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. 2020. hal-02941429v1

HAL Id: hal-02941429

https://inria.hal.science/hal-02941429v1

Preprint submitted on 17 Sep 2020 (v1), last revised 20 Jan 2022 (v3)

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.

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

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

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

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

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

1. Introduction.

We consider the unconstrained minimization of an objective function $f : \mathbb{R}^d \to \mathbb{R}$ without the use of derivatives. We more precisely assume that the algorithm minimizing the function f sees it as a *black-box oracle* that can be queried at points $x \in \mathbb{R}^d$ and returns as response the function value $f(x)$, but no gradient. This problem statement is referred to as zero-order black-box optimization [12, 53, 54] or derivative-free optimization [16]. This context is particularly useful for solving real-world applications where the function is computed for instance through a simulation where an estimate of the gradient is hard to derive or when differentiability cannot be assumed.

Black-box problems can be advantageously approached by randomized algorithms that can typically be more robust to noise, non-convexity and irregularities of the objective function than deterministic algorithms. Recently, there has been a vivid interest in randomized derivative-free algorithms: direct search methods have been analyzed assuming that the set of polling directions includes a descent direction with probability p [26]; randomized trust-region methods have been investigated [9, 27]; randomized model-based method with line search have been analyzed under the assumption that the algorithm under study delivers a sufficiently good model with high enough probability [13], and a similar analysis was undertaken in [57] with a focus on line search. Additionally, Nesterov and Spokoiny [55] have analyzed a randomized method where the directions follow a Gaussian distributions. For strongly-convex

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

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

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

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

functions with Lipschitz gradient they obtain a runtime bound of order $\mathcal{O}\big(d \cdot \log(1/\epsilon)\big)$ instead of $\mathcal{O}\big(d^2 \cdot \log(1/\epsilon)\big)$ for deterministic direct search [44]. The quadratic scaling is due to the problem that finding a suitable descent direction takes linear time. Note however that the method analyzed in [55] does not assume a black-box setting and explicitly uses the Lipschitz constant and the target accuracy ϵ for choosing the parameters of the algorithm. We refer to [46] for an in-depth survey including the references of this paragraph and additional ones.

Random Pursuit [62] is another randomized derivative-free method where a line search is performed starting from an incumbent solution in a direction sampled uniformly at random. Stich et al. analyzed the random pursuit with approximate line search on convex functions, which realizes a geometric decrease of the function value regret $f(x_t) - f(x^*)$ (and hence linear convergence) on strongly convex functions for exact line search and also for the more realistic case of approximate line search with bounded relative error. More recently, Golovin et al. [25] analyzed a randomized algorithm where a search direction is sampled from the standard Gaussian. For the step-size, they consider as approximate line search all possible value of a grid spanning an interval with uniform spacing on a log-scale. They show a geometric decrease of the function value regret, up to a constant defined by the minimum step-size, on strongly convex functions with Lipschitz smooth gradients.

All of the previously cited studies have either taken well known deterministic derivative-free algorithms and modified their convergence and complexity analysis to handle randomized components, or built novel randomized algorithms with a clear route towards (tractable) convergence proofs and complexity bounds in mind. In some cases, the analysis concerns parametrized frameworks rather than concrete algorithms that could be directly implemented and used.

We take here a different approach. We investigate practical algorithms that belong to the class of Evolution Strategies (ES) that are among the oldest randomized derivative-free or zero-order black-box methods [17, 58, 61]. In contrast to previously cited studies, establishing formal convergence proofs has never been the primary concern when designing ES methods. Consequently, establishing their convergence turns out to be often a difficult task requiring sometimes to extend current mathematical frameworks [3, 8, 14].

Nowadays, a specific ES called covariance-matrix-adaptation ES (CMA-ES) [35] is among the best solvers to address *difficult* black-box problems. It performs well on many ill-conditioned, non-convex, non-smooth, and non-separable problems [34, 60] with various published successful applications in different domains by researchers unrelated to the algorithm designers [5, 11, 21–23, 29, 45, 50, 66, 67][1]. The algorithm is affine-invariant and implements complex adaptation mechanisms for the sampling covariance matrix and step-size. It learns second order information of the objective function [33]. Part of the covariance matrix update derives from a stochastic natural gradient step of an optimization problem formulated on the statistical manifold formed by the family of Gaussian distributions [4, 56, 68][2]. The CMA-ES should always be preferred over the ES variant we analyze in this paper for solving real-world

[1]The pycma python implementation has regularly more than 2000 downloads per day according to https://pypistats.org/packages/cma as of May 2020.

[2]This natural gradient should not be confused with the gradient of the objective function. Additionally, the vector formed by the update of the mean vector $m_{t+1} - m_t$ is not meant to be proportional to the gradient of f in m_t. Indeed the step-size is typically large compared to what would be needed to obtain a good approximation of the gradient. This feature helps to solve problems with local irregularities like noisy functions with noise-to-signal ratio bounded.

problems. Yet, even the simple ES variants analyzed achieve faster linear convergence on well-conditioned problems when compared to algorithms with established complexity analysis (see [62, Table 6.3 and Figure 6.1] and [8, Figure B.4] where the random pursuit algorithm and the (1+1)-ES algorithms are compared, and also Appendix A).

Because Evolution Strategies are important randomized derivative-free algorithms in practice, it is essential to study them from a theoretical convergence perspective, which is the aim of this paper.

In this context, we analyze the arguably simplest and oldest adaptive ES, denoted (1+1)-ES. Its pseudo code is given below:

Algorithm 1.1 (1+1)-ES with 1/5 success rule

input $m \in \mathbb{R}^d$, $\sigma > 0$
set $\alpha_\uparrow \approx 1.5$, $\alpha_\downarrow \approx (1.5)^{-1/4}$
while stopping criterion not met **do**
  $x = m + \sigma \cdot \mathcal{N}(0, I_d)$ ▷ sample candidate solution following $\mathcal{N}(m, \sigma^2 I_d)$
  if $f(x) \leqslant f(m)$ **then**
    $m \leftarrow x$ ▷ move to the better solution
    $\sigma \leftarrow \alpha_\uparrow \sigma$ ▷ increase step-size
  else
    $\sigma \leftarrow \alpha_\downarrow \sigma$ ▷ decrease step-size .

Our analysis will also allow to cover an (abstract) adaptation of the covariance matrix provided the condition number of the matrix stays bounded.

The pseudo-code above shares some similarities with the simplified direct search whose complexity analysis has been presented in [44]. Yet there are some fundamental differences that we wish to highlight:

- The algorithm is comparison-based: the decisions to accept a candidate solution are based only on the order relation of $f(x)$ and $f(m)$. In contrast, in direct search algorithms, the acceptance is based on the sufficient decrease condition $f(x) \leqslant f(m) - c\sigma^2$ where c is a constant which is a parameter of the algorithm. In ES, sufficient decrease is not guaranteed when accepting a better solution. On the positive side, the decision is invariant to strictly increasing transformation of f.
- In contrast to the simplified direct-search algorithm, the step-size can increase and can thus recover from bad initial choices of the step-size. This recovery is practically fast. Hence the method does not require any tuning of the initial parameters to work efficiently.
- Algorithm 1.1 is rotational invariant. In contrast, direction search methods construct candidate solutions along the vectors of a predefined set. As a consequence, performance can be very different if the function is separable with respect to the elements of the predefined set of vectors or not.

The comparison-based nature of the (1+1)-ES is important as it implies invariance to strictly increasing transformations of the objective function. Invariance properties naturally add robustness to a method. Also, practically, the method does not require any tuning of the essential parameters and converges globally, independent of the initial conditions. The absence of parameters that need problem-specific tuning is shared by CMA-ES, and it is probably a key to its success. Yet, while powerful in practice, this property poses great challenges for the analysis because we cannot rely on sufficient decrease of f.

Prior theoretical studies of the (1+1)-ES with 1/5 success rule have established the global linear convergence on differentiable positively homogeneous functions (composed with a strictly increasing function) with a single optimum [7, 8]. Those results establish the almost sure linear convergence for all initial states (σ and m). They however do not provide the dependency of the convergence rate with respect to the dimension. A more specific study on the sphere function $f(x) = \frac{1}{2}\|x\|^2$ establishes lower and upper bounds on the expected hitting time of an ϵ-ball of the optimum in $\Theta(\log(d\|m_0 - x^*\|/\epsilon))$, where x^* is the optimum of the function [2]. Prior to that, a variant of the $(1+1)$-ES with one-fifth success rule had been analyzed on the sphere and certain convex quadratic functions establishing bounds on the expected hitting time or with overwhelming probability in $\Theta(\log(\kappa_f d\|m_0 - x^*\|/\epsilon))$, where κ_f is the condition number (the ratio between the greatest and smallest eigenvalues) of the Hessian [39–42].

Recently, the class of functions where the convergence of the (1+1)-ES was proven has been extended to continuously differentiable functions. This analysis does not address the question of linear convergence focusing only on the convergence (which is possibly sublinear) [24].

In this context, our main objective is to establish linear convergence of a generalized version of the (1+1)-ES algorithm on a function class which is as wide as possible. Our main contribution is as follows:

For a generalized version of the (1+1)-ES with one-fifth success rule, we prove bounds on the expected hitting time akin to linear convergence, i.e., hitting an ϵ-ball in $\Theta(\log \|m_0 - x^*\|/\epsilon)$ iterations on a quite general class of functions. This class of functions includes all composites of L-smooth strongly convex functions with a strictly increasing transformation. This latter transformation allows to include some non-continuous functions, and even functions with non-smooth level sets. We additionally deduce linear convergence with probability one.

Our analysis relies on finding an appropriate Lyapunov function with lower and upper-bounded expected drift. It is building on classical fundamental ideas presented by Hajek [30] and widely used to analyze stochastic hill-climbing algorithms on discrete search spaces [48].

Preliminary versions of this work were published in [2, 52]. Those studies were concerned with an algorithm without covariance matrix adaptation. Here we extend the analysis accordingly and also include almost sure convergence results. Additionally the presentation of this paper includes an introduction to the methods analyzed, put in the context of derivative-free optimization, as well as thorough discussions of the assumptions and of the obtained results.

This paper is organized as follows. In Section 2 we present the class of algorithms analyzed, we provide central definitions, establish preliminary results, and finish by presenting and discussing the class of objective functions under study. In Section 3 we present our main methodology based on drift analysis to prove bounds on the expected first hitting time of an ϵ-ball of the optimum. We prove lower-bounds and upper bounds akin to linear convergence in Section 4.

