{N}(0, \Sigma_t)\| 1_{\{f(m_{t+1}) \leq f(m_t)\}}$.
 464 If the normalized step-size is too large, the success probability is close to zero and
 465 produces no progress with high probability. If we would use $\log f_\mu(m)$ as a potential
 466 function instead of $V(\theta)$ then the progress is arbitrarily small in such situations, which
 467 prevents the application of drift arguments. The above potential function penalizes
 468 such situations, and guarantees a certain progress in the penalized quantity since the
 469 step-size will be increased or decreased, respectively, with high probability, leading to
 470 a certain decrease of $V(\theta)$. We illustrate in [Figure 2.1](#) that $\log(f_\mu(m))$ cannot work
 471 alone as a potential function while $V(\theta)$ does: when we start from a too small or too
 472 large step-size, $\log(f_\mu(m))$ looks constant (dotted line in green and blue). Only when
 473 the step-size is started at 1, we see progress in $\log(f_\mu(m))$. Also, the step size can
 474 always get arbitrarily worse, with a very small probability, which forces us to handle
 475 the case of badly adapted step size properly. Yet the simulation of $V(\theta)$ shows that in
 476 all three situations (small, large and well adapted step-sizes compared to the distance
 477 to the optimum), we observe a geometric decrease of $V(\theta)$.

478 **4.4. Upper Bound of the First Hitting Time.** We are now ready to establish
 479 that the potential function defined in (4.5) satisfies a (truncated)-drift condition from
 480 [Theorem 3.2](#). This will in turn imply an upper bound on the expected hitting time of
 481 $f_\mu(m)$ to $[0, \epsilon]$ provided $a \leq \epsilon$. The proof follows the same line of argumentation as the
 482 proof of [[3](#), Proposition 4.2], which was restricted to the case of spherical functions. It
 483 was generalized under similar assumptions as in this paper, but for a fixed covariance
 484 matrix equal to the identity, in [[46](#), Proposition 6]. The detailed proof is given in
 485 [Appendix B.5](#).

486 **PROPOSITION 4.3.** *Consider the (1+1)-ES $_\kappa$ described in [Algorithm 2.1](#) with state*
 487 *$\theta_t = (m_t, \sigma_t, \Sigma_t)$. Assume that the minimized objective function f satisfies [A1](#) and*
 488 *[A2](#) for some $0 \leq a < b \leq \infty$. Let p_u and p_ℓ be constants satisfying $0 < p_u < p_{\text{target}} <$*
 489 *$p_\ell < p^{\text{limit}}$ and $p_\ell + p_u = 2p_{\text{target}}$. Then, there exists $\ell \leq \bar{\sigma}_\ell$ and $u \geq \bar{\sigma}_u$ such that*
 490 *$u/\ell \geq \alpha_\uparrow/\alpha_\downarrow$, where $\bar{\sigma}_\ell$ and $\bar{\sigma}_u$ are defined in [Lemma 4.2](#). For any $A > 0$, taking v*
 491 *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*
 492 *have*

$$493 \quad (4.6) \quad \mathbb{E}[\max\{V(\theta_{t+1}) - V(\theta_t), -A\} 1_{\{m_t \in \mathcal{X}_a^b\}} \mid \mathcal{F}_t] \leq -B 1_{\{m_t \in \mathcal{X}_a^b\}}$$

494 where

$$495 \quad (4.7) \quad B = \min \left\{ A p_r^* - v \log \left(\frac{\alpha_\uparrow}{\alpha_\downarrow} \right), v \frac{p_\ell - p_u}{2} \log \left(\frac{\alpha_\uparrow}{\alpha_\downarrow} \right) \right\},$$

496 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,
 497 for any $A > 0$ there exists v such that $B < A$ is positive.

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

501 **LEMMA 4.4.** *Let $\{\theta_t : t \geq 0\}$ be the sequence of parameters $\theta_t = (m_t, \sigma_t, \Sigma_t)$
 502 defined by the (1+1)-ES $_{\kappa}$ with the initial condition $\theta_0 = (m_0, \sigma_0, \Sigma_0)$ optimizing a
 503 measurable function f . Set $X_t = V(\theta_t)$ as defined in [\(4.5\)](#) and define the process Y_t
 504 as defined in [Theorem 3.2](#). Then, for any $A > 0$, $\{Y_t : t \geq 0\}$ is integrable, i.e.,
 505 $\mathbb{E}[|Y_t|] < \infty$ for each t .*

506 *Proof of Lemma 4.4.* The drift $Y_{t+1} = Y_t + \max\{V(\theta_{t+1}) - V(\theta_t), -A\} 1_{\{T_{\beta}^x > t\}} -$
 507 $B 1_{\{T_{\beta}^x \leq t\}}$ is by construction bounded by $-A$ from below. It is also bounded by a
 508 constant from above. Indeed, from the proof of [Proposition 4.3](#), it is easy to find
 509 the upper bound, say C , of the truncated one-step change, Δ_t in the proof of [Propo-](#)
 510 [sition 4.3](#), without using [A1](#) and [A2](#). Let $D = \max\{A, C\}$. Then, by recursion,
 511 $|V(\theta_t)| \leq |V(\theta_0)| + |V(\theta_t) - V(\theta_0)| \leq |Y_0| + Dt$. Hence $\mathbb{E}[|Y_t|] \leq |Y_0| + Dt < \infty$ for
 512 all t . \square

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

514 **THEOREM 4.5.** *Consider the same situation as described in [Proposition 4.3](#). Let
 515 $T_{\epsilon} = \min\{t : f_{\mu}(m_t) \leq \epsilon\}$ be the first hitting time of $f_{\mu}(m_t)$ to $[0, \epsilon]$. Choose $a \leq$
 516 $\epsilon < f_{\mu}(m_t) \leq b$, where a and b appear in [Definition 2.5](#). If $m_0 \in \mathcal{X}_a^b$, the first hitting
 517 time is upper bounded by $\mathbb{E}[T_{\epsilon}] \leq (V(\theta_0) - \log(\epsilon) + A)/B$ for $A > B > 0$ described in
 518 [Proposition 4.3](#), where $V(\theta)$ is the potential function defined in [\(4.5\)](#). Equivalently,
 519 we have $\mathbb{E}[T_{\epsilon}] \leq C_T + C_R^{-1} \log(f_{\mu}(m_0)/\epsilon)$, where*

$$520 \quad C_T = \frac{A}{B} + \frac{v}{B} \max \left\{ 0, \log \left(\frac{\alpha^{\uparrow} \ell f_{\mu}(m_0)}{\sigma_0} \right), \log \left(\frac{\sigma_0}{\alpha^{\downarrow} u f_{\mu}(m_0)} \right) \right\}, \quad C_R = B.$$

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

524 *Proof.* [Theorem 3.2](#) with [Proposition 4.3](#) and [Lemma 4.4](#) together bounds the
 525 expected first hitting time of $V(\theta_t)$ to $(-\infty, \log(\epsilon)]$ by $(V(\theta_0) - \log(\epsilon) + A)/B$. Since
 526 $\log f_{\mu}(m_t) \leq V(\theta_t)$, T_{ϵ} is bounded by the first hitting time of $V(\theta_t)$ to $(-\infty, \log(\epsilon)]$.
 527 The inequality is preserved if we take the expectation. The last claim is trivial from
 528 the inequality $\|x - x^*\| \leq 2C_u f_{\mu}(x)$, which holds under [A1](#). \square

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

537 **4.5. Lower Bound of the First Hitting Time.** We derive a general lower
 538 bound of the expected first hitting time of $\|m_t - x^*\|$ to $[0, \epsilon]$. The following results
 539 hold for an arbitrary measurable function f and for a (1+1)-ES $_{\kappa}$ with an arbitrary
 540 σ -control mechanism. The following lemma provides the lower bound of the expected
 541 one-step progress measured by the logarithm of the distance to the optimum.

542 **LEMMA 4.6.** *We consider the process $\{\theta_t : t \geq 0\}$ generated by a (1+1)-ES $_{\kappa}$ algo-*
 543 *rithm with an arbitrary step-size adaptation mechanism and an arbitrary covariance*

544 *matrix update optimizing an arbitrary measurable function f . We assume $d \geq 2$*
 545 *and $\kappa_t = \text{Cond}(\Sigma_t) \leq \kappa$. We consider the natural filtration \mathcal{F}_t . Then, the expected*
 546 *single-step progress is lower-bounded by*

$$547 \quad (4.8) \quad \mathbb{E}[\min(\log(\|m_{t+1} - x^*\|/\|m_t - x^*\|), 0) \mid \mathcal{F}_t] \geq -\kappa_t^{d/2}/d .$$

548 *Proof of Lemma 4.6.* Note first that $\log(\|m_{t+1} - x^*\|/\|m_t - x^*\|) = \log(\|x_t -$
 549 $x^*\|/\|m_t - x^*\|)1_{\{f(x_t) \leq f(m_t)\}}$. This value can be positive since $f(x_t) \leq f(m_t)$ does
 550 not imply $\|x_t - x^*\| \leq \|m_t - x^*\|$ in general. Clipping the positive part to zero,
 551 we obtain a lower bound, which is the RHS of the above equality times the indica-
 552 tor $1_{\{\|x_t - x^*\| \leq \|m_t - x^*\|\}}$. Since the quantity is non-positive, dropping the indicator
 553 $1_{\{f(x_t) \leq f(m_t)\}}$ only decreases the lower bound. Hence, we have $\min(\log(\|m_{t+1} -$
 554 $x^*\|/\|m_t - x^*\|), 0) \geq \log(\|x_t - x^*\|/\|m_t - x^*\|)1_{\{\|x_t - x^*\| \leq \|m_t - x^*\|\}}$. Then,

$$555 \quad \mathbb{E}[\min(\log(\|m_{t+1} - x^*\|) - \log(\|m_t - x^*\|), 0) \mid \mathcal{F}_t]$$

$$556 \quad \geq \mathbb{E}[\log(\|x_t - x^*\|/\|m_t - x^*\|)1_{\{\|x_t - x^*\| \leq \|m_t - x^*\|\}} \mid \mathcal{F}_t] .$$

559 We rewrite the lower bound of the drift. The RHS of the above inequality is the
 560 integral of $\log(\|x - x^*\|/\|m_t - x^*\|)$ in the integral domain $\tilde{\mathcal{B}}(x^*, \|m_t - x^*\|)$ under the
 561 probability measure $\Phi(; m_t, \sigma_t^2 \Sigma_t)$. Performing a variable change (through rotation
 562 and scaling) so that $m_t - x^*$ becomes $e_1 = (1, 0, \dots, 0)$ and letting $\tilde{\sigma}_t = \sigma_t/\|m_t - x^*\|$,
 563 we can further rewrite it as the integral of $\log(\|x\|)$ in $\tilde{\mathcal{B}}(0, 1)$ under $\Phi(; e_1, \tilde{\sigma}_t^2 \Sigma_t)$.
 564 With $\kappa_t = \text{Cond}(\Sigma_t)$, we have $\varphi(; e_1, \tilde{\sigma}_t^2 \Sigma_t) \leq \kappa_t^{d/2} \varphi(; e_1, \kappa_t \tilde{\sigma}_t^2 \mathbf{I})$, see Lemma B.1. Al-
 565 together, we obtain the lower bound $\mathbb{E}[\log(\|x_t - x^*\|/\|m_t - x^*\|)1_{\{\|x_t - x^*\| \leq \|m_t - x^*\|\}} \mid$
 566 $\mathcal{F}_t] \geq \kappa_t^{d/2} \int_{\tilde{\mathcal{B}}(0, 1)} \log(\|x\|) \varphi(; e_1, \kappa_t \tilde{\sigma}_t^2 \mathbf{I}) dx$. The RHS is equivalent to $-\kappa_t^{d/2}$ times
 567 the single step progress of the (1+1)-ES on the spherical function at $m_t = e_1$ and
 568 $\sigma = \sqrt{\kappa} \tilde{\sigma}_t$, which is proven in the proof of Lemma 4.4 of [3] to be lower bounded by
 569 $1/d$ for $d \geq 2$. This completes the proof. \square

570 The following theorem proves that the expected first hitting time of (1+1)-ES $_{\kappa}$ is
 571 $\Omega(\log(\|m_0 - x^*\|/\epsilon))$ for any measurable function f , implying that it can not converge
 572 faster than linearly. In case of $\kappa = 1$ the lower runtime bound becomes $\Omega(d(\log(\|m_0 -$
 573 $x^*\|/\epsilon)))$, meaning that the runtime scales linearly with respect to d . The proof is a
 574 direct application of Lemma 4.6 to Theorem 3.3.

575 **THEOREM 4.7.** *We consider the process $\{\theta_t : t \geq 0\}$ generated by a (1+1)-ES $_{\kappa}$*
 576 *described in Algorithm 2.1 and assume that f is a measurable function with $d \geq 2$. Let*
 577 $T_{\epsilon} = \inf\{t : \|m_t - x^*\| \leq \epsilon\}$ *be the first hitting time of $[0, \epsilon]$ by $\|m_t - x^*\|$. Then, the*
 578 *expected first hitting time is lower bounded by $\mathbb{E}[T_{\epsilon}] \geq -(1/2) + \frac{d}{4\kappa^{d/2}} \log(\|m_0 - x^*\|/\epsilon)$.*
 579 *The bound holds for arbitrary step-size adaptation mechanisms. If A1 holds, it gives*
 580 *a lower bound for the expected first hitting time bound of $f_{\mu}(m_t)$ to $[0, 2C_{\ell}\epsilon]$.*

581 *Proof of Theorem 4.7.* Let $X_t = \log\|m_t - x^*\|$ for $t \geq 0$. Define Y_t iteratively as
 582 $Y_0 = X_0$ and $Y_{t+1} = Y_t + \min(X_{t+1} - X_t, 0)$. Then, it is easy to see that $Y_t \leq X_t$ and
 583 $Y_{t+1} \leq Y_t$ for all $t \geq 0$. Note that $\mathbb{E}[Y_{t+1} - Y_t \mid \mathcal{F}_t] = \mathbb{E}[\min(X_{t+1} - X_t, 0) \mid \mathcal{F}_t] =$
 584 $\mathbb{E}[\min(\log(\|m_{t+1} - x^*\|/\|m_t - x^*\|), 0) \mid \mathcal{F}_t]$, where the RMS is lower bounded in light
 585 of Lemma 4.6. Then, applying Theorem 3.3, we obtain the lower bound. The last
 586 statement directly follows from $\|x - x^*\| \leq 2C_{\ell} f_{\mu}(x)$ under A1. \square

587 **4.6. Almost Sure Linear Convergence.** Additionally to the expected first
 588 hitting time bound, we can deduce from Proposition 4.3, almost sure linear conver-
 589 gence as stated in the following proposition.

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

$$592 \quad (4.9) \quad \Pr \left[\limsup_{t \rightarrow \infty} \frac{1}{t} \log f_\mu(m_t) \leq -B \right] = \Pr \left[\limsup_{t \rightarrow \infty} \frac{1}{t} \log \|m_t - x^*\| \leq -B \right] = 1 ,$$

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

595 *Proof of Proposition 4.8.* Let V be defined in (4.5). Let $Y_0 = V(\theta_0)$ and $Y_{t+1} =$
596 $Y_t + \max(-A, V(\theta_{t+1}) - V(\theta_t))$. Define $Z_t = Y_t - \mathbb{E}_{t-1}[Y_t]$ for $t \geq 0$. Then, $\{Z_t\}$ is
597 a martingale difference sequence on the filtration $\{\mathcal{F}_t\}$ produced by $\{\theta_t\}$. We hence
598 have $\frac{1}{t} \log f_\mu(m_t) \leq \frac{1}{t} V(\theta_t) \leq \frac{1}{t} Y_t$, and from Proposition 4.3 we obtain

$$599 \quad Y_t = \mathbb{E}_{t-1}[Y_t] + Z_t = Y_{t-1} + \mathbb{E}_{t-1}[Y_t - Y_{t-1}] + Z_t \leq Y_{t-1} - B + Z_t .$$

601 By repeatedly applying the above inequality and dividing it by t , we obtain $\frac{1}{t} Y_t \leq$
602 $-B + \frac{1}{t} Y_0 + \frac{1}{t} \sum_{i=1}^t Z_i$, where $\lim_{t \rightarrow \infty} \frac{1}{t} Y_0 = 0$ and $\sum_{i=1}^t Z_i$ is a martingale sequence.
603 In light of the strong law of large numbers for martingales [14], if $\sum_{t=1}^{\infty} \mathbb{E}[Z_t^2]/t^2 < \infty$,
604 we have $\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^t Z_i = 0$ almost surely. By the definition of $V(\theta_t)$ and the
605 working mechanism of the (1+1)-ES $_\kappa$, we have $V(\theta_i) - V(\theta_{i-1}) \leq v \log(\alpha_\uparrow/\alpha_\downarrow)$. Hence,
606 $\mathbb{E}[Z_i^2] = \mathbb{E}[(Y_i - \mathbb{E}_{i-1}[Y_i])^2] = \mathbb{E}[\max(-A, V(\theta_i) - V(\theta_{i-1}))^2] \leq \max(A, v \log(\alpha_\uparrow/\alpha_\downarrow))^2$.
607 Hence, we have $\limsup_{t \rightarrow \infty} \frac{1}{t} \log f_\mu(m_t) \leq -B + \lim_{t \rightarrow \infty} \frac{1}{t} Y_0 + \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^t Z_i =$
608 $-B$ almost surely. Along with $\|x - x^*\| \leq 2C_u f_\mu(x)$, we obtain Equation (4.9). \square

609 **4.7. Wrap-up of the Results: Global Linear Convergence.** As a corollary
610 to the lower-bound from Theorem 4.7, the upper bound from Theorem 4.5, Proposi-
611 tion 4.8 stating the almost sure linear convergence and the fact that different assump-
612 tions discussed in Section 2.3 imply A1 and A2, we summarize our linear convergence
613 results in the following theorem.

614 THEOREM 4.9 (Global Linear Convergence). We consider the (1+1)-ES $_\kappa$ opti-
615 mizing an objective function f . Suppose either

- 616 (a) f satisfies A1 and A2 for $a = 0$, $p^{\text{limit}} > p^{\text{target}}$, and $m_0 \in \mathcal{X}_0^b$; or
617 (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
618 $[0, \epsilon]$ is $\Theta(\log(\|m_0 - x^*\|/\epsilon))$ for all $\epsilon > 0$. Moreover, both $f_\mu(m_t)$ and $\|m_t - x^*\|$
619 linearly converge almost surely, i.e.