Notation. Throughout the paper, we use the following notations. The set of natural numbers $\{1, 2, \dots, \}$ is denoted $\mathbb{N}$. The real field is denoted by $\mathbb{R}$. Open, closed, and left open intervals on $\mathbb{R}$ are denoted by $(\cdot, \cdot)$, $[\cdot, \cdot]$, and $(\cdot, \cdot]$, respectively. The set of strictly positive real numbers is denoted by $\mathbb{R}_>$. The absolute value on $\mathbb{R}$ is denoted by $|\cdot|$. The Euclidean d-dimensional real vector space is denoted by $\mathbb{R}^d$. The Euclidean norm on $\mathbb{R}^d$ is denoted by $\|\cdot\|$. Open and closed balls with center c and radius r are

denoted as $\mathcal{B}(c,r) = \{x \in \mathbb{R}^d : \|x-c\| < r\}$ and $\bar{\mathcal{B}}(c,r) = \{x \in \mathbb{R}^d : \|x-c\| \leqslant r\}$, respectively. Lebesgue measures on $\mathbb{R}$ and $\mathbb{R}^d$ are both denoted by the same symbol μ. A multivariate normal distribution with mean m and covariance matrix Σ is denoted by $\mathcal{N}(m,\Sigma)$. Its probability measure and its induced probability density under Lebesgue measure are denoted by $\Phi(\cdot; m, \Sigma)$ and $\varphi(\cdot; m, \Sigma)$. The indicator function of a set or condition C is denoted by $1_{\{C\}}$. For a random object X on some probability space $\mathcal{X} \sim P$ following a probability measure P, the probability and the expectation are denoted by $\Pr_{X\sim P}$ and $\mathbb{E}_{X\sim P}$, and we drop $X \sim P$ from the subscripts unless ambiguous. We use Bachmann-Landau notations: o, O, Θ, Ω, ω.

2. Algorithm, Definitions and Objective Function Assumptions. In this section we present the (1+1)-ES$_\kappa$ with a success-based step-size adaptation and a covariance matrix adaptation with bounded condition number that we will analyze. We then introduce preliminary definitions for the analysis and state the objective function assumptions as well as discuss their relation to typical classes of functions where numerical optimization algorithms are usually analyzed.

2.1. Algorithm: (1+1)-ES with Success-based Step-size Control. We analyze in this paper a generalized version of the (1+1)-ES with one-fifth success rule sketched in the introduction (Algorithm 1.1) which implements one of the oldest idea to adapt the step-size in randomized optimization methods [17, 58, 61]. The specific implementation was proposed in [43]. The algorithm analyzed is written in Algorithm 2.1 where we use notations with a time index that are convenient to analyze the algorithm. Before to explain the generalization of Algorithm 1.1 considered, we explain briefly the (1+1)-ES algorithm sketched in the introduction (we use however notations with a time index). At each iteration, a candidate solution x_t is sampled. It is centered in the current incumbent m_t and follows a multivariate normal distribution with mean vector m_t and covariance matrix equal to $\sigma_t^2 I_d$ where I_d denotes the identity matrix. The candidate solution is accepted, that is m_t becomes x_t, if and only if x_t is better than m_t (i.e. $f(x_t) \leqslant f(m_t)$). In this case, we say that the candidate solution is successful. The step-size σ_t is adapted so as to maintain a probability of success of roughly 1/5. The step-size is increased in case of success (which is an indication that the step-size is likely to be too small) and decreased otherwise. The decrease and increase factors are linked such that a probability of success of roughly 1/5 is maintained. This probability of success has been shown to be close to optimal, which gives nearly optimal (linear) convergence rate on the sphere function [17, 58]. To do so, assume an increase factor denoted $\alpha_\uparrow$ and a decrease factor denoted $\alpha_\downarrow$, we want the following relation to be satisfied:

$$\frac{1}{5}\log(\alpha_\uparrow) + \left(1 - \frac{1}{5}\right)\log(\alpha_\downarrow) = 1 \ . \tag{2.1}$$

If we set $\alpha_\uparrow$ to be equal to 1.5 like in Line 2 of Algorithm 1.1, then this implies that $\alpha_\downarrow = 1.5^{-1/4}$. We will denote as

$$p_{\text{target}} := \frac{\log(1/\alpha_\downarrow)}{\log(\alpha_\uparrow/\alpha_\downarrow)} \ , \tag{2.2}$$

the underlying target success probability that we try to maintain, i.e. that satisfies (2.1) when we replace 1/5 by p_{target}. We generalize the simple algorithm above by replacing the I_d matrix for the sampling of candidate solutions by a covariance matrix Σ_t to be adapted in the set $\mathcal{S}_\kappa$ of positive-definite symmetric matrices with determinant $\det(\Sigma) = 1$ and condition number $\text{Cond}(\Sigma) \leqslant \kappa$. We do not assume any specific

update mechanism for Σ, but we assume that the update of Σ is invariant to any strictly increasing transformation of f. We call such an update comparison-based (see Line 11 of Algorithm 2.1). Then, our algorithm behaves exact-equally on f and on $g \circ f$ for all strictly increasing functions $g : \mathbb{R} \to \mathbb{R}$ $\big($i.e., $g(s) \lesseqgtr g(t) \Leftrightarrow s \lesseqgtr t\big)$. This defines a class of comparison-based randomized algorithms and we denote it as (1+1)-ES$_\kappa$.

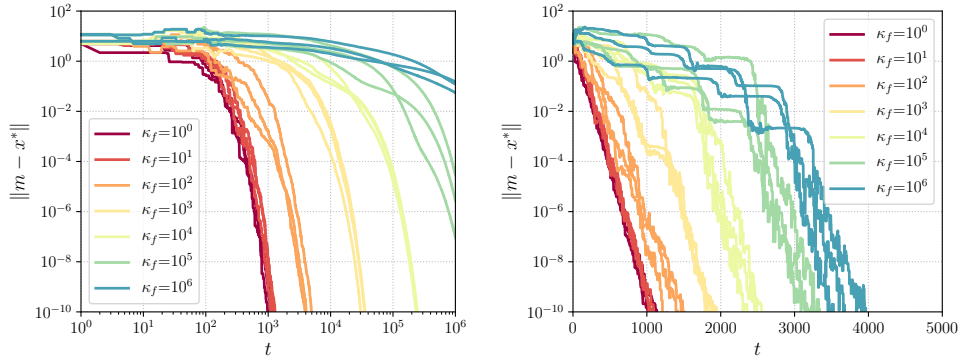
Algorithm 2.1 (1+1)-ES$_\kappa$ with success-based step-size adaptation

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

Note that $\alpha_\uparrow$ and $\alpha_\downarrow$ are not meant to be tuned depending on the function properties. How to choose such constants for $\Sigma_t = I_d$ is well-known and is related to the so-called evolution window [59]. The covariance matrix is updated in $\mathcal{S}_\kappa$. Hereunder we write $\theta = (m, \sigma, \Sigma)$ as the state of the algorithm, $\theta_t = (m_t, \sigma_t, \Sigma_t)$ and the state-space is denoted as Θ.

Remark that while the algorithm analyzed considers an adaptive covariance matrix, the assumption that the condition number of the matrix should stay bounded is restrictive and does not allow to encompass the state-of-the-art evolution strategies that adapt Σ_t with the *covariance matrix adaptation* (CMA) mechanism [31,35]. The resulting algorithm is called CMA-ES. In this case the update of Σ_t [31,35] is affine-invariant which implies that the condition number of Σ_t cannot be bounded. Another major difference between the default CMA-ES algorithm and the one analyzed here relates to the number of sampled solutions at each iteration and the update of m_t. Here we sample one candidate solution whereas $\lambda > 1$ are sampled. The update of the mean also disregards previous solutions and is simply a weighted average of the $\mu = \lfloor \lambda/2 \rfloor$ best solutions. Yet a version of CMA-ES in the (1+1)-ES setting presented here has been proposed [38]. It is however not recommended for practical applications: the default CMA-ES that is sampling λ candidate solutions and where $f(m_t)$ is not monotonous should be preferred.

For the CMA-ES versions mentioned above [31,35,38], it is empirically observed that Σ_t approaches the inverse Hessian $\nabla^2 f(m_t)$ of the objective function up to the scalar factor if the objective function is convex quadratic. Figure 2.1 shows typical runs of the (1+1)-ES and the (1+1)-CMA-ES on a 10-dimensional ellipsoidal function with different condition numbers (denoted as κ_f) of the Hessian. The runtime of (1+1)-ES scales linearly with κ_f (notice that the x-axis is in logarithmic scale), while the runtime of the (1+1)-CMA-ES does not scale. We take the (1+1)-CMA-ES algorithm proposed in [6]. The transient time at the beginning where the (1+1)-CMA-ES adapts Σ is roughly proportional to the logarithm of κ_f. Once the Hessian is well approximated by Σ (up to the scaler factor), it approaches the global optimum



(a) (1+1)-ES (b) (1+1)-CMA-ES

Fig. 2.1: Convergence graphs for the (1+1)-ES and the (1+1)-CMA-ES on 10 dimensional ellipsoidal function $f(x) = \frac{1}{2}\sum_{i=1}^{d} \kappa_f^{\frac{i-1}{d-1}} x_i^2$ with $\kappa_f = 10^0, 10^1, \ldots, 10^6$. The y-axis displays the distance to the optimum and not the function value which explains the non-monotonous behavior when κ_f is larger than 1. We employ the covariance matrix adaptation mechanism proposed by [6], 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 right plot.

geometrically at the same rate for different values of κ_f.

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

2.2. Preliminary definitions.

2.2.1. Spatial Suboptimality Function. The algorithms studied are comparison-based (under the assumption that the update of the covariance matrix is comparison-based) and thus invariant to strictly increasing transformations of f. If the convergence of the algorithms is measured in terms of f, say by investigating the convergence or hitting time of the sequence $f(m_t)$, this will not reflect the invariance to monotonic transformations of f because the smallest iteration t_0 such that $f(m_{t_0}) \leqslant \epsilon$ is not equal to the smallest iteration t_0' such that $g(f(m_{t_0'})) \leqslant \epsilon$ for some $\epsilon > 0$. For this reason, we introduce the quality measure called *spatial suboptimality function* [24]. It is the dth root of the volume of the sub-levelset where the function value is better or equal to $f(x)$:

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

$$f_\mu(x) = \sqrt[d]{\mu\left(f^{-1}\left((-\infty, f(x)]\right)\right)} = \sqrt[d]{\mu\left(\left\{y \in \mathbb{R}^d \,\middle|\, f(y) \leqslant f(x)\right\}\right)} \,. \tag{2.3}$$

Remark that for any f, the suboptimality function f_μ is greater or equal to zero. For

any f and any strictly increasing function $g : \mathrm{Im}(f) \to \mathbb{R}$, f and its composite $g \circ f$ have the same spatial suboptimality function such that hitting time of f_μ smaller than epsilon will be the same for f or $g \circ f$. Moreover, there exists a strictly increasing function g such that $f_\mu(x) = g(f(x))$ holds μ-almost everywhere [24, Lemma 1].