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

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

620 **4.8. Tightness in the Sphere Function Case.** Now we consider a specific
621 convex quadratic function, namely the sphere function $f(x) = \frac{1}{2} \|x\|^2$ where the spa-
622 tial suboptimality function equals $f_\mu(x) = V_d \|x\|$. In Theorem 4.9 we have formul-
623 ated that the expected hitting time of a ball of radius ϵ for the (1+1)-ES $_\kappa$ equals
624 $\Theta(\log \|m_0 - x^*\|/\epsilon)$. Yet, this statement does not give information on how the con-
625 stants hidden in the Θ -notation scale with the dimension. In particular the conver-
626 gence rate of the algorithm is upper-bounded by $\exp(-B)$ where B is given in (4.7),
627 see Theorem 4.5. In this section, we estimate precisely the scaling of B in Proposi-
628 tion 4.3 with respect to the dimension and compare it with the general lower bound

629 of the expected first hitting time given in Theorem 4.7. We then conclude that the
 630 bound is tight with respect to the scaling with d in the case of the sphere function.

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

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

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

639 Two conditions on the choice of α_{\uparrow} and α_{\downarrow} : $p_{\text{target}} = \log(1/\alpha_{\downarrow})/\log(\alpha_{\uparrow}/\alpha_{\downarrow}) \in$
 640 $\Theta(1)$ and $\log(\alpha_{\uparrow}/\alpha_{\downarrow}) \in \omega(1/d)$, are understood as follows. The first condition implies
 641 that the target success probability p_{target} must be independent of d . In the $1/5$ success
 642 rule, α_{\uparrow} and α_{\downarrow} are set so that $p_{\text{target}} = 1/5$ independent of d . The second condition
 643 implies that the factors of the step-size increase and decrease must be $\log(\alpha_{\uparrow}) \in \omega(1/d)$
 644 and $\log(1/\alpha_{\downarrow}) \in \omega(1/d)$. Note that on the sphere function the normalized step-size
 645 $\bar{\sigma} \propto \sigma/\|m - x^*\|$ is kept around a constant during the search. It implies that the
 646 convergence speed of $\|m - x^*\|$ and σ must agree. Therefore the speed of the adaptation
 647 of the step-size must not be too small to achieve $\Theta(d)$ scaling of the expected first
 648 hitting time.

649 Proposition 4.10 and Theorem 4.5 imply $\mathbb{E}[T_{\epsilon}] \in O(d \log(\|m_0\|/\epsilon))$ and Theo-
 650 rem 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
 651 result shows i) that the runtime of the (1+1)-ES on the sphere function is propor-
 652 tional to d as long as $\log(\alpha_{\uparrow}/\alpha_{\downarrow}) \in \omega(1/d)$, and ii) that from our methodology one
 653 can derive a tight bound of the runtime in some cases. The result is formally stated
 654 as follows.

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

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

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

- 670 (i) strongly convex functions with Lipschitz gradient, where linear convergence
 671 of numerical optimization algorithm is usually analyzed,
- 672 (ii) continuously differentiable positively homogenous functions, where previous
 673 linear convergence results had been introduced, and
- 674 (iii) functions with non-smooth level sets as illustrated in Figure 2.2.

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

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

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

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

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

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

706 Covariance matrix adaptation is also important for optimizing functions with non-
 707 smooth level sets. On continuously differentiable functions, we can always set α_{\uparrow} and
 708 α_{\downarrow} so that $p = \frac{\log(1/\alpha_{\downarrow})}{\log(\alpha_{\uparrow}/\alpha_{\downarrow})} < p^{\text{limit}} = 1/2$. This is the rationale behind the 1/5 success
 709 rule, where $p = 1/5$. Indeed, $p = 1/5$ is known to approximate the optimal situation on
 710 the sphere function where the expected one-step progress is maximized [50]. Therefore,
 711 one does not need to tune these parameters in a problem-specific manner. However,
 712 if the objective is not continuously differentiable and levelsets are non-smooth, then
 713 p^{limit} is in general smaller than 1/2. For example, it can be as low as $p^{\text{limit}} = 1/2^d$ on
 714 $f(x) = \|x\|_{\infty} = \max_{i=1,\dots,n} |x_i|$. Without an appropriate adaptation of the covariance
 715 matrix the success probability will be smaller than $p = 1/5$ and one must tune α_{\uparrow} and
 716 α_{\downarrow} in order to converge to the optimum, which requires information about p^{limit} . By
 717 adapting the covariance matrix appropriately, the success probability can be increased
 718 arbitrary close to 1/2 (by elongating steps in the direction of the success domain) and
 719 α_{\uparrow} and α_{\downarrow} do not require tuning.

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

724 **Acknowledgement.** We gratefully acknowledge support by Dagstuhl seminar
 725 17191 “Theory of Randomized Search Heuristics”. We would like to thank Per Kris-

726 tian Lehre, Carsten Witt, and Johannes Lengler for valuable discussions and advice
 727 on drift theory. Y. A. is supported by JSPS KAKENHI Grant Number 19H04179.

728

REFERENCES

- 729 [1] M.A. ABRAMSON, C. AUDET, J.E. DENNIS JR, AND S. LE DIGABEL, *OrthoMADS: A deter-*
 730 *ministic MADS instance with orthogonal directions*, SIAM Journal on Optimization, 20
 731 (2009), pp. 948–966.
- 732 [2] Y. AKIMOTO, *Analysis of a natural gradient algorithm on monotonic convex-quadratic-*
 733 *composite functions*, in GECCO, 2012, pp. 1293–1300.
- 734 [3] Y. AKIMOTO, A. AUGER, AND T. GLASMACHERS, *Drift theory in continuous search spaces:*
 735 *expected hitting time of the $(1+1)$ -ES with $1/5$ success rule*, in GECCO, 2018, pp. 801–
 736 808.
- 737 [4] S. ALVERNAZ AND J. TOGELIUS, *Autoencoder-augmented neuroevolution for visual doom play-*
 738 *ing*, in IEEE CIG, 2017, pp. 1–8.
- 739 [5] D. V. ARNOLD AND N. HANSEN, *Active covariance matrix adaptation for the $(1+1)$ -CMA-ES*,
 740 in GECCO, 2010, pp. 385–392.
- 741 [6] C. AUDET AND J.E. DENNIS JR, *Mesh adaptive direct search algorithms for constrained opti-*
 742 *mization*, SIAM Journal on Optimization, 17 (2006), pp. 188–217.
- 743 [7] A. AUGER AND N. HANSEN, *Linear convergence on positively homogeneous functions of a com-*
 744 *parison based step-size adaptive randomized search: the $(1+1)$ -ES with generalized one-*
 745 *fifth success rule*, 2013, <https://arxiv.org/abs/1310.8397>.
- 746 [8] A. AUGER AND N. HANSEN, *Linear convergence of comparison-based step-size adaptive ran-*
 747 *domized search via stability of Markov chains*, SIAM Journal on Optimization, 26 (2016),
 748 pp. 1589–1624.
- 749 [9] A. S. BANDEIRA, K. SCHEINBERG, AND L. N. VICENTE, *Convergence of trust-region methods*
 750 *based on probabilistic models*, SIAM Journal on Optimization, 24 (2014), pp. 1238–1264.
- 751 [10] B. BARITOMPA AND M. STEEL, *Bounds on absorption times of directionally biased random*
 752 *sequences*, Random Structures & Algorithms, 9 (1996), pp. 279–293.
- 753 [11] P. BONTRAGER, A. ROY, J. TOGELIUS, N. MEMON, AND A. ROSS, *DeepMasterPrints: Gener-*
 754 *ating MasterPrints for Dictionary Attacks via Latent Variable Evolution*, in IEEE BTAS,
 755 2018, pp. 1–9.
- 756 [12] S. BUBECK, *Convex optimization: Algorithms and complexity*, 2014, [https://arxiv.org/abs/](https://arxiv.org/abs/1405.4980)
 757 [1405.4980](https://arxiv.org/abs/1405.4980).
- 758 [13] C. CARTIS AND K. SCHEINBERG, *Global convergence rate analysis of unconstrained optimization*
 759 *methods based on probabilistic models*, Mathematical Programming, 169 (2018), pp. 337–
 760 375.
- 761 [14] Y. S. CHOW, *On a strong law of large numbers for martingales*, Ann. Math. Statist., 38 (1967),
 762 p. 610.
- 763 [15] A. R. CONN, K. SCHEINBERG, AND L. N. VICENTE, *Introduction to Derivative-Free Optimiza-*
 764 *tion*, SIAM, 2009.
- 765 [16] L. DEVROYE, *The compound random search*, in International Symposium on Systems Engineer-
 766 *ing and Analysis*, 1972, pp. 195–110.
- 767 [17] Y. DIOUANE, S. GRATTON, AND L. N. VICENTE, *Globally convergent evolution strategies*, Math-
 768 *ematical Programming*, 152 (2015), pp. 467–490.
- 769 [18] B. DOERR AND L. A. GOLDBERG, *Adaptive drift analysis*, Algorithmica, 65 (2013), pp. 224–250.
- 770 [19] B. DOERR, D. JOHANNSEN, AND C. WINZEN, *Multiplicative drift analysis*, Algorithmica, 64
 771 (2012), pp. 673–697.
- 772 [20] Y. DONG, H. SU, B. WU, Z. LI, W. LIU, T. ZHANG, AND J. ZHU, *Efficient decision-based*
 773 *black-box adversarial attacks on face recognition*, in CVPR, 2019.
- 774 [21] G. FUJII, M. TAKAHASHI, AND Y. AKIMOTO, *CMA-ES-based structural topology optimization*
 775 *using a level set boundary expression—application to optical and carpet cloaks*, Computer
 776 *Methods in Applied Mechanics and Engineering*, 332 (2018), pp. 624 – 643.
- 777 [22] T. GEIJTENBEEK, M. VAN DE PANNE, AND A. F. VAN DER STAPPEN, *Flexible muscle-based*
 778 *locomotion for bipedal creatures*, ACM Transactions on Graphics (TOG), 32 (2013), pp. 1–
 779 11.
- 780 [23] T. GLASMACHERS, *Global convergence of the $(1+1)$ Evolution Strategy to a critical point*, Evo-
 781 *lutionary Computation*, 28 (2020), pp. 27–53.
- 782 [24] D. GOLOVIN, J. KARRO, G. KOCHANSKI, C. LEE, X. SONG, AND Q. ZHANG, *Gradientless de-*
 783 *scend: High-dimensional zeroth-order optimization*, in ICLR, 2020.
- 784 [25] S. GRATTON, C. W. ROYER, L. N. VICENTE, AND Z. ZHANG, *Direct search based on probabilistic*

- 785 descent, *SIAM Journal on Optimization*, 25 (2015), pp. 1515–1541.
- 786 [26] S. GRATTON, C. W. ROYER, L. N. VICENTE, AND Z. ZHANG, *Complexity and global rates of*
787 *trust-region methods based on probabilistic models*, *IMA Journal of Numerical Analysis*, 38
788 (2017), pp. 1579–1597.
- 789 [27] D. HA AND J. SCHMIDHUBER, *Recurrent world models facilitate policy evolution*, in *NeurIPS*,
790 2018, pp. 2450–2462.
- 791 [28] B. HAJEK, *Hitting-time and occupation-time bounds implied by drift analysis with applications*,
792 *Advances in Applied Probability*, 14 (1982), pp. 502–525.
- 793 [29] N. HANSEN, A. AUGER, R. ROS, S. FINCK, AND P. POŠÍK, *Comparing results of 31 algorithms*
794 *from the black-box optimization benchmarking bbob-2009*, in *GECCO*, 2010, pp. 1689–1696.
- 795 [30] N. HANSEN AND A. OSTERMEIER, *Completely derandomized self-adaptation in evolution strate-*
796 *gies*, *Evolutionary Computation*, 9 (2001), pp. 159–195.
- 797 [31] J. HE AND X. YAO, *Drift analysis and average time complexity of evolutionary algorithms*,
798 *Artificial intelligence*, 127 (2001), pp. 57–85.
- 799 [32] J. HE AND X. YAO, *A study of drift analysis for estimating computation time of evolutionary*
800 *algorithms*, *Natural Computing*, 3 (2004), pp. 21–35.
- 801 [33] J. JÄGERSKÜPPER, *Analysis of a simple evolutionary algorithm for minimization in Euclidean*
802 *spaces*, *Automata, Languages and Programming*, 2003, pp. 188–188.
- 803 [34] J. JÄGERSKÜPPER, *Rigorous runtime analysis of the (1+1)-ES: 1/5-rule and ellipsoidal fitness*
804 *landscapes*, in *FOGA*, 2005, pp. 260–281.
- 805 [35] J. JÄGERSKÜPPER, *How the (1+1)-ES using isotropic mutations minimizes positive definite*
806 *quadratic forms*. *Theoretical Computer Science*, 361 (2006), pp. 38–56.
- 807 [36] J. JÄGERSKÜPPER, *Algorithmic analysis of a basic evolutionary algorithm for continuous opti-*
808 *mization*, *Theoretical Computer Science*, 379 (2007), pp. 329–347.
- 809 [37] S. KERN, S. D. MÜLLER, N. HANSEN, D. BÜCHE, J. OCENASEK, AND P. KOUMOUTSAKOS, *Learn-*
810 *ing probability distributions in continuous evolutionary algorithms—a comparative review*,
811 *Natural Computing*, 3 (2004), pp. 77–112.
- 812 [38] J. KONEČNÝ AND P. RICHTÁRIK, *Simple complexity analysis of simplified direct search*, 2014,
813 <https://arxiv.org/abs/1410.0390>.
- 814 [39] I. KRIEST, V. SAUERLAND, S. KHATIWALA, A. SRIVASTAV, AND A. OSCHLIES, *Calibrating a*
815 *global three-dimensional biogeochemical ocean model (mops-1.0)*, *Geoscientific Model De-*
816 *velopment*, 10 (2017), p. 127.
- 817 [40] J. LARSON, M. MENICKELLY, AND S. M. WILD, *Derivative-free optimization methods*, *Acta*
818 *Numerica*, 28 (2019), pp. 287–404.
- 819 [41] P. K. LEHRE AND C. WITT, *General drift analysis with tail bounds*, 2013, <https://arxiv.org/abs/1307.2559>.
- 820 [42] J. LENGLER, *Drift analysis*, in *Theory of Evolutionary Computation*, Springer, 2020, pp. 89–
821 131.
- 822 [43] J. LENGLER AND A. STEGER, *Drift analysis and evolutionary algorithms revisited*, 2016, <https://arxiv.org/abs/1608.03226>.
- 823 [44] P. MACALPINE, S. BARRETT, D. URIELI, V. VU, AND P. STONE, *Design and optimization of an*
824 *omnidirectional humanoid walk: A winning approach at the RoboCup 2011 3D simulation*
825 *competition*, in *AAAI*, 2012.
- 826 [45] B. MITAVSKIY, J. ROWE, AND C. CANNINGS, *Theoretical analysis of local search strategies to*
827 *optimize network communication subject to preserving the total number of links*, *Internation-*
828 *al Journal of Intelligent Computing and Cybernetics*, 2 (2009), pp. 243–284.
- 829 [46] D. MORINAGA AND Y. AKIMOTO, *Generalized drift analysis in continuous domain: linear con-*
830 *vergence of (1+1)-ES on strongly convex functions with lipschitz continuous gradients*, in
831 *FOGA*, 2019, pp. 13–24.
- 832 [47] A. NEMIROVSKI, *Information-based complexity of convex programming*, *Lecture Notes*, (1995).
- 833 [48] Y. NESTEROV, *Lectures on convex optimization*, vol. 137, Springer, 2018.
- 834 [49] C. PAQUETTE AND K. SCHEINBERG, *A stochastic line search method with convergence rate*
835 *analysis*, 2018, <https://arxiv.org/abs/1807.07994>.
- 836 [50] I. RECHENBERG, *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der*
837 *biologischen Evolution*, Frommann-Holzboog, 1973.
- 838 [51] I. RECHENBERG, *Evolutionsstrategie '94*, frommann-holzboog, 1994.
- 839 [52] L. M. RIOS AND N. V. SAHINIDIS, *Derivative-free optimization: a review of algorithms and com-*
840 *parison of software implementations*, *Journal of Global Optimization*, 56 (2013), pp. 1247–
841 1293.
- 842 [53] M. SCHUMER AND K. STEIGLITZ, *Adaptive step size random search*, *Automatic Control, IEEE*
843 *Transactions on*, 13 (1968), pp. 270–276.
- 844 [54] S. U. STICH, C. L. MULLER, AND B. GARTNER, *Optimization of convex functions with random*
845

- 847 *pursuit*, SIAM Journal on Optimization, 23 (2013), pp. 1284–1309.
- 848 [55] S. U. STICH, C. L. MÜLLER, AND B. GÄRTNER, *Variable metric random pursuit*, Mathematical
849 Programming, 156 (2016), pp. 549–579.
- 850 [56] J. UHLENDORF, A. MIERMONT, T. DELAVEAU, G. CHARVIN, F. FAGES, S. BOTTANI, G. BATT,
851 AND P. HERSEN, *Long-term model predictive control of gene expression at the popula-
852 tion and single-cell levels*, Proceedings of the National Academy of Sciences, 109 (2012),
853 pp. 14271–14276.
- 854 [57] V. VOLZ, J. SCHRUM, J. LIU, S. M. LUCAS, A. SMITH, AND S. RISI, *Evolving Mario levels in
855 the latent space of a deep convolutional generative adversarial network*, in GECCO, 2018,
856 pp. 221–228.