We will investigate the expected first hitting time of $\|m_t - x^*\|$ to $\epsilon > 0$. For this, we will bound the first hitting time of $\|m_t - x^*\|$ to ϵ by the first hitting time of $f_\mu(m_t)$ to a constant times ϵ. To understand why, consider first a strictly convex quadratic function f. As is stated formally in Proposition 2.8, we have $\sqrt{f(x) - f(x^*)} \propto f_\mu(x)$ for all $x \in \mathbb{R}^d$, which is also proven in [1]. This implies that the first hitting time of $f_\mu(m_t)$ translates to the first hitting time of $\sqrt{f(m_t) - f(x^*)}$. Moreover, since $\sqrt{\lambda_{\min}}\|x - x^*\| \leqslant \sqrt{f(x) - f(x^*)} \leqslant \sqrt{\lambda_{\max}}\|x - x^*\|$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimal and maximal eigenvalues of the Hessian of f, it also translates to the first hitting time of $\|m_t - x^*\|$. More generally, we will formalize an assumption on f later on (Assumption A1), which allow us to bound $\|x - x^*\|$ by a constant times $f_\mu(x)$ from above and below (see (2.11)), implying that the first hitting time of $\|m_t - x^*\|$ to ϵ is bounded by that of $f_\mu(m_t)$ to ϵ times a constant.

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

We start with the definition of the *success domain with rate r* and the *success probability with rate r*.[3]

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

$$S_r(m) = \{x \in \mathbb{R}^d \mid f_\mu(x) \leqslant (1 - r) f_\mu(m)\} \ . \tag{2.4}$$

The probability to sample in the r-success domain is called success probability with rate r. When $r = 0$ we simply talk about success probability.

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

$$p_r^{\mathrm{succ}}(\bar\sigma; m, \Sigma) = \Pr_{z \sim \mathcal{N}(0,\Sigma)} \left[m + f_\mu(m)\bar\sigma z \in S_r(m)\right] \ . \tag{2.5}$$

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

[3] For $r = 0$, the success domain $S_0(m)$ is not necessarily equivalent to the sub-levelset $S'_0(m) := \{x \in \mathbb{R}^d \mid f(x) \leqslant f(m)\}$, where it always holds that $S'_0(m) \subseteq S_0(m)$. However, since it is guaranteed that $\mu(S_0(m) \setminus S'_0(m)) = 0$ by [24, 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 $\Pr_{z \sim \mathcal{N}(0,\Sigma)} \left[m + (f_\mu(m)/c)\bar\sigma z \in S'_0(m)\right]$.

LEMMA 2.4. *If* $f(x) = \frac{1}{2}\|x - x^*\|^2$, *then*

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

where $V_d = \frac{\pi^{1/2}}{\Gamma^{1/d}(d/2+1)}$ *and* $e_1 = (1, 0, \ldots, 0)$.

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

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

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

$$p_{(a,b]}^{\mathrm{lower}}(\bar{\sigma}) = \inf_{m\in\mathcal{X}_a^b}\inf_{\Sigma\in\mathcal{S}_\kappa} p_0^{\mathrm{succ}}(\bar{\sigma}; m, \Sigma) \ , \tag{2.6}$$

$$p_{(a,b]}^{\mathrm{upper}}(\bar{\sigma}) = \sup_{m\in\mathcal{X}_a^b}\sup_{\Sigma\in\mathcal{S}_\kappa} p_0^{\mathrm{succ}}(\bar{\sigma}; m, \Sigma) \ . \tag{2.7}$$

A central quantity for our analysis is the limit for $\bar{\sigma}$ to 0 of the success probability $p_0^{\mathrm{succ}}(\bar{\sigma}; m, \Sigma)$. Intuitively, if this limit is too small for a given m (compared to p_{target}), because the ruling principle of the algorithm is to decrease the step-size if the probability of success is smaller than p_{target}, the step-size will be kept decreasing, causing undesired convergence. Following Glasmachers [24], we introduce the concepts of *p-improbability* and *p-criticality*. They are defined in [24] by the probability of sampling a better point from the isotropic normal distribution in the limit of the step-size to zero. Here, we define p-improvability and p-criticality for a general multivariate normal distribution.

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

$$p = \liminf_{\bar{\sigma}\to+0} p_0^{\mathrm{succ}}(\bar{\sigma}; m, \Sigma) \ . \tag{2.8}$$

Otherwise, it is called p*-critical.*

The connection to the classical definition of the critical points for continuously differentiable functions is summarized in the following proposition, which is an extension of Lemma 4 in [24], taking a non-identity covariance matrix into account. Its proof is provided in Appendix B.2.

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

$$p = \Pr_{z\sim\mathcal{N}(0,\Sigma)}\left(z^{\mathrm{T}}\nabla^2 h(m) z \leqslant 0\right) > 0 \tag{2.9}$$

if $\nabla^2 h(x)$ *has at least one strictly negative eigenvalue.*

2.3. Main Assumptions on the Objective Functions. Given a and b positive real numbers satisfying $0 \leqslant a < b \leqslant +\infty$, and a measurable objective function, we denote $\mathcal{X}_a^b$ the set of points of the search space with suboptimality function between a and b, more precisely

$$\mathcal{X}_a^b = \{x \in \mathbb{R}^d | a < f_\mu(x) \leqslant b\} \ .$$

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

A1 We assume that f is a measurable function and that there exists $a \geqslant 0$ and $b > a$ such that there exist universal constants $C_\ell \leqslant C_u$ such that for any $m \in \mathcal{X}_a^b$, there exist an open ball $\mathcal{B}_\ell$ with radius $C_\ell f_\mu(m)$ and a closed ball $\bar{\mathcal{B}}_u$ with radius $C_u f_\mu(m)$ such that $\mathcal{B}_\ell \subseteq \{x \in \mathbb{R}^d \mid f_\mu(x) < f_\mu(m)\}$ and $\{x \in \mathbb{R}^d \mid f_\mu(x) \leqslant f_\mu(m)\} \subseteq \bar{\mathcal{B}}_u$.

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

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

A2 Let f be a measurable function, we assume that there exists $\mathcal{X}_a^b$ and there 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 function f is p-improvable for some $p \geqslant p^{\text{limit}}$, i.e.,

$$\liminf_{\bar{\sigma} \downarrow 0} p^{\text{lower}}_{(a,b]}(\bar{\sigma}) \geqslant p^{\text{limit}} \ . \tag{2.10}$$

The property is illustrated in Figure 4.2. This assumption implies in particular for a continuous function that $\mathcal{X}_a^b$ does not contain any local optimum. This latter assumption is required to obtain global convergence [24, Theorem 2] even without any covariance matrix adaptation (i.e. with $\kappa = 1$) and it can be intuitively understood. If we have a point which is p-improvable with $p < p_{\text{target}}$ and which is not a local minimum of the function, then, starting with a small step-size, the success-based step-size control may keep decreasing the step-size at such a point and the (1+1)-ES$_\kappa$ will prematurely converge to a point which is not a local optimum.

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

$$C_\ell f_\mu(x) \leqslant \|x - x^*\| \leqslant C_u f_\mu(x) \ . \tag{2.11}$$

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

Under A1 and A2, an upper bound on the expected first hitting time of $[0, \epsilon]$ by $f_\mu(m_t)$ will be derived provided $a \leqslant \epsilon \leqslant f_\mu(m_0) \leqslant b$. Yet, the obtained upper bound will translate into linear convergence only if we can choose ϵ as small as possible, that is on functions where $a = 0$. Similarly, it will translate as global linear convergence (i.e. independently of where the starting point m_0 is) only if $b = \infty$. We will also

derive a lower bound for the expected hitting time which holds for any measurable function.

We remark that A1 and A2 satisfied for $a = 0$ allow to include non-smooth functions with non-convex sublevel sets as illustrated in Figures 4.1 and 4.2.

We now give some examples of functions that satisfy A1 and A2 including function classes where linear convergence of numerical optimization algorithms are typically analyzed.

Convex-quadratic functions. We consider composite of convex-quadratic functions as formalized in the next assumption.

A3 We assume that $f = g \circ h$ where $g : \mathrm{Im}(h) \to \mathbb{R}$ is strictly increasing and h is a convex-quadratic function, that is $h(x) = \frac{1}{2}(x - x^*)^{\mathrm{T}} H (x - x^*)$ where H is a $n \times n$ matrix which is symmetric positive definite.

In the next proposition, we give for a function f satisfying A3, the expression of the spatial suboptimality function. We also prove that it satisfies A1 and provide tight bounds for the constants C_ℓ and C_u. We explicit that it satisfies A2. We provide its proof in Appendix B.3.

PROPOSITION 2.8. *Let f be a composite of a convex-quadratic function h by a strictly increasing function g as in A3. Then the spatial suboptimality function $f_\mu(x)$ equals*

$$f_\mu(x) = V_d \cdot \left[\frac{2h(x)}{\det(H)^{1/d}} \right]^{1/2} \tag{2.12}$$

where $V_d = \frac{\pi^{1/2}}{\Gamma^{1/d}(d/2+1)}$ is the dth root of the volume of the d-dimensional unit hypersphere. Assumptions A1 is satisfied for $a = 0$ and $b = \infty$ with

$$C_\ell = V_d^{-1} \big(\det(H)^{1/d} / \lambda_{\max}(H) \big)^{1/2} \quad \text{and} \quad C_u = V_d^{-1} \big(\det(H)^{1/d} / \lambda_{\min}(H) \big)^{1/2} \ .$$

The function f is also measurable and differentiable satisfying thus A2 for $a = 0$ and $b = \infty$.

Convex-quadratic functions are a particular case of strongly-convex and L-smooth functions discussed in the next section.

Strongly-convex and L-smooth functions. The set of strongly-convex and L-smooth functions constitutes a function class where the linear convergence of numerical optimization methods in often analyzed. Let us remind that a function h is L_ℓ-strongly convex on $\mathcal{M} \subseteq \mathbb{R}^d$ if it is differentiable and there exists $L_\ell > 0$ such that for all $x, y \in \mathcal{M}$,

$$h(x) + \langle y - x, \nabla h(x) \rangle + \frac{L_\ell}{2} \|x - y\|^2 \leqslant h(y) \ . \tag{2.13}$$

A differentiable function h is *L_u-smooth on $\mathcal{M}$* if the gradient $\nabla h(x)$ is Lipschitz continuous with Lipschitz constant L_u. If h is L_u-smooth, we have for any $x, y \in \mathcal{M}$,

$$h(y) \leqslant h(x) + \langle y - x, \nabla h(x) \rangle + \frac{L_u}{2} \|x - y\|^2 \ . \tag{2.14}$$

We pose the following assumption.

A4 We assume that $f = g \circ h$ where $g : \mathrm{Im}(h) \to \mathbb{R}$ is strictly increasing and h is L_ℓ-strongly convex on $\mathbb{R}^d$ and L_u-smooth.

Functions satisfying A4 are measurable as the composition of two measurable functions and they satisfy A1 and A2 for $a = 0$ and $b = \infty$. This can be shown as a consequence of the fact that the functions are then quadratically bounded (see Lemma 2.10) given in the next paragraph).

Quadratically bounded functions. We remark that (2.11) implied by A1 can be rewritten into a quadratic upper and lower bound for $f_\mu(x)^2$, i.e.

$$\frac{1}{C_u^2}\|x - x^*\|^2 \leqslant f_\mu(x)^2 \leqslant \frac{1}{C_\ell^2}\|x - x^*\|^2 \ .$$

Conversely, a measurable function which is quadratically bounded will satisfy A1. Indeed let us pose first the quadratically bounded assumption as follows.

A5 We assume that $f = g \circ h$ where g is a strictly increasing function and h is a measurable with the optimum x^* where h is quadratically bounded around x^*, i.e.,

$$\frac{L_\ell}{2}\|x - x^*\|^2 \leqslant h(x) - h(x^*) \leqslant \frac{L_u}{2}\|x - x^*\|^2 \tag{2.15}$$

for $x \in \left\{x : h(x) - h(x^*) \in \left(0, \frac{L_u b^2}{2V_d^2}\right]\right\}$ with $b > 0$.

We note that strongly-convex and L-smooth functions are quadratically bounded functions around the optimum x^* of h where $\nabla f(x) = 0$. The following lemma shows that A5 implies A1 with $a = 0$. The proof of the lemma in presented in Appendix B.4.

LEMMA 2.9. *Let f satisfy A5. Then, f satisfies A1 with $a = 0$ and b given in A5 and*

$$C_\ell = \frac{1}{V_d}\sqrt{\frac{L_\ell}{L_u}} \qquad \text{and} \qquad C_u = \frac{1}{V_d}\sqrt{\frac{L_u}{L_\ell}} \ .$$

The assumption of quadratically bounded functions includes non-convex functions. In particular, the function can oscillate and we can have local optima.

As a direct consequence of Proposition 2.7, we know that a continuously differentiable function with a unique critical point satisfies A2. Hence a differentiable quadratically bounded function with a unique optimum x^* will satisfy both A1 and A2. We pose this as an assumption.

A6 We assume that $f = g \circ h$ where $g : \mathrm{Im}(h) \to \mathbb{R}$ is strictly increasing and h is continuously differentiable and quadratically bounded satisfying (2.15) around x^* assumed to be the unique optimum of f.

Combining Proposition 2.7 and Lemma 2.9, functions satisfying A6 satisfy A1 and A2. Also, since strongly-convex and L-smooth functions are quadratically bounded, functions satisfying A4 satisfy A1 and A2. We formalize this result in the next lemma.

LEMMA 2.10. *If A6 holds, then f satisfies A1 and A2 with $a = 0$ and b given in A5. Moreover, if A4 holds, then f satisfies A1 and A2 with $a = 0$ and $b = \infty$.*

Positively Homogeneous Functions. A positively homogeneous function h with a unique optimum x^* satisfies the following property:

$$h(x^* + \gamma x) = h(x^*) + \gamma\left(h(x^* + x) - h(x^*)\right) \tag{2.16}$$

for $\gamma > 0$ The levelsets of a positively homogeneous function are all geometrically similar around x^*. The sublevel sets can be non-convex such that the function class includes non quasi-convex functions (see [8, Figure 3.1]). Linear convergence of evolution strategies including the (1+1)-ES has been proven on such functions [7, 8].

A7 We assume that $f = g \circ h$ where h is positively homogeneous with a unique optimum x^* satisfying (2.16) and continuously differentiable.

PROPOSITION 2.11. *Let f be positively homogeneous satisfying A7, then the sub-optimality function $f_\mu(x)$ is proportional to $h(x) - h(x^*)$ and satisfies A1 for $a = 0$ and $b = \infty$ with $C_u = \sup\{\|x - x^*\| : f_\mu(x) = 1\}$ and $C_\ell = \inf\{\|x - x^*\| : f_\mu(x) = 1\}$. It also satisfies A2 for $a = 0$ and $b = \infty$.*

The proof of this proposition is provided in Appendix B.5.

3. Methodology: Additive Drift on Unbounded Continuous Domains. We analyze the linear convergence by bounding the expected first hitting time of a ball of radius ϵ around the optimum. In Section 3.1, we define the first hitting of a process and connect bounds on this hitting time and linear convergence. In Section 3.2.1 and Section 3.2.2 we present two theorems that translate drift conditions into upper bounds and lower bounds on the expected first hitting time.

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

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

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

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

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

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

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

When we have an equality in the previous statement, we say that $\exp(-C)$ is the convergence rate.[4]

[4]The link between (3.1) and (3.2) can be easily understood. Assume a deterministic sequence

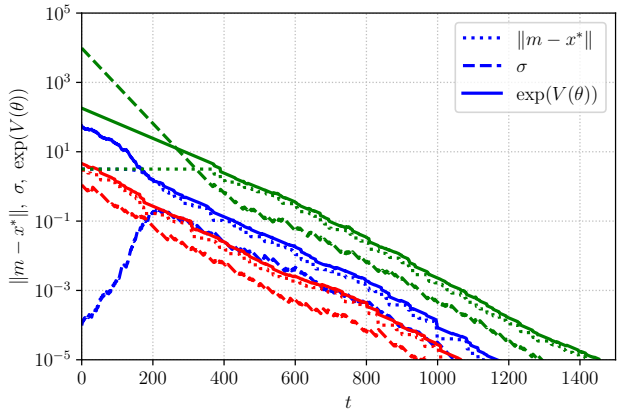


Fig. 3.1: Three runs of (1+1)-ES ($\alpha_\uparrow = e^{0.1}$ and $\alpha_\uparrow = e^{-0.025}$) on 10 dimensional spherical function $f(x) = \frac{1}{2}\|x - x^*\|^2$ with initial step-size $\sigma_0 = 10^{-4}$, 1, and 10^4 (in blue, red, green, respectively). Plotted are the distance to the optimum (dotted line), the step-size (dashed line), and the potential function $V(\theta)$ defined in (4.7) (solid line) with $v = 4/d$, $\ell = \alpha_\uparrow^{-10}$, and $u = \alpha_\downarrow^{-10}$.

To deal with the stochastic setting, we consider the expected hitting time while for the asymptotic convergence we consider almost sure results. Upper-bounds on the hitting time are typically harder to derive than lower-bounds. While upper-bounds inform us about the (linear) convergence, the lower-bound helps understanding whether the upper bound is tight.

Figure 3.1 visualizes three different runs of the (1+1)-ES on a function with spherical level sets with different initial step-sizes. First of all we clearly observe linear convergence. The first hitting time of $\mathcal{B}(x^*, \epsilon)$ scales linearly with $\log(1/\epsilon)$ for a sufficiently small $\epsilon > 0$. Second, its convergence speed is independent of the initial condition. Therefore, we expect to have universal constants C_R and $\tilde{C}_R$ independent of the initial state. Last, depending on the initial step-size, the transient time can differ. If the initial step-size is too large or too small, it does not produce progress in terms of $\|m_t - x^*\|$ until the step-size is well adapted. Therefore, C_T and $\tilde{C}_T$ depend on the initial condition, with a logarithmic dependency on the initial multiplicative mismatch.

---

satisfying (3.2) with a convergence rate $\exp(-C)$. Then for all C' with $0 < C' < C$, there exits τ such that for all $t > \tau$

$$\log \|m_t - x^*\| < -tC' + \log \|m_0 - x^*\|$$

and thus there exists $C_\tau > 0$ such that for all $t \geqslant 0$

$$\log \|m_t - x^*\| < -tC' + \log \|m_0 - x^*\| + C_\tau \ .$$

Then the first hitting time to $\mathcal{B}(x^*, \epsilon)$ denoted T_ϵ is upper bounded by T such that $-TC' + \log \|m_0 - x^*\| + C_\tau = \log \epsilon$ that is $T = (\log(1/\epsilon) + \log \|m_0 - x^*\| + C_\tau)/C'$ (this T might not be an integer) and thus

$$T_\epsilon \leqslant \frac{\log(\|x_0 - x^*\|/\epsilon) + C_\tau}{C'}.$$

Hence we have shown that the asymptotic linear convergence in (3.2) implies the upper bound on the hitting time given in (3.1).

3.2. Bounds of the Hitting Time via Drift Conditions. We are going to use *drift analysis* that consists in deducing properties on a sequence $\{X_t : t \geqslant 0\}$ (adapted to a natural filtration $\{\mathcal{F}_t : t \geqslant 0\}$) from its drift defined as $\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] - X_t$ [30]. Drift analysis has been widely used to analyze hitting times of evolutionary algorithms defined on discrete search spaces (mainly on binary search spaces). In this context, specific drift conditions have been derived. Depending on the form of the bound of the drift, different types of expected first hitting time bounds are available, such as additive drift [36, 37], variable drift [10, 51] and multiplicative drift [19, 20]. Though they were developed mainly for finite search spaces, the drift theorems can naturally be generalized to continuous domains [47, 49]. Indeed, Jägersküpper's work [39, 41, 42] are based on the same idea, while the link to the drift analysis is unveiled in the literature.

Since many drift conditions have been developed for analyzing algorithms on discrete domains, the domain of X_t is often implicitly assumed to be bounded. However, this assumption is violated in our situation, where we will use $X_t = \log\left(f_\mu(m_t)\right)$ as the quality measure, which takes values in $\mathbb{R} \cup \{-\infty\}$, and is meant to approach $-\infty$. To see the technical difficulty when the random process is unbounded, consider a process X_t defined as $X_{t+1} = X_t - 1/p$ with probability $p < 1$ and $X_{t+1} = X_t$ with probability $1 - p$. The drift $\mathbb{E}[X_{t+1}|\mathcal{F}_t] - X_t$ equals -1 no matter how small p is. However, it is easy to see that the expected first hitting time of this process can be arbitrarily large as p is chosen very small, while the expected single step progress is constant. This example reveals that bounding the drift is not sufficient to derive the expected first hitting time. This problem does not happen if the domain of X_t is bounded from below as the maximal decrease of X_t in one step is also bounded. In general, translating expected progress requires bounding the tail of the progress distribution, as formalized in [30].

3.2.1. Additive Drift for the Upper Runtime Bound. To overcome the above mentioned difficulty, we construct a stochastic process $\{Y_t : t \geqslant 0\}$ iteratively as follows: $Y_0 = X_0$ and

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

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

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

$$\mathbb{E}\left[\max\left\{X_{t+1} - X_t, -A\right\} \cdot 1_{\{T_\beta^X > t\}} \,\middle|\, \mathcal{F}_t\right] \leqslant -B \cdot 1_{\{T_\beta^X > t\}} \ , \tag{3.4}$$

then the expectation of T_β^X satisfies

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

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

We will prove that

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

We start from

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

and bound the different terms:

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

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

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

where we have used condition (3.4). Hence, by injecting (3.8) and (3.9) into (3.7), we obtain (3.6).