857 Appendix A. Some Numerical Results.

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

861 All algorithms were started at the initial search point $x_0 = \frac{1}{\sqrt{d}}(1, \dots, 1) \in \mathbb{R}^d$. We
862 implemented the algorithms as follows, with their parameters tuned where necessary:
863 The ES always uses the setting $\alpha_\uparrow = \exp(4/d)$ and $\alpha_\downarrow = \alpha_\uparrow^{-1/4}$ for step size adaptation.
864 We set the constant c in the sufficient decrease condition of Simplified Direction
865 Search to $\frac{1}{10}$, and we employed the standard basis as well as the negatives of these
866 vectors as candidate directions. In each iteration we looped over the set of directions
867 in random order. Randomizing the order greatly boosted performance over a fixed
868 order. Random Pursuit was implemented with a golden section line search in the range
869 $[-2\sigma, 2\sigma]$ with a rather loose target precision of $\sigma/2$, where σ is either the initial step
870 size or the length of the previous step. For Gradientless Descent we used the initial
871 step size as the maximal step size and defined a target precision of 10^{-10} . This target
872 is reached by the ES in all cases. The experiments are designed to demonstrate several
873 different effects: (a) We perform all experiments in $d = 10$ and $d = 50$ dimensions to
874 investigate dimension-dependent effects. (b) We investigate best-case performance by
875 running the algorithms on the spherical function $\|x\|^2$, i.e., on the separable convex
876 quadratic function with minimal condition number. The initial step size is set to
877 $\sigma_0 = 1$. All algorithms have a budget of $100d$ function evaluations. (c) We investigate
878 the dependency of the performance on initial parameter settings by repeating the
879 same experiment as above, but with an initial step size of $\sigma_0 = \frac{1}{1000}$. All algorithms
880 have a budget of $700d$ function evaluations. (d) We investigate the dependence on
881 problem difficulty by running the algorithms on an ellipsoid problem with a moderate
882 condition number of $\kappa_f = 100$. The eigenvalues of the Hessian are evenly distributed
883 on a log-scale. We use $\sigma_0 = 1$ like in the first experiment. All algorithms have a budget
884 of $500d$ function evaluations. The experimental results are presented in Figure A.1.

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

896 Appendix B. Proofs.

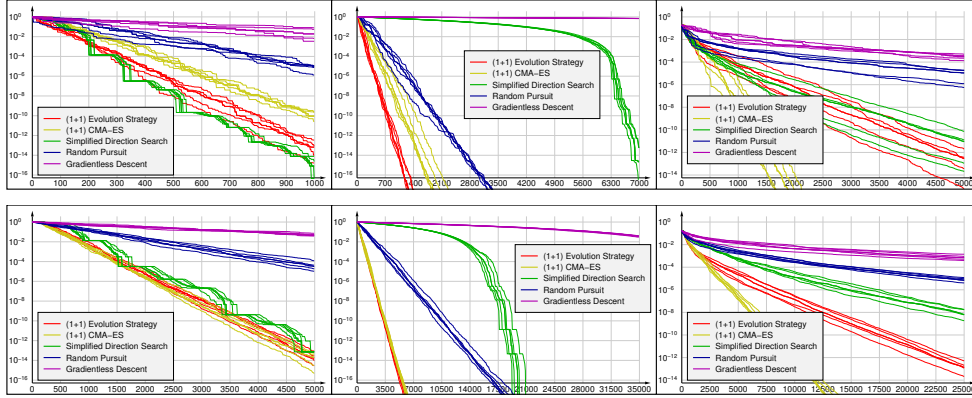


Fig. A.1: Comparison of (1+1)-ES with and without covariance matrix adaptation with three well-analyzed derivative-free optimization algorithms on two convex quadratic functions. The left column of plots shows the performance on the sphere function $\|x\|^2$ in dimensions 10 (top) and 50 (bottom). The 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).

897 **B.1. Proof of Lemma 2.8.** Since f_μ is invariant to g , without loss of generality we assume $f(x) = h(x) - h(x^*)$ in this proof. Inequality (2.7) implies that
 898 $f(y) \leq f(x) \Rightarrow (L_\ell/2)\|y - x^*\|^2 \leq f(x)$, meaning that $\{y : f(y) \leq f(x)\} \subseteq$
 899 $\bar{\mathcal{B}}(x^*, \sqrt{\frac{f(x)}{L_\ell/2}})$. Since $f_\mu(x)$ is the d th root of the volume of the left-hand side
 900 of the above relation, we find $f_\mu(x) \leq \mu^{\frac{1}{d}} \left(\bar{\mathcal{B}}(x^*, \sqrt{\frac{f(x)}{L_\ell/2}}) \right) = V_d \sqrt{\frac{f(x)}{L_\ell/2}}$. Analo-
 901 gously, we obtain $\mathcal{B}(x^*, \sqrt{\frac{f(x)}{L_u/2}}) \subseteq \{y : f(y) < f(x)\}$ and $f_\mu(x) \geq V_d \sqrt{\frac{f(x)}{L_u/2}}$.
 902 From these inequalities, we obtain $\{y : f(y) \leq f(x)\} \subseteq \bar{\mathcal{B}}(x^*, \sqrt{\frac{L_u}{L_\ell} \frac{f_\mu(x)}{V_d}})$ and
 903 $\mathcal{B}(x^*, \sqrt{\frac{L_\ell}{L_u} \frac{f_\mu(x)}{V_d}}) \subseteq \{y : f(y) < f(x)\}$. This implies A1 for \mathcal{X}_0^∞ . A2 is immedi-
 904 ately implied by Proposition 2.7. This completes the proof.
 905

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

909 It is easy to see that the spatial suboptimality function $f_\mu(x)$ is proportional
 910 to $h(x) - h(x^*)$. Let $f_\mu(x) = c(h(x) - h(x^*))$ for some $c > 0$. Then, f_μ is also a
 911 homogeneous function. Since it is homogeneous, A1 reduces to that there are open
 912 and closed balls with radius C_ℓ and C_u satisfying the conditions described in the
 913 assumption with $f_\mu(m) = 1$. Such constants are obtained by $C_u = \sup\{\|x - x^*\| :$
 914 $f_\mu(x) = 1\}$ and $C_\ell = \inf\{\|x - x^*\| : f_\mu(x) = 1\}$.

915 Due to the continuity of f there exists an open ball B around x^* such that
 916 $h(x) < h(x^*) + 1/c$ for all $x \in B$. Then, it holds that $f_\mu(x) < 1$ for all $x \in B$. It
 917 implies that C_ℓ is no smaller than the radius of B , which is positive. Hence, $C_\ell > 0$.

918 We show the finiteness of C_u by a contradiction argument. Suppose $C_u = \infty$.
 919 Then, there is a direction v such that $f_\mu(x^* + Mv) \leq 1$ with an arbitrarily large
 920 $M > 0$. Since f_μ is homogeneous, we have $f_\mu(x^* + v) \leq 1/M$ and this must hold for
 921 any $M > 0$. This implies $f_\mu(x^* + v) = c(h(x) - h(x^*)) = 0$, which contradicts the
 922 assumption that x^* is the unique global optimum. Hence, $C_u < \infty$.

923 The above argument proves that **A1** holds with the above constants for $a = 0$ and
 924 $b = \infty$. **Proposition 2.7** proves **A2**.

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

$$\begin{aligned} 927 \quad p_{(a,b]}^{\text{upper}}(\bar{\sigma}) &= \sup_{m \in \mathcal{X}_a^b} \sup_{\Sigma \in \mathcal{S}_\kappa} \int_{S_0(m)} \varphi(x; m, (f_\mu(m)\bar{\sigma})^2 \Sigma) dx \\ 928 \quad &\leq \sup_{m \in \mathcal{X}_a^b} \sup_{\Sigma \in \mathcal{S}_\kappa} \underbrace{\int_{\bar{\mathcal{B}}_u} \varphi(x; m, (f_\mu(m)\bar{\sigma})^2 \Sigma) dx}_{(*)} . \end{aligned}$$

930 The integral is maximized if the ball is centered at m . By a variable change ($x \leftarrow$
 931 $x - m$),

$$\begin{aligned} 932 \quad (*) &\leq \int_{\|x\| \leq C_u f_\mu(m)} \varphi(x; 0, (f_\mu(m)\bar{\sigma})^2 \Sigma) dx = \int_{\|x\| \leq C_u / \bar{\sigma}} \varphi(x; 0, \Sigma) dx \\ 933 \quad &\leq \kappa^{d/2} \Phi(\bar{\mathcal{B}}(0, \frac{C_u}{\bar{\sigma} \kappa^{1/2}}); 0, \mathbf{I}) . \end{aligned}$$

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

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

$$\begin{aligned} 939 \quad p_{(a,b]}^{\text{lower}}(\bar{\sigma}) &= \inf_{m \in \mathcal{X}_a^b} \inf_{\Sigma \in \mathcal{S}_\kappa} \int_{S_0(m)} \varphi(x; m, (f_\mu(m)\bar{\sigma})^2 \Sigma) dx \\ 940 \quad &\geq \inf_{m \in \mathcal{X}_a^b} \inf_{\Sigma \in \mathcal{S}_\kappa} \underbrace{\int_{\mathcal{B}_\ell} \varphi(x; m, (f_\mu(m)\bar{\sigma})^2 \Sigma) dx}_{(*)} . \end{aligned}$$

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

$$\begin{aligned} 944 \quad (*) &\geq \int_{\|x - ((2C_u - C_\ell)f_\mu(m))e_m\| \leq C_\ell f_\mu(m)} \varphi(x; 0, (f_\mu(m)\bar{\sigma})^2 \Sigma) dx \\ 945 \quad &= \int_{\|x - ((2C_u - C_\ell)/\bar{\sigma})e_m\| \leq C_\ell / \bar{\sigma}} \varphi(x; 0, \Sigma) dx \\ 946 \quad &\geq \kappa^{-d/2} \Phi(\bar{\mathcal{B}}\left(\left(\frac{(2C_u - C_\ell)\kappa^{1/2}}{\bar{\sigma}}\right)e_m, \frac{C_\ell \kappa^{1/2}}{\bar{\sigma}}\right); 0, \mathbf{I}) . \end{aligned}$$

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

951 **LEMMA B.1.** For all $\Sigma \in \mathcal{S}_\kappa$, $\kappa^{-d/2} \varphi(x; 0, \kappa^{-1} \mathbf{I}) \leq \varphi(x; 0, \Sigma) \leq \kappa^{d/2} \varphi(x; 0, \kappa \mathbf{I})$
 952 and $\kappa^{-d/2} \Phi(\mathcal{B}(\sqrt{\kappa}c, \sqrt{\kappa}r); 0, \mathbf{I}) \leq \Phi(\mathcal{B}(c, r); 0, \Sigma) \leq \kappa^{d/2} \Phi(\mathcal{B}(c/\sqrt{\kappa}, r/\sqrt{\kappa}); 0, \mathbf{I})$.