From (3.6), by taking the expectation we deduce

$$\mathbb{E}[\bar{Y}_{t+1}] \leqslant \mathbb{E}[\bar{Y}_t] - B \cdot \Pr[T^Y_\beta > t] \ . \tag{3.10}$$

Following the same approach as [49, Theorem 1], since T^Y_β is a random variable taking values in $\mathbb{N}$, it can be rewritten as $\mathbb{E}[T^Y_\beta] = \sum_{t=0}^{+\infty} \Pr[T^Y_\beta > t]$ and thus it holds

$$\begin{aligned} B \cdot \mathbb{E}\left[T^Y_\beta\right] \overset{\tilde{t}\to\infty}{\longleftarrow} \sum_{t=0}^{\tilde{t}} B \cdot \Pr\left[T^Y_\beta > t\right] &\leqslant \sum_{t=0}^{\tilde{t}} \left(\mathbb{E}[\bar{Y}_t] - \mathbb{E}[\bar{Y}_{t+1}]\right) \\ &\leqslant \mathbb{E}[\bar{Y}_0] - \mathbb{E}[\bar{Y}_{\tilde{t}}] = \beta_0 - \mathbb{E}[\bar{Y}_{\tilde{t}}] \ . \end{aligned} \tag{3.11}$$

Since $Y_{t+1} \geqslant Y_t - A$, then $Y_{T^Y_\beta} \geqslant \beta - A$ and given that $\bar{Y}_t \geqslant Y_{T^Y_\beta}$, we deduce that $E[\bar{Y}_{\tilde{t}}] \geqslant \beta - A$ for all $\tilde{t}$, which implies

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

With $\mathbb{E}[T^X_\beta] \leqslant \mathbb{E}[T^Y_\beta]$ this proves the upper bound. ∎

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

3.2.2. Additive Drift for the Lower Runtime Bound. Jägersküpper [41, Theorem 2] established a general lower bound of the expected first hitting time of the (1+1)-ES. We borrow the same idea to prove the following general theorem for a lower bound of the expected first hitting time, which generalizes [42, Lemma 12].

THEOREM 3.3. *Let $\{X_t : t \geqslant 0\}$ be a sequence of real-valued random variables adapted to a filtration $\{\mathcal{F}_t : t \geqslant 0\}$ and integrable such that*

$$X_0 = \beta_0, \quad X_{t+1} \leqslant X_t, \quad \text{and} \quad \mathbb{E}[X_{t+1} \mid \mathcal{F}_t] - X_t \geqslant -C$$

for $C > 0$. For $\beta < \beta_0$ we define $T_\beta^X = \min\{t : X_t \leqslant \beta\}$. Then the expected hitting time is lower bounded by

$$\mathbb{E}\left[T_\beta^X\right] \geqslant -(1/2) + (4C)^{-1} \cdot (\beta_0 - \beta) \ .$$

Proof of Theorem 3.3. After $T = \lfloor(\beta_0 - \beta)/(2C)\rfloor$ iterations it holds $\mathbb{E}[\beta_0 - X_T] \leqslant C \cdot T \leqslant (\beta_0 - \beta)/2$. From Markov's inequality we conclude $\Pr[\beta_0 - X_T \geqslant \beta_0 - \beta] \leqslant \frac{1}{2}$ and thus $\Pr[\beta_0 - X_T \leqslant \beta_0 - \beta] \geqslant \frac{1}{2}$, which is equivalent to $\Pr[T_\beta^X \geqslant T] \geqslant \frac{1}{2}$. Applying the Markov inequality once more we obtain

$$\mathbb{E}[T_\beta^X] = \sum_{t=1}^{\infty} \Pr\left[T_\beta^X \geqslant t\right] \geqslant \sum_{t=1}^{T} \Pr\left[T_\beta^X \geqslant t\right] \geqslant \Pr\left[T_\beta^X \geqslant T\right] \cdot T \geqslant \frac{T}{2} \geqslant \frac{\beta_0 - \beta}{4C} - \frac{1}{2} \ .$$

This completes the proof. ∎

4. Main Result: Expected First Hitting Time Bound. We derive in this section the expected first hitting time bounds of $\log\|m_t - x^*\|$ to $(-\infty, \log\epsilon)$ by using the methodology introduced in the previous section under the hypothesis that the objective functions satisfy Assumptions A1 and A2. For this purpose, after some preliminary results where we bound the probability of success in Section 4.2, we define in Section 4.3 a potential function $V(\theta_t)$ such that $\log f_\mu(m_t) \leqslant V(\theta_t)$ for any $\theta_t = (m_t, \sigma_t, \Sigma_t)$. In Section 4.4, we establish the drift on the potential $V(\theta)$ and derive an upper bound of the expected first hitting time of $X_t = V(\theta_t)$ to $(-\infty, \beta]$. This immediately provides the bounds for the expected first hitting time of $\log f_\mu(m_t)$ and $\log\|m_t - x^*\|$. Finally in Section 4.5 we derive a general lower bound of the expected first hitting time of $\log\|m_t - x^*\|$ for (1+1)-ES$_\kappa$ with an arbitrary step-size control mechanism.

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

4.2. Preliminaries. We present two preliminary results. In Assumption A1, we assume that for $m \in \mathcal{X}_a^b$, we can include into the sublevel set $S_0(m)$ a ball of radius $C_\ell f_\mu(m)$ and embed $S_0(m)$ into a ball of radius $C_u f_\mu(m)$. This allows us to upper bound and lower bound the probability of success for all $m \in \mathcal{X}_a^b$, for all $\Sigma \in \mathcal{S}_\kappa$ by the probability to sample into balls of radius $C_u f_\mu(m)$ and $C_\ell f_\mu(m)$ with appropriate center. From this we can upper-bound $p_{(a,b]}^{\text{upper}}(\bar\sigma)$ by a function of $\bar\sigma$. Similarly we can lower-bound $p_{(a,b]}^{\text{lower}}(\bar\sigma)$ by a function of $\bar\sigma$. Its proof is given in Appendix B.6.

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

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

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

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

We use the previous proposition to establish the next lemma that guarantees the existence of a finite range of normalized step-size that leads to the success probability into some range (p_u, p_ℓ) independent of m and Σ, and provides a lower bound on the success probability with rate r when the normalized step-size is in the above range. Its proof is provided in Appendix B.7.

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

$$\bar{\sigma}_\ell = \sup\left\{\bar{\sigma} > 0 : p^{\mathrm{lower}}_{(a,b]}(\bar{\sigma}) \geqslant p_\ell\right\} \quad , \tag{4.3}$$

$$\bar{\sigma}_u = \inf\left\{\bar{\sigma} > 0 : p^{\mathrm{upper}}_{(a,b]}(\bar{\sigma}) \leqslant p_u\right\} \quad , \tag{4.4}$$

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

$$p^*_r := \inf_{\ell \leqslant \bar{\sigma} \leqslant u} \inf_{m \in \mathcal{X}^b_a} \inf_{\Sigma \in \mathcal{S}_\kappa} p^{\mathrm{succ}}_r(\bar{\sigma}; m, \Sigma) \tag{4.5}$$

is lower bounded by

$$p^*_r \geqslant \min_{\ell \leqslant \bar{\sigma} \leqslant u} \kappa^{-d/2}\Phi\left(\mathcal{B}\left(\left(\frac{(2C_u - (1-r)C_\ell)\kappa^{1/2}}{\bar{\sigma}}\right)e_1, \frac{(1-r)C_\ell\kappa^{1/2}}{\bar{\sigma}}\right); 0, \mathrm{I}\right) > 0 \ . \tag{4.6}$$

4.3. Potential Function. Lemma 4.2 divides the domain of the normalized step-size into three disjoint subsets: $\bar{\sigma} \in (0, \ell)$ is a too small normalized step-size situation where we have $p^{\mathrm{succ}}_0(\bar{\sigma}; m, \Sigma) \geqslant p_\ell$ for all $m \in \mathcal{X}^b_a$ and $\Sigma \in \mathcal{S}_\kappa$; $\bar{\sigma} \in (u, \infty)$ is a too large normalized step-size situation where we have $p^{\mathrm{succ}}_0(\bar{\sigma}; m, \Sigma) \leqslant p_u$ for all $m \in \mathcal{X}^b_a$ and $\Sigma \in \mathcal{S}_\kappa$; and $\bar{\sigma} \in [\ell, u]$ is a reasonable normalized step-size situation where the success probability with rate r is lower bounded by (4.6). Since $p_{\mathrm{target}} \in [p_u, p_\ell]$, the normalized step-size is supposed to be maintained in the reasonable range.

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

$$V(\theta) = \log(f_\mu(m)) + \max\left\{0, v\log\left(\frac{\alpha_\uparrow \cdot \ell \cdot f_\mu(m)}{\sigma}\right), v\log\left(\frac{\sigma}{\alpha_\downarrow \cdot u \cdot f_\mu(m)}\right)\right\} \ . \tag{4.7}$$

The rationale behind the second term on the RHS is as follows. The second and third terms inside max are positive only if the normalized step-size $\bar{\sigma} = \sigma/f_\mu(m)$ is smaller than $\ell\alpha_\uparrow$ and greater than $u\alpha_\downarrow$, respectively. The potential value is $\log f_\mu(m)$

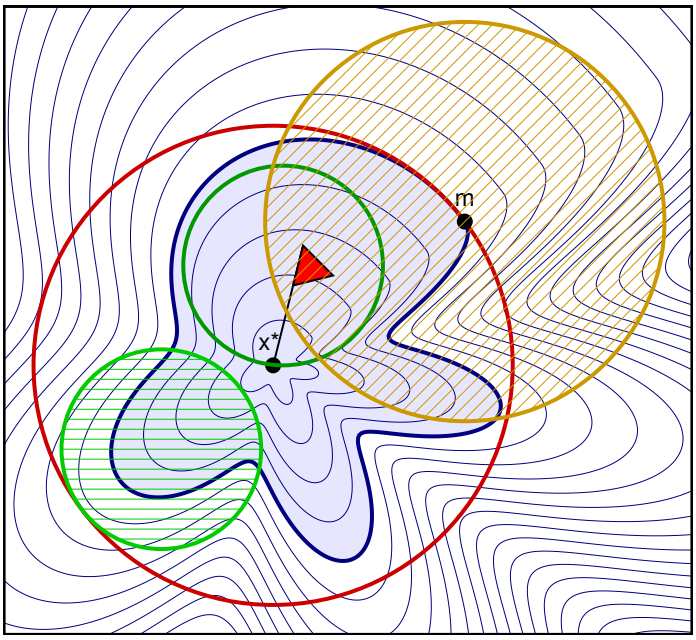


Fig. 4.1: Illustration of property A1. The sampling distribution is indicated by the mean m and the shaded orange circle, indicating one standard deviation. The blue set is the sub-levelset $S_0(m)$ of points improving upon m. It is enclosed in the red (outer) ball of radius $C_u f_\mu(m)$ and contains the dark green (inner) ball of radius $C_\ell f_\mu(m)$. The shaded light green ball indicates the worst case situation captured by the bound, namely that the small ball is positioned within the large ball at maximal distance to m.

if the normalized step-size is in $[\ell\alpha_\uparrow, u\alpha_\downarrow]$ and it is penalized if the normalized step-size is too small or too large. We need this penalization for the following reason. If the normalized step-size is too small, the success probability is close to $1/2$ for non-critical points, assuming $f = g \circ h$ where h is a continuously differentiable function but the progress in per step is very small because the step-size directly controls the progress for instance measured as $\|m_{t+1} - m_t\| = \sigma_t \|\mathcal{N}(0, \Sigma_t)\| 1_{\{f(m_{t+1}) \leqslant f(m_t)\}}$. If the normalized step-size is too large, the success probability is close to zero and produces no progress with high probability. If we would use $\log f_\mu(m)$ as a potential function instead of $V(\theta)$ then the progress is arbitrarily small in such situations, which prevents the application of drift arguments. The above potential function penalizes such situations, and guarantees a certain progress in the penalized quantity since the step-size will be increased or decreased, respectively, with high probability, leading to a certain decrease of $V(\theta)$. We illustrate in Figure 3.1 that $\log(f_\mu(m))$ cannot work alone as potential function while $V(\theta)$ does: when we start from a too small or too large step-size, $\log(f_\mu(m))$ looks constant (doted line in green and blue). Only when the step-size is started at 1, we see progress in $\log(f_\mu(m))$. Also, the step size can always get arbitrarily worse, with a very small probability, which forces us to handle the case of badly adapted step size properly. Yet the simulation of $V(\theta)$ shows that in all three situations (small, large and well adapted step-sizes compared to the distance

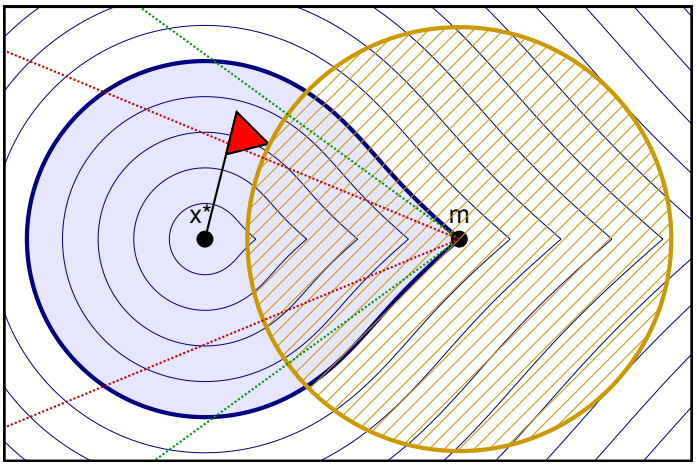


Fig. 4.2: The sampling distribution is indicated by the mean m and the shaded orange circle, indicating one standard deviation. The blue set is the sub-levelset $S_0(m)$ of points improving upon m. Although the level set has a kink at m, there exists a cone centered at m covering a probability mass of 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}.

to the optimum), we observe a geometric decrease of $V(\theta)$.

4.4. Upper Bound of the First Hitting Time. We are now ready to establish that the potential function defined in (4.7) satisfies a (truncated)-drift condition from Theorem 3.2. This will in turn imply an upper bound on the expected hitting time of $f_\mu(m)$ to $[0, \epsilon]$ provided $a \leqslant \epsilon$. The proof follows the same line of argumentation as the proof of [2, Proposition 4.2], which was restricted to the case of spherical functions. It was generalized under similar assumptions as in this paper, but for a fixed covariance matrix equal to the identity, in [52, Proposition 6]. The detailed proof is given in Appendix B.8.

PROPOSITION 4.3. *Consider the (1+1)-ES$_\kappa$ (Algorithm 2.1) with state $\theta_t = (m_t, \sigma_t, \Sigma_t)$. Assume that the minimized objective function f satisfies A1 and A2 for some $0 \leqslant a < b \leqslant \infty$. Let p_u and p_ℓ be constants satisfying $0 < p_u < p_{\text{target}} < p_\ell < p^{\text{limit}}$ and $p_\ell + p_u = 2p_{\text{target}}$. Then, there exists $\ell \leqslant \bar{\sigma}_\ell$ and $u \geqslant \bar{\sigma}_u$ such that $u/\ell \geqslant \alpha_\uparrow/\alpha_\downarrow$, where $\bar{\sigma}_\ell$ and $\bar{\sigma}_u$ are defined in Lemma 4.2. For any $A > 0$, taking v satisfying $0 < v < \min\left\{1, \frac{A}{\log(1/\alpha_\downarrow)}, \frac{A}{\log(\alpha_\uparrow)}\right\}$, and the potential function (4.7), we have*

$$\mathbb{E}\left[\max\{V(\theta_{t+1}) - V(\theta_t), -A\} \cdot 1_{\{m_t \in \mathcal{X}_a^b\}} \mid \mathcal{F}_t\right] \leqslant -B \cdot 1_{\{m_t \in \mathcal{X}_a^b\}} \tag{4.8}$$

where

$$B = \min\left\{A \cdot p_r^* - v \cdot \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right),\ v \cdot \frac{p_\ell - p_u}{2} \cdot \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right)\right\}\ , \tag{4.9}$$

and

$$p_r^* = \inf_{\bar{\sigma} \in [\ell, u]} \inf_{m \in \mathcal{X}_a^b} \inf_{\Sigma \in \mathcal{S}_\kappa} p_r^{\text{succ}}(\bar{\sigma}; m, \Sigma) \ \textit{with}\ r = 1 - \exp\left(-\frac{A}{1-v}\right)\ .$$

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

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

LEMMA 4.4. *Let $\{\theta_t : t \geqslant 0\}$ be the sequence of parameters $\theta_t = (m_t, \sigma_t, \Sigma_t)$ defined by the (1+1)-ES$_\kappa$ with the initial condition $\theta_0 = (m_0, \sigma_0, \Sigma_0)$ optimizing a measurable function f. Let $X_t = V(\theta_t)$ as defined in (4.7) and define the process Y_t as defined in Theorem 3.2. Then, for any $A > 0$, $\{Y_t : t \geqslant 0\}$ is integrable, i.e., $\mathbb{E}[|Y_t|] < \infty$ for each t.*

Proof of Lemma 4.4. The drift

$$Y_{t+1} = Y_t + \max\left\{V(\theta_{t+1}) - V(\theta_t), -A\right\} \cdot 1_{\{T^X_\beta > t\}} - B \cdot 1_{\{T^X_\beta \leqslant t\}}$$

is by construction bounded by $-A$ from below. It is also bounded by a constant from above. Indeed, from the proof of Proposition 4.3, it is easy to find the upper bound, say C, of the truncated one-step change, Δ_t in the proof of Proposition 4.3, without using A1 and A2. Let $D = \max\{A, C\}$. Then, by recursion, $|V(\theta_t)| \leqslant |V(\theta_0)| + |V(\theta_t) - V(\theta_0)| \leqslant |Y_0| + D \cdot t$. Hence $\mathbb{E}[|Y_t|] \leqslant |Y_0| + D \cdot t < \infty$ for all t. ∎

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

THEOREM 4.5. *Consider the same situation as described in Proposition 4.3. Let $T_\epsilon = \min\{t : f_\mu(m_t) \leqslant \epsilon\}$ be the first hitting time of $f_\mu(m_t)$ to $[0, \epsilon]$. Choose $a \leqslant \epsilon < f_\mu(m_t) \leqslant b$, where a and b appear in Definition 2.5. If $m_0 \in \mathcal{X}^b_a$, the first hitting time is upper bounded by*

$$\mathbb{E}[T_\epsilon] \leqslant \frac{V(\theta_0) - \log(\epsilon) + A}{B} \tag{4.10}$$

for $A > B > 0$ described in Proposition 4.3, where $V(\theta)$ is the potential function defined in (4.7). Equivalently, we have $\mathbb{E}[T_\epsilon] \leqslant C_T + C_R^{-1} \cdot \log(f_\mu(m_0)/\epsilon)$, where

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

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

Proof. Theorem 3.2 with Proposition 4.3 and Lemma 4.4 together bounds the expected first hitting time of $V(\theta_t)$ to $(-\infty, \log(\epsilon)]$ by the RHS of (4.10) with A and B as defined in Proposition 4.3. Since $\log f_\mu(m_t) \leqslant V(\theta_t)$, T_ϵ is bounded by the first hitting time of $V(\theta_t)$ to $(-\infty, \log(\epsilon)]$. The inequality is preserved if we take the expectation. The last claim is trivial from the inequality $\|x - x^*\| \leqslant 2C_u f_\mu(x)$, which holds under A1. ∎

Theorem 4.5 shows an upper bound on the expected hitting time of the (1+1)-ES$_\kappa$ with success-based step-size adaptation for linear convergence towards the global optimum x^* on functions satisfying A1 and A2 with $a = 0$. Moreover, with the choice if $b = \infty$, this bound holds from all initial search points m_0. We state this result in the following corollary.

COROLLARY 4.6 (Global Linear Convergence Upper Bound). *Consider the (1+1)-ES$_\kappa$ algorithm given in Algorithm 2.1. Assume that the minimized objective function f satisfies A1 and A2 for $a = 0$ and $b = \infty$. Then, for all ϵ, for all $m_0 \in \mathbb{R}^d$, the expected first hitting time $T_\epsilon = \min\{t : f_\mu(m_t) \leqslant \epsilon\}$ satisfies a bound for linear convergence*

$$\mathbb{E}[T_\epsilon] \leqslant C_T + C_R^{-1} \cdot \log(f_\mu(m_0)/\epsilon) \ ,$$

where C_T and C_R are given in Theorem 4.5. Moreover, $C_T + C_R^{-1} \cdot \log(f_\mu(m_0)/\epsilon)$ is an upper bound on the expected hitting time of $\|m_t - x^*\|$ to $[0, 2C_u\epsilon]$.

If $a > 0$, the bound in Theorem 4.5 does not translate into linear convergence, but we still obtain an upper bound on the expected first hitting time of the target accuracy $\epsilon \geqslant a$. This is useful for understanding the behavior of (1+1)-ES$_\kappa$ on multimodal functions, and on functions with degenerated Hessian matrix at the optimum.