953 *Proof.* For $\Sigma \in \mathcal{S}_\kappa$, we have $\det(\Sigma) = 1$ and $\text{Cond}(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma) \leq$
 954 κ . Since $\det(\Sigma) = 1$ and $\det(\Sigma) = \prod_{i=1}^d \lambda_i(\Sigma)$, we have $\lambda_{\max}(\Sigma) \geq 1 \geq \lambda_{\min}(\Sigma)$.
 955 Therefore, we have $\lambda_{\min}(\Sigma) \geq \lambda_{\max}/\kappa \geq \kappa^{-1}$ and $\lambda_{\max}(\Sigma) \leq \kappa \lambda_{\min}(\Sigma) \leq \kappa$. Then we
 956 obtain $\kappa^{-1} x^T \mathbf{I} x \leq x^T \Sigma^{-1} x \leq \kappa x^T \mathbf{I} x$. With this inequality we have
 957

$$\begin{aligned} 958 \quad \varphi(x; 0, \Sigma) &= (2\pi)^{-d/2} \exp(-x^T \Sigma^{-1} x / 2) \leq (2\pi)^{-d/2} \exp(-x^T \mathbf{I} x / (2\kappa)) \\ 959 \quad &= \kappa^{d/2} (2\pi \kappa)^{-d/2} \exp(-x^T \mathbf{I} x / (2\kappa)) = \kappa^{d/2} \varphi(x; 0, \kappa \mathbf{I}) . \end{aligned}$$

961 Analogously, we obtain $\varphi(x; 0, \Sigma) \geq \kappa^{-d/2} \varphi(x; 0, \kappa^{-1}\mathbf{I})$. Taking the integral over
962 $\mathcal{B}(c, r)$, we obtain the second statement. \square

963 **B.4. Proof of Lemma 4.2.** The upper bound of $p_{(a,b]}^{\text{upper}}$ given in (4.1) is strictly
964 decreasing in $\bar{\sigma}$ and converges to zero when $\bar{\sigma}$ goes to infinity. This guarantees the
965 existence of $\bar{\sigma}_u$ as a finite value. The existence of $\bar{\sigma}_\ell > 0$ is obvious under A2.
966 A1 guarantees that there exists an open ball B_ℓ with radius $C_\ell(1-r)f_\mu(m)$ such
967 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
968 Proposition 4.1, the success probability with rate r is lower bounded by

$$969 \quad (\text{B.1}) \quad p_r^{\text{succ}}(\bar{\sigma}; m, \Sigma) \geq \kappa^{-d/2} \Phi \left(\mathcal{B} \left(\left(\frac{(2C_u - (1-r)C_\ell)\kappa^{1/2}}{\bar{\sigma}} \right) e_1, \frac{(1-r)C_\ell\kappa^{1/2}}{\bar{\sigma}} \right); 0, \mathbf{I} \right).$$

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

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

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

$$977 \quad (\text{B.2}) \quad V(\theta_t) = \log(f_\mu(m_t)) + v \log^+ \left(\frac{\alpha_\uparrow \ell f_\mu(m_t)}{\sigma_t} \right) + v \log^+ \left(\frac{\sigma_t}{\alpha_\downarrow u f_\mu(m_t)} \right).$$

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

$$980 \quad V(\theta_{t+1}) = \log f_\mu(m_{t+1}) + v \underbrace{\log^+ \frac{\ell f_\mu(m_{t+1})}{\sigma_t} 1_{\{\sigma_{t+1} > \sigma_t\}}}_{P_2} + v \underbrace{\log^+ \frac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} 1_{\{\sigma_{t+1} < \sigma_t\}}}_{P_3}$$

$$981 \quad + v \underbrace{\log^+ \frac{\alpha_\uparrow \sigma_t}{\alpha_\downarrow u f_\mu(m_{t+1})} 1_{\{\sigma_{t+1} > \sigma_t\}}}_{P_4} + v \underbrace{\log^+ \frac{\sigma_t}{u f_\mu(m_t)} 1_{\{\sigma_{t+1} < \sigma_t\}}}_{P_5}.$$

982 We want to estimate the conditional expectation

$$983 \quad (\text{B.3}) \quad \mathbb{E}[\max\{V(\theta_{t+1}) - V(\theta_t), -A\} \mid \theta_t].$$

984 We partition the possible values of θ_t into three sets: first the set of θ_t such that
985 $\sigma_t < \ell f_\mu(m_t)$ (σ_t is small), second the set of θ_t such that $\sigma_t > u f_\mu(m_t)$ (σ_t is large),
986 and last the set of θ_t such that $\ell f_\mu(m_t) \leq \sigma_t \leq u f_\mu(m_t)$ (reasonable σ_t). In the
987 following, we bound (B.3) for each of the three cases and in the end our bound B
988 will equal the minimum of the three bounds obtained for each case.

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

$$993 \quad P_3 = v \log^+ \left(\frac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} \right) 1_{\{\sigma_{t+1} < \sigma_t\}},$$

$$994 \quad P_4 = v \left[\log \left(\frac{\alpha_\uparrow \sigma_t}{\alpha_\downarrow u f_\mu(m_t)} \right) - \log \left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)} \right) \right] 1_{\left\{ \frac{\alpha_\downarrow u f_\mu(m_{t+1})}{\alpha_\uparrow \sigma_t} < 1 \right\}} 1_{\{\sigma_{t+1} > \sigma_t\}}.$$

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

999

$$\begin{aligned}
 1000 \quad (B.4) \quad \Delta_t &\leq \left(1 - v \mathbb{1} \left\{ \frac{\alpha_{\downarrow} u f_{\mu}(m_t)}{\alpha_{\uparrow} \sigma_t} < 1 \right\} \mathbb{1} \{ \sigma_{t+1} > \sigma_t \} \right) \log \left(\frac{f_{\mu}(m_{t+1})}{f_{\mu}(m_t)} \right) \\
 1001 &\quad + v \log^+ \left(\frac{\alpha_{\uparrow} \ell f_{\mu}(m_t)}{\alpha_{\downarrow} \sigma_t} \right) \mathbb{1} \{ \sigma_{t+1} < \sigma_t \} + v \log^+ \left(\frac{\alpha_{\uparrow} \sigma_t}{\alpha_{\downarrow} u f_{\mu}(m_t)} \right) \mathbb{1} \{ \sigma_{t+1} > \sigma_t \} \\
 1002 &\leq (1 - v) \log \frac{f_{\mu}(m_{t+1})}{f_{\mu}(m_t)} + v \log^+ \frac{\alpha_{\uparrow} \ell f_{\mu}(m_t)}{\alpha_{\downarrow} \sigma_t} \mathbb{1} \{ \sigma_{t+1} < \sigma_t \} + v \log^+ \frac{\alpha_{\uparrow} \sigma_t}{\alpha_{\downarrow} u f_{\mu}(m_t)} \mathbb{1} \{ \sigma_{t+1} > \sigma_t \} . \\
 1003
 \end{aligned}$$

1004 The truncated one-step change $\max\{\Delta_t, -A\}$ is upper bounded by

1005

$$\begin{aligned}
 1006 \quad (B.5) \quad \max\{\Delta_t, -A\} &\leq (1 - v) \max \left\{ \log \left(\frac{f_{\mu}(m_{t+1})}{f_{\mu}(m_t)} \right), -\frac{A}{1-v} \right\} \\
 1007 &\quad + v \log^+ \left(\frac{\alpha_{\uparrow} \ell f_{\mu}(m_t)}{\alpha_{\downarrow} \sigma_t} \right) \mathbb{1} \{ \sigma_{t+1} < \sigma_t \} + v \log^+ \left(\frac{\alpha_{\uparrow} \sigma_t}{\alpha_{\downarrow} u f_{\mu}(m_t)} \right) \mathbb{1} \{ \sigma_{t+1} > \sigma_t \} . \\
 1008
 \end{aligned}$$

1009 To consider the expectation of the above upper bound, we need to compute the
 1010 expectation of the maximum of $\log \left(\frac{f_{\mu}(m_{t+1})}{f_{\mu}(m_t)} \right)$ and $-\frac{A}{1-v}$. Let $a \leq 0$ and $b \in \mathbb{R}$
 1011 then $\max(a, b) = a \mathbb{1} \{ a > b \} + b \mathbb{1} \{ a \leq b \} \leq b \mathbb{1} \{ a \leq b \}$. Applying this and taking the
 1012 conditional expectation, a trivial upper bound for the conditional expectation of
 1013 $\max \left\{ \log \left(\frac{f_{\mu}(m_{t+1})}{f_{\mu}(m_t)} \right), -\frac{A}{1-v} \right\}$ is $-\frac{A}{1-v}$ times the probability of $\log \left(\frac{f_{\mu}(m_{t+1})}{f_{\mu}(m_t)} \right)$ being
 1014 no greater than $-\frac{A}{1-v}$. The latter condition is equivalent to $f_{\mu}(m_{t+1}) \leq (1-r)f_{\mu}(m_t)$
 1015 corresponding to successes with rate $r = 1 - \exp \left(-\frac{A}{1-v} \right)$ or better. That is,

$$1016 \quad (B.6) \quad (1 - v) \mathbb{E} \left[\max \left\{ \log \left(\frac{f_{\mu}(m_{t+1})}{f_{\mu}(m_t)} \right), -\frac{A}{1-v} \right\} \right] \leq -A p_r^{\text{succ}} \left(\frac{\sigma_t}{f_{\mu}(m_t)}; m_t, \Sigma_t \right) .$$

1017 Note also that the expected value of $\mathbb{1} \{ \sigma_{t+1} > \sigma_t \}$ is the success probability, namely,
 1018 $p_0^{\text{succ}} \left(\frac{\sigma_t}{f_{\mu}(m_t)}; m_t, \Sigma_t \right)$. We obtain an upper bound for the conditional expectation of
 1019 $\max\{\Delta_t, -A\}$ in the case of reasonable σ_t as

1020

$$\begin{aligned}
 1021 \quad (B.7) \quad \mathbb{E} [\max\{\Delta_t, -A\} | \theta_t] &\leq -A p_r^{\text{succ}} \left(\frac{\sigma_t}{f_{\mu}(m_t)}; m_t, \Sigma_t \right) \\
 1022 &\quad + \left(\log \left(\frac{\alpha_{\uparrow}}{\alpha_{\downarrow}} \right) + \underbrace{\log \left(\frac{\ell f_{\mu}(m_t)}{\sigma_t} \right)}_{\leq 0} \right) v \left(1 - p_0^{\text{succ}} \left(\frac{\sigma_t}{f_{\mu}(m_t)}; m_t, \Sigma_t \right) \right) \\
 1023 &\quad + \left(\log \left(\frac{\alpha_{\uparrow}}{\alpha_{\downarrow}} \right) + \underbrace{\log \left(\frac{\sigma_t}{u f_{\mu}(m_t)} \right)}_{\leq 0} \right) v p_0^{\text{succ}} \left(\frac{\sigma_t}{f_{\mu}(m_t)}; m_t, \Sigma_t \right) \leq -A p_r^* + v \log \left(\frac{\alpha_{\uparrow}}{\alpha_{\downarrow}} \right) . \\
 1024
 \end{aligned}$$

1025 *Small σ_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.
 1026 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
 1027 2nd summand in (B.2) is positive for $V(\theta_{t+1})$ as well. If $\sigma_{t+1} > \sigma_t$, any regime can

1028 happen. Then, $V(\theta_{t+1}) - V(\theta_t) =$

1029 $= \log \frac{f_\mu(m_{t+1})}{f_\mu(m_t)} - v \log \frac{\alpha_\uparrow \ell f_\mu(m_t)}{\sigma_t} + v \log \frac{\ell f_\mu(m_{t+1})}{\sigma_t} 1 \left\{ \frac{\ell f_\mu(m_{t+1})}{\sigma_t} > 1 \right\} 1 \{ \sigma_{t+1} > \sigma_t \}$

1030 $+ v \log \frac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} 1 \left\{ \frac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} > 1 \right\} 1 \{ \sigma_{t+1} < \sigma_t \}$

1031 $+ v \log \frac{\alpha_\uparrow \sigma_t}{\alpha_\downarrow u f_\mu(m_{t+1})} 1 \left\{ \frac{\alpha_\downarrow u f_\mu(m_{t+1})}{\alpha_\uparrow \sigma_t} < 1 \right\} 1 \{ \sigma_{t+1} > \sigma_t \}$

1032 $= \log \left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)} \right) \left[1 + v \left(1 \left\{ \frac{\ell f_\mu(m_{t+1})}{\sigma_t} > 1 \right\} - 1 \left\{ \frac{\alpha_\downarrow u f_\mu(m_{t+1})}{\alpha_\uparrow \sigma_t} < 1 \right\} \right) 1 \{ \sigma_{t+1} > \sigma_t \} \right]$

1033 $- v \log \left(\frac{\alpha_\downarrow u f_\mu(m_t)}{\alpha_\uparrow \sigma_t} \right) 1 \left\{ \frac{\alpha_\downarrow u f_\mu(m_{t+1})}{\alpha_\uparrow \sigma_t} < 1 \right\} 1 \{ \sigma_{t+1} > \sigma_t \}$

1034 $- v \log \left(\frac{\ell f_\mu(m_t)}{\sigma_t} \right) \left[1 - 1 \left\{ \frac{\ell f_\mu(m_{t+1})}{\sigma_t} > 1 \right\} 1 \{ \sigma_{t+1} > \sigma_t \} - 1 \left\{ \frac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} > 1 \right\} 1 \{ \sigma_{t+1} < \sigma_t \} \right]$

1035 $- v \left(\log(\alpha_\uparrow) - \log \left(\frac{\alpha_\uparrow}{\alpha_\downarrow} \right) 1 \left\{ \frac{\alpha_\uparrow \ell f_\mu(m_t)}{\alpha_\downarrow \sigma_t} > 1 \right\} 1 \{ \sigma_{t+1} < \sigma_t \} \right) .$

1037 On the RMS of the above equality, the first term is guaranteed to be non-positive
 1038 since $v \in (0, 1)$. The second and third terms are non-positive as well since $\frac{\alpha_\downarrow u f_\mu(m_t)}{\alpha_\uparrow \sigma_t} >$
 1039 $\frac{\alpha_\downarrow u}{\alpha_\uparrow \ell} \geq 1$ 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
 1040 term provides an upper bound. Altogether, we obtain

1041
$$\Delta_t = V(\theta_{t+1}) - V(\theta_t) \leq -v (\log(\alpha_\uparrow) - \log(\alpha_\uparrow/\alpha_\downarrow) 1 \{ \sigma_{t+1} < \sigma_t \}) .$$

1042 Note that the RHS is larger than $-A$ since it is lower bounded by $-v \log(\alpha_\uparrow)$ and
 1043 $v \leq A/\log(\alpha_\uparrow)$. Then, the conditional expectation of $\max\{\Delta_t, -A\}$ is

1044

1045 (B.8) $\mathbb{E} [\max\{\Delta_t, -A\} | \mathcal{F}_t] \leq -v \left(\log \left(\frac{\alpha_\uparrow}{\alpha_\downarrow} \right) p_0^{\text{succ}} \left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t \right) + \log(\alpha_\downarrow) \right)$

1046 $\leq -v \left(\log \left(\frac{\alpha_\uparrow}{\alpha_\downarrow} \right) p_\ell + \log(\alpha_\downarrow) \right) = -v \log \left(\frac{\alpha_\uparrow}{\alpha_\downarrow} \right) (p_\ell - p_{\text{target}}) = -v \frac{p_\ell - p_u}{2} \log \left(\frac{\alpha_\uparrow}{\alpha_\downarrow} \right) .$

1048 Here we used $\mathbb{E}[1\{\sigma_{t+1} < \sigma_t\} | \mathcal{F}_t] = 1 - p_0^{\text{succ}} \left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t \right)$ for the first inequality,
 1049 $p_0^{\text{succ}} \left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t \right) > p_\ell$ for the second inequality, and $p_{\text{target}} = \log \left(\frac{1}{\alpha_\downarrow} \right) / \log \left(\frac{\alpha_\uparrow}{\alpha_\downarrow} \right) =$
 1050 $(p_u + p_\ell)/2$ for the last equality.