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

LEMMA 4.7. *We consider the process $\{\theta_t : t \geqslant 0\}$ generated by a (1+1)-ES$_\kappa$ algorithm with an arbitrary step-size adaptation mechanism and an arbitrary covariance matrix update optimizing an arbitrary measurable function f. We assume $d \geqslant 2$ and $\kappa_t = \mathrm{Cond}(\Sigma_t) \leqslant \kappa$. We consider the natural filtration $\mathcal{F}_t$. Then, the expected single-step progress is lower-bounded by*

$$\mathbb{E}[\min(\log(\|m_{t+1} - x^*\|/\|m_t - x^*\|), 0) \mid \mathcal{F}_t] \geqslant -\kappa_t^{\frac{d}{2}}/d \ . \tag{4.11}$$

Proof of Lemma 4.7. Note first that

$$\log(\|m_{t+1} - x^*\|/\|m_t - x^*\|) = \log(\|x_t - x^*\|/\|m_t - x^*\|)1_{\{f(x_t) \leqslant f(m_t)\}} \ .$$

This value can be positive since $f(x_t) \leqslant f(m_t)$ does not imply $\|x_t - x^*\| \leqslant \|m_t - x^*\|$ in general. Clipping the positive part to zero, we obtain a lower bound, which is the RHS of the above equality times the indicator $1_{\{\|x_t - x^*\| \leqslant \|m_t - x^*\|\}}$. Since the quantity is non-positive, dropping the indicator $1_{\{f(x_t) \leqslant f(m_t)\}}$ only decrease the lower bound. Hence, we have

$$\begin{aligned}\min(\log(\|m_{t+1} - x^*\|/\|m_t - x^*\|), 0) \\ \geqslant \log(\|x_t - x^*\|/\|m_t - x^*\|)1_{\{\|x_t - x^*\| \leqslant \|m_t - x^*\|\}} \ .\end{aligned}$$

Then,

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

We rewrite the lower bound of the drift. The RHS of the above inequality is the integral of $\log(\|x - x^*\|/\|m_t - x^*\|)$ in the integral domain $\bar{\mathcal{B}}(x^*, \|m_t - x^*\|)$ under the probability measure $\Phi\left(\cdot; m_t, \sigma_t^2\Sigma_t\right)$. Performing a variable change (through rotation and scaling) so that $m_t - x^*$ becomes $e_1 = (1, 0, \cdots, 0)$ and letting $\tilde{\sigma}_t = \sigma_t/\|m_t - x^*\|$, we can further rewrite it as the integral of $\log(\|x\|)$ in $\bar{\mathcal{B}}(0, 1)$ under $\Phi\left(\cdot; e_1, \tilde{\sigma}_t^2\Sigma_t\right)$.

With $\kappa_t = \mathrm{Cond}(\Sigma_t)$, we have $\varphi\left(\cdot; e_1, \tilde{\sigma}_t^2 \Sigma_t\right) \leqslant \kappa_t^{\frac{d}{2}} \varphi\left(\cdot; e_1, \kappa_t \tilde{\sigma}_t^2 \mathrm{I}\right)$, see Lemma C.1. Altogether, we obtain the lower bound

$$\mathbb{E}[\log(\|x_t - x^*\|/\|m_t - x^*\|) 1_{\{\|x_t - x^*\| \leqslant \|m_t - x^*\|\}} \mid \mathcal{F}_t] \\ \geqslant \kappa_t^{\frac{d}{2}} \int_{\bar{\mathcal{B}}(0,1)} \log(\|x\|) \varphi\left(\cdot; e_1, \kappa_t \tilde{\sigma}_t^2 \mathrm{I}\right) \mathrm{d}x \ .$$

The RHS is equivalent to $-\kappa_t^{\frac{d}{2}}$ times the single step progress of the (1+1)-ES on the spherical function at $m_t = e_1$ and $\sigma = \sqrt{\kappa}\tilde{\sigma}_t$.

To further bound the RHS of the above inequality, we adopt the argumentation of Jägersküpper [41, Theorem 2]. The point is as follows. We rewrite $\delta = (x - e_1)/\gamma_t$, where $\gamma_t = \sqrt{\kappa_t}\tilde{\sigma}_t$, then δ_t is $\mathcal{N}(0, \mathrm{I})$ distributed. Now we use the inequality $\|x\| = \|e_1 + \gamma_t \cdot \delta\| \geqslant \|e_1 + \gamma^*(\delta) \cdot \delta\|$, where $\gamma^*(\delta) = \mathrm{argmin}_\gamma \|e_1 + \gamma \cdot \delta\|$, and we obtain a lower bound of the RHS of the above inequality as $\kappa_t^{\frac{d}{2}} \int \log(\|e_1 + \gamma^*(\delta) \cdot \delta\|) \varphi(\delta; 0, \mathrm{I}) \mathrm{d}\delta$.

The integrand of the RHS of the above lower bound amounts to $-\log(\sin(\theta)) \cdot 1_{\{\theta \leqslant \pi/2\}}$, where $\theta \in [0, \pi)$ is the angle between δ and e_1. This easily follows from a geometric interpretation of the optimal scenario (see [2]). Let $W_d = \int_0^{\pi/2} \sin^d(\theta) d\theta$ denote the Wallis integral. Then the density of θ is $(2W_{d-2})^{-1} |\sin(\theta)|^{d-2}$. The integral is written as

$$\frac{-1}{2W_{d-2}} \int_0^{\pi/2} \log\left(\sin(\theta)\right) \sin^{d-2}(\theta) \mathrm{d}\theta = -\left(2(d-1)^2 W_{d-2}\right)^{-1} \int_0^1 \frac{\log(r)}{\sqrt{1 - r^{2/(d-1)}}} \mathrm{d}r \ .$$

Here we applied the change of variables $\sin(\theta)^{d-1} = r$. When considering r as a random variable uniformly distributed on $[0, 1]$, then $\log(r)$ and $1/\sqrt{1 - r^{2/(d-1)}}$ are positively correlated [65, Chapter 1, eq. (2.1)]. Therefore, the integral on the RHS is lower bounded by the product of the integrals of the two terms, which reads

$$\int_0^1 \frac{\log(r)}{\sqrt{1 - r^{2/(d-1)}}} \mathrm{d}r \geqslant \int_0^1 \log(r)\, dr \int_0^1 \sqrt{\frac{1}{1 - r^{2/(d-1)}}} \mathrm{d}r \ ,$$

where the 1st and 2nd integral on the RHS are -1 and $(d-1)W_{d-2}$, respectively. Using $d \leqslant 2(d-1)$ for all $d \geqslant 2$ concludes the proof. ∎

The following theorem proves that the expected first hitting time of (1+1)-ES$_\kappa$ is $\Omega(\log(\|m_0 - x^*\|/\epsilon))$ for any measurable function f, implying that it can not converge faster than linearly. In case of $\kappa = 1$ the lower runtime bound becomes $\Omega(d \cdot (\log(\|m_0 - x^*\|/\epsilon)))$, meaning that the runtime scales linearly with respect to d. The proof is a direct application of Lemma 4.7 to Theorem 3.3.

THEOREM 4.8. *We consider the process $\{\theta_t : t \geqslant 0\}$ generated by a (1+1)-ES$_\kappa$ described in Algorithm 2.1 and assume that f is a measurable function with $d \geqslant 2$. Let $T_\epsilon = \inf\{t : \|m_t - x^*\| \leqslant \epsilon\}$ be the first hitting time of $[0, \epsilon]$ by $\|m_t - x^*\|$. Then, the expected first hitting time is lower bounded by*

$$\mathbb{E}[T_\epsilon] \geqslant -(1/2) + \frac{d}{4\kappa^{d/2}} \log(\|m_0 - x^*\|/\epsilon) \ .$$

The bound holds for arbitrary step-size adaptation mechanisms. If A1 holds, it gives a lower bound for the expected first hitting time bound of $f_\mu(m_t)$ to $[0, 2C_\ell \epsilon]$.

Proof of Theorem 4.8. Let $X_t = \log\|m_t - x^*\|$ for $t \geqslant 0$. Define Y_t iteratively as $Y_0 = X_0$ and $Y_{t+1} = Y_t + \min(X_{t+1} - X_t, 0)$. Then, it is easy to see that $Y_t \leqslant X_t$ and $Y_{t+1} \leqslant Y_t$ for all $t \geqslant 0$. Note that $\mathbb{E}[Y_{t+1} - Y_t \mid \mathcal{F}_t] = \mathbb{E}[\min(X_{t+1} - X_t, 0) \mid \mathcal{F}_t] = \mathbb{E}[\min(\log(\|m_{t+1} - x^*\|/\|m_t - x^*\|), 0) \mid \mathcal{F}_t]$, where the RMS is lower bounded in light of Lemma 4.7. Then, applying Theorem 3.3, we obtain the lower bound. The last statement directly follows from $\|x - x^*\| \leqslant 2C_\ell f_\mu(x)$ under A1. ∎

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

PROPOSITION 4.9. *Consider the same situation as described in Proposition 4.3, where $a = 0$ and $0 < b \leqslant \infty$. Then, for any $m_0 \in \mathcal{X}_0^b$, $\sigma_0 > 0$ and $\Sigma \in \mathcal{S}_\kappa$, we have*

$$\Pr\left[\limsup_{t\to\infty} \frac{1}{t} \log f_\mu(m_t) \leqslant -B\right] = 1 \ , \tag{4.12}$$

and

$$\Pr\left[\limsup_{t\to\infty} \frac{1}{t} \log \|m_t - x^*\| \leqslant -B\right] = 1 \ , \tag{4.13}$$

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

Proof of Proposition 4.9. Let V be defined in (4.7). Let $Y_0 = V(\theta_0)$ and $Y_{t+1} = Y_t + \max(-A, V(\theta_{t+1}) - V(\theta_t))$. Define $Z_t = Y_t - \mathbb{E}_{t-1}[Y_t]$ for $t \geqslant 0$. Then, $\{Z_t\}$ is a martingale difference sequence on the filtration $\{\mathcal{F}_t\}$ produced by $\{\theta_t\}$. Then, we have $\frac{1}{t}\log f_\mu(m_t) \leqslant \frac{1}{t}V(\theta_t) \leqslant \frac{1}{t}Y_t$, and we have from Proposition 4.3

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

By repeatedly applying the above inequality and dividing it by t, we obtain $\frac{1}{t}Y_t \leqslant -B + \frac{1}{t}Y_0 + \frac{1}{t}\sum_{i=1}^t Z_i$, where $\lim_{t\to\infty}\frac{1}{t}Y_0 = 0$ and $\sum_{i=1}^t Z_i$ is a martingale sequence. In light of the strong law of large numbers for martingales [15], if $\sum_{t=1}^\infty \mathbb{E}[Z_t^2]/t^2 < \infty$, we have $\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^t Z_i = 0$ almost surely. By the definition of $V(\theta_t)$, we have $V(\theta_i) - V(\theta_{i-1}) \leqslant v\log(\alpha_\uparrow/\alpha_\downarrow)$. Hence,

$$\mathbb{E}[Z_i^2] = \mathbb{E}[(Y_i - \mathbb{E}_{i-1}[Y_i])^2] = \mathbb{E}[\max(-A, V(\theta_i) - V(\theta_{i-1}))^2] \leqslant \max(A, v\log(\alpha_\uparrow/\alpha_\downarrow))^2 \ .$$

Therefore, we have

$$\limsup_{t\to\infty} \frac{1}{t}\log f_\mu(m_t) \leqslant -B + \lim_{t\to\infty}\frac{1}{t}Y_0 + \lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^t Z_i = -B \quad a.s.,$$

which proves (4.12). The proof of (4.13) is then immediate using that $\|x - x^*\| \leqslant 2C_u f_\mu(x)$. ∎

4.7. Wrap-up of the Results: Global Linear Convergence. As a corollary to the lower-bound from Theorem 4.8, the upper bound from Theorem 4.5, Proposition 4.9 stating the almost sure linear convergence and the fact that different assumptions discussed in Section 2.3 imply A1 and A2, we summarize our linear convergence results in the following theorem.