1051 *Large σ_t case:* $\frac{f_\mu(m_t)}{\sigma_t} < \frac{1}{u}$. Since $\frac{f_\mu(m_{t+1})}{\sigma_{t+1}} \leq \frac{f_\mu(m_t)}{\alpha_\downarrow \sigma_t} < \frac{1}{\alpha_\downarrow u}$, the 3rd summand in
 1052 (B.2) is positive in both $V(\theta_t)$ and $V(\theta_{t+1})$. For the 2nd summand in (B.2), recall that
 1053 $\alpha_\uparrow \ell f_\mu(m_t)/\sigma_t < \alpha_\uparrow \ell/u \leq \alpha_\downarrow < 1$ since we have assumed that $u/\ell \geq \alpha_\uparrow/\alpha_\downarrow$. Hence,
 1054 for $V(\theta_t)$ the 2nd summand in (B.2) is zero. Also, $\alpha_\uparrow \ell \|m_{t+1}\|/\sigma_{t+1} \leq \alpha_\uparrow \ell/(\alpha_\downarrow u) =$
 1055 $(\alpha_\uparrow/\alpha_\downarrow)\ell/u \geq 1$ and thus for $V(\theta_{t+1})$ the 2nd summand in (B.2) also equals 0. We
 1056 obtain

1057
$$V(\theta_{t+1}) - V(\theta_t) = (1 - v) (\log(f_\mu(m_{t+1})) - \log(f_\mu(m_t))) + v \log(\sigma_{t+1}/\sigma_t) .$$

1058 The first term on the RHS is guaranteed to be non-positive since $v < 1$, yielding
 1059 $\Delta_t \leq v \log(\sigma_{t+1}/\sigma_t)$. On the other hand,

1060
$$v \log(\sigma_{t+1}/\sigma_t) = v (\log(\alpha_\uparrow) 1 \{ \sigma_{t+1} > \sigma_t \} + \log(\alpha_\downarrow) 1 \{ \sigma_{t+1} < \sigma_t \})$$

1061 $= v (\log(\alpha_\uparrow/\alpha_\downarrow) 1 \{ \sigma_{t+1} > \sigma_t \} - \log(1/\alpha_\downarrow))$

1062 $\geq -v \log(1/\alpha_\downarrow) \geq -A ,$

1064 where the last inequality comes from the prerequisite $v \leq A/\log(1/\alpha_\downarrow)$. Hence,

$$1065 \quad \max\{\Delta_t, -A\} \leq \max\{v \log(\sigma_{t+1}/\sigma_t), -A\} = v \log(\sigma_{t+1}/\sigma_t) .$$

1066 Then, the conditional expectation of $\max\{\Delta_t, -A\}$ is

$$1067 \quad \mathbb{E}[\max\{\Delta_t, -A\} | \theta_t] \leq v \left(\log(\alpha_\downarrow) + \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) p_0^{\text{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right) \right) \\ 1068 \leq v \left(\log(\alpha_\downarrow) + \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) p_u \right) = v \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) (-p_{\text{target}} + p_u) = -v \frac{p_\ell - p_u}{2} \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) .$$

1071 Here we used $p_0^{\text{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right) \leq p_u$.

1072 *Conclusion.* Inequalities (B.7)–(B.9) together cover all possible cases and we
1073 hence obtain (4.7).

1074 Finally, we prove the positivity of B for an arbitrary $A > 0$. **Lemma 4.2**
1075 guarantees the positivity of p_r^* for any choice of A since $r = 1 - \exp(-A/(1 -$
1076 $v)) \in (0, 1)$ for any $A > 0$ and $v < 1$. Therefore, $Ap_r^* > 0$ for any A and $v \leq$
1077 $\min(1, A/\log(1/\alpha_\downarrow), A/\log(\alpha_\uparrow))$. Moreover, for a sufficiently small v , p_r^* is strictly
1078 positive for any $A > 0$. Therefore, one can take a sufficiently small v that satisfies
1079 $Ap_r^* > v \log(\alpha_\uparrow/\alpha_\downarrow)$. The first term in the minimum in (4.7) is positive. The second
1080 term therein is clearly positive for $v > 0$. This completes the proof.

1081 **B.6. Proof of Proposition 4.10.** Consider $d \geq 2$. We set $A = 1/d$. We bound
1082 B from below by taking a specific value for $v \in (0, \min(1, A/\log(1/\alpha_\downarrow), A/\log(\alpha_\uparrow)))$
1083 instead of considering sup for v . Our candidate is $v = \frac{Ap'}{\log(\alpha_\uparrow/\alpha_\downarrow)(2+p_\ell-p_u)}$, where
1084 $p' = \inf_{\bar{\sigma} \in [\ell, u]} p_{r'}(\bar{\sigma})$ and $r' = 1 - \exp\left(-A\left(1 - \frac{1}{d \log(\alpha_\uparrow/\alpha_\downarrow)}\right)^{-1}\right)$. It holds $v < \frac{1}{d \log(\alpha_\uparrow/\alpha_\downarrow)}$
1085 and hence $r' > r$, from which we obtain $p' < p^*$.

1086 We bound the terms in (4.7) as: $Ap^* - v \log(\alpha_\uparrow/\alpha_\downarrow) = \frac{p'}{d} \left(\frac{p^*}{p'} - \frac{2}{2+p_\ell-p_u} \right) \geq$
1087 $\frac{p'}{d} \left(\frac{p_\ell - p_u}{2+p_\ell-p_u} \right)$ 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 $B \geq \frac{p'}{d} \frac{p_\ell - p_u}{2+p_\ell-p_u}$.
1088 Note that one can take $p_\ell - p_u \in \Theta(1)$ since the only condition is $p_{\text{target}} = (p_\ell + p_u)/2 \in$
1089 $\Theta(1)$. To obtain $B \in \Omega(1/d)$, it is sufficient to show $p' \in \Theta(1)$ for $d \rightarrow \infty$.

1090 Fix p_ℓ and p_u independently of d . In the light of Lemma 3.1 in [3], we have that
1091 $p_0 : \mathbb{R}_> \rightarrow (0, 1/2)$ is continuous and strictly decreasing from $1/2$ to 0 for all $d \in \mathbb{N}$.
1092 Therefore, for each $d \in \mathbb{N}$ there exists an inverse map $p_0^{-1} : (0, 1/2) \rightarrow \mathbb{R}_>$. Define
1093 $\hat{\sigma}_\ell^d = dV_d p_0^{-1}(p_\ell)$ and $\hat{\sigma}_u^d = dV_d p_0^{-1}(p_u)$ for each $d \in \mathbb{N}$. It follows from Lemma 3.2
1094 in [3] that $p_0^{\text{lim}} : \bar{\sigma} \mapsto \lim_{d \rightarrow \infty} p_0(\bar{\sigma})$ is also strictly decreasing, hence invertible. The
1095 existence of $\lim_{d \rightarrow \infty} p_0(\cdot)$ is also proved in [3]. We let $\hat{\sigma}_\ell^\infty = (p_0^{\text{lim}})^{-1}(p_\ell)$ and $\hat{\sigma}_u^\infty =$
1096 $(p_0^{\text{lim}})^{-1}(p_u)$. Because of the pointwise convergence of $p_0(\bar{\sigma} = \hat{\sigma}/(dV_d))$ to $p_0^{\text{lim}}(\hat{\sigma})$, we
1097 have $\hat{\sigma}_\ell^d \rightarrow \hat{\sigma}_\ell^\infty$ and $\hat{\sigma}_u^d \rightarrow \hat{\sigma}_u^\infty$ for $d \rightarrow \infty$. Hence, for any $\hat{u} > \hat{\sigma}_u^\infty$ and $\hat{\ell} < \hat{\sigma}_\ell^\infty$ with
1098 $u/\ell \geq \alpha_\uparrow/\alpha_\downarrow$, there exists $D \in \mathbb{N}$ such that for all $d \geq D$ we have $\hat{u} > \hat{\sigma}_u^d$ and $\hat{\ell} < \hat{\sigma}_\ell^d$.
1099 Now we fix \hat{u} and $\hat{\ell}$ in this way. This amounts to selecting $u = dV_d \hat{u}$ and $\ell = dV_d \hat{\ell}$.

1100 We have $\lim_{d \rightarrow \infty} dr' = 1$ since $\lim_{d \rightarrow \infty} d \log(\alpha_\uparrow/\alpha_\downarrow) = \infty$ and hence according to
1101 Lemma 3.2 in [3] we have

$$1102 \quad \liminf_{d \rightarrow \infty} p' = \liminf_{d \rightarrow \infty} \min_{\bar{\sigma} \in [\ell, u]} \{p_{r'}(\bar{\sigma})\} = \liminf_{d \rightarrow \infty} \min_{\hat{\sigma} \in [\hat{\ell}, \hat{u}]} p_{r'}\left(\frac{\hat{\sigma}}{dV_d}\right) \\ 1103 \stackrel{(*)}{=} \min_{\hat{\sigma} \in [\hat{\ell}, \hat{u}]} \lim_{d \rightarrow \infty} \left(p_{r'}\left(\frac{\hat{\sigma}}{dV_d}\right) \right) = \min_{\hat{\sigma} \in [\hat{\ell}, \hat{u}]} \Psi\left(-\frac{1}{\hat{\sigma}} - \frac{\hat{\sigma}}{2}\right) ,$$

1105 where the equality (\star) follows from the pointwise convergence of $p_{r'}$ to $\lim_{d \rightarrow \infty} p_{r'}$
 1106 and the continuity of $p_{r'}$ and $\lim_{d \rightarrow \infty} p_{r'}$.² This completes the proof.

²Let $\{f_n : n \geq 1\}$ 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 $f_n(x_n) \leq f_n(x^*)$, hence $\liminf_n f_n(x_n) \leq \liminf_n f_n(x^*) = f(x^*)$. Let $\{n_i : i \geq 1\}$ be the sub-sequence of the indices such that $\liminf_n f_n(x_n) = \lim_i f_{n_i}(x_{n_i})$. Since $\{x_{n_i} : i \geq 1\}$ is a bounded sequence, Bolzano-Weierstraß theorem provides a convergent sub-sequence $\{x_{n_{i_k}} : k \geq 1\}$ 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 $\{f_n : n \geq 1\}$ 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, $\liminf_n f_n(x_n) = f(x_*) \leq f(x^*)$. Since x^* is the minimizer of f in $[\ell, u]$ and $x_* \in [\ell, u]$, it must hold $f(x_*) \geq f(x^*)$. Hence, $\liminf_n f_n(x_n) = f(x^*)$.