THEOREM 4.10 (Global Linear Convergence). *We consider the (1+1)-ES$_\kappa$ optimizing an objective function f.*

(a) If one of the two following conditions holds:
- *f satisfies A1 and A2 for $a = 0$ and $p^{\text{limit}} > p^{\text{target}}$,*
- *f is a composite of a quadratically-bounded continuously differentiable function with a strictly increasing function (i.e. satisfying A6) and $p^{\text{target}} < 1/2$,*

then, for all m_0 in $\mathcal{X}_0^b$, $\sigma_0 > 0$, and $\Sigma_0 \in \mathcal{S}_\kappa$, for all ϵ, the expected hitting time $\mathbb{E}[T_\epsilon]$ of $\|m_t - x^*\|$ to $[0, \epsilon]$ is $\Theta\big(\log(\|m_0 - x^*\|/\epsilon)\big)$.

(b) Moreover, if $p^{\text{target}} < 1/2$ and one of the following assumption holds
- *f is the composite of a convex-quadratic function with a strictly increasing function (i.e. satisfying A3),*
- *f is the composite of a strongly-convex and smooth function on $\mathbb{R}^d$ with a strictly increasing function (i.e. satisfying A4),*
- *f is the composite of a positively homogeneous continuously differentiable function (with unique optimum) with a strictly increasing function (i.e. satisfying A7),*

then, for all m_0 in $\mathbb{R}^d$, $\sigma_0 > 0$, and $\Sigma_0 \in \mathcal{S}_\kappa$, for all ϵ, the expected hitting time $\mathbb{E}[T_\epsilon]$ of $\|m_t - x^*\|$ to $[0, \epsilon]$ is $\Theta\big(\log(\|m_0 - x^*\|/\epsilon)\big)$.

In the case (a) for all m_0 in $\mathcal{X}_0^b$, $\sigma_0 > 0$, and $\Sigma_0 \in \mathcal{S}_\kappa$ and in the case (b) for all m_0 in $\mathbb{R}^d$, $\sigma_0 > 0$, both $f_\mu(m_t)$ and $\|m_t - x^*\|$ converge almost surely, i.e.

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

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

The linear convergence of the $(1+1)$-ES (with $\kappa = 1$) on positively homogeneous functions satisfying A7 has been previously analyzed [7, 8]. The result is asymptotic and holds almost surely. It does not imply directly a bound on the expected hitting time to reach an ϵ-ball of the optimum. The present results complements that analysis.

4.8. Tightness in the Sphere Function Case. Now we consider a specific convex quadratic function, namely the sphere function $f(x) = \frac{1}{2}\|x\|^2$ where the spatial suboptimality function equals $f_\mu(x) = V_d\|x\|$. We have formulated in Theorem 4.10 that the expected hitting time of a ball of radius ϵ for the (1+1)-ES$_\kappa$ equals $\Theta(\log \|m_0 - x^*\|/\epsilon)$. Yet this statement does not give information on how the constants hidden in the Θ-notation scale with the dimension. In particular the convergence rate of the algorithm is upper-bounded by $\exp(-B)$ where B is given in Equation (4.9), see Theorem 4.5.

In this section, we estimate precisely the scaling of B with respect to the dimension and compare it with the general lower bound of the expected first hitting time given in Theorem 4.8. We then conclude that the bound is tight with respect to the scaling with d in the case of the sphere function.

Let us assume $\kappa = 1$. That is, we consider the (1+1)-ES without covariance matrix adaptation ($\Sigma = I$). Then, $p^{\text{lower}}_{(a,b]}(\bar\sigma) = p^{\text{upper}}_{(a,b]}(\bar\sigma) = p^{\text{succ}}_r(\bar\sigma; m, \Sigma) = p_r(\bar\sigma)$ with

$$p_r(\bar\sigma) = \Pr_{z\sim\mathcal{N}(0,I)}\left(\|e_1 + V_d\bar\sigma z\| < (1-r)\right) = \Phi\left(\mathcal{B}\left(\frac{e_1}{V_d\bar\sigma}, \frac{1-r}{V_d\bar\sigma}\right)\right) \ , \tag{4.14}$$

where $e_1 = (1, 0, \ldots, 0)$. This means that the success probability is solely controlled by the normalized step-size $\bar{\sigma}$.

We first investigate the properties of the success probability, proved in Appendix B.9.

LEMMA 4.11. *For all $d \in \mathbb{N}$ and $r \in [0, 1)$, p_r is positive and continuous. For $r = 0$ it is strictly decreasing and thus bijective. For all $d \in \mathbb{N}$, the image of $\bar{\sigma} \in \mathbb{R}_> \mapsto p_0(\bar{\sigma})$ is $(0, 1/2)$.*

We now investigate the asymptotic limit of the success probability for d to infinity, which is proved in Appendix B.10.

LEMMA 4.12. *For $r = r(d)$ fulfilling $\lim_{d\to\infty} d \cdot r(d) = \rho$, the limit $p^{\lim}_\rho(\hat{\sigma}) := \lim_{d\to\infty} p_r\left(\frac{\hat{\sigma}}{dV_d}\right)$ exists, and it equals $p^{\lim}_\rho(\hat{\sigma}) = \Psi\left(-\frac{\rho}{\hat{\sigma}} - \frac{\hat{\sigma}}{2}\right)$, where Ψ is the cumulative density function of the standard normal distribution. For $\rho = 0$, the function $p^{\lim}_0$ is continuous and strictly decreasing and the image of $p^{\lim}_\rho$ is $(0, 1/2)$.*

The following proposition states that the convergence speed is $\Omega(1/d)$, hence the expected first hitting time scales as $O(1/d)$. The proof is provided in Appendix B.11.

PROPOSITION 4.13. *For $A = 1/d$, $p_{\text{target}} \in \Theta(1)$ and $\log(\alpha_\uparrow/\alpha_\downarrow) \in \omega(1/d)$, we have $B \in \Omega(1/d)$.*

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

Proposition 4.13 and Theorem 4.5 imply $\mathbb{E}[T_\epsilon] \in O(d \cdot \log(\|m_0\|/\epsilon))$ and Theorem 4.8 implies $\mathbb{E}[T_\epsilon] \in \Omega(d \cdot \log(\|m_0\|/\epsilon))$. They yield $T_\epsilon \in \Theta(d \cdot \log(\|m_0\|/\epsilon))$. This result shows i) that the runtime of the (1+1)-ES on the sphere function is proportional to d as long as $\log(\alpha_\uparrow/\alpha_\downarrow) \in \omega(1/d)$ and ii) that from our methodology one can derive a tight bound of the runtime in some cases. The result is formally stated as follows.

THEOREM 4.14. *The (1+1)-ES described in Algorithm 2.1 with $\kappa = 1$ and $p^{\text{target}} < 1/2$ converges globally and linearly in terms of $\log\|m_t - x^*\|$ from any starting point $m_0 \in \mathbb{R}^d$, $\sigma_0 > 0$, and $\Sigma_0 = I$ on any function $f(x) = g(\|x - x^*\|)$, where g is a strictly increasing function. Moreover, if $p \in \Theta(1)$ and $\log(\alpha_\uparrow/\alpha_\downarrow) \in \omega(1/d)$, the expected first hitting time T_ϵ of $\log\|m_t - x^*\|$ to $(-\infty, \log(\epsilon)]$ is $\Theta(d \cdot \log(\|m_0\|/\epsilon))$ and the almost sure convergence rate is upper-bounded by $\exp(-\Theta(1/d))$.*

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

5. Discussion. We have established the global linear convergence of the (1+1)-ES$_\kappa$ almost surely and also expressed as a bound on the expected hitting time of an ϵ-neighborhood of the solution. Assumption A1 has been the key to obtaining the expected first hitting time bound of (1+1)-ES$_\kappa$ in the form of Equation (3.1). The

convergence results hold on a wide class of functions. It includes

(i) strongly convex functions with Lipschitz gradient where linear convergence of numerical optimization algorithm is usually analyzed,
(ii) continuously differentiable positively homogenous functions where previous linear convergence results had been introduced, and
(iii) functions with non-smooth level sets as illustrated in Figure 4.2.

Because the analyzed algorithms are invariant to strictly monotonic transformations of the objective functions, *all results that hold on f also hold on $g \circ f$ where $g : \mathrm{Im}(f) \to \mathbb{R}$ is a strictly increasing transformation that can thus introduce discontinuities on the objective function.*

In contrast to the previous result establishing the convergence of CMA-ES [18] by adding a step to enforce a sufficient decrease (which works well for direct search methods, but which is unnatural for ESs), we did not need to modify the adaptation mechanism of the (1+1)-ES to achieve our convergence proofs. We believe that this is crucial, since it allows our analysis to reflect the main mechanism that make the algorithm work well in practice.

Theorem 4.14 proves that we can derive a tight convergence rate with Proposition 4.3 on the sphere function in the case where $\kappa = 1$, i.e., without covariance matrix adaptation. This partially supports the utility of our methodology. However, the derivation of the tight upper bound (Lemmas 4.11 and 4.12 and Proposition 4.13) relies on the fact that both the level sets of the objective function and the equal-density curves of the sampling distribution are isotropic, and hence does not generalize immediately. Moreover, the lower bound (Theorem 4.8) seems to be loose even for $\kappa = 1$ on convex quadratic functions, where we empirically observe that the logarithmic convergence rate scales like $\Theta(1/\operatorname{Cond}(\nabla\nabla f))$, see Figure 2.1a, while its dependency on the dimension is tight. A better lower bound of the expected first hitting time and a handy way to estimate the convergence rate are relevant directions of future work.

Our main result in Theorem 4.10 states that (1+1)-ES$_\kappa$ converges linearly. That is, the convergence rate, $\exp(-B)$, is guaranteed to be strictly smaller than one under A1 and A2. This puts evolution strategies into the same category as many other direct search methods like direction search [44], the method of Nesterov and Spokoiny [55], random pursuit [62], and even many gradient-based methods. Furthermore, evolution strategies enjoy linear scaling with the search space dimension, which is optimal for comparison-based methods [64]. Moreover, our convergence result is not only asymptotic in nature. It provides a non-asymptotic bound of the expected first hitting time.

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

On convex quadratic functions, we empirically observe that the covariance matrix approaches a stable distribution that is closely concentrated around the inverse Hessian up to a scalar factor, and the convergence speed on all convex quadratic functions is equal to that on the sphere function (see Figure 2.1b). This behavior is not described by our result. Indeed in this paper, we did not assume any specific mechanism for the covariance matrix adaptation, and the results therefore hold for the worst situation, where the covariance matrix is wrongly adapted within $\mathcal{S}_\kappa$.

Covariance matrix adaptation is also important for optimizing functions with non-smooth level sets. On continuously differentiable functions, we can always set $\alpha_\uparrow$ and

$\alpha_\downarrow$ so that $p = \frac{\log(1/\alpha_\downarrow)}{\log(\alpha_\uparrow/\alpha_\downarrow)} < p^{\text{limit}} = 1/2$. This is the rationale behind the 1/5 success rule, where $p = 1/5$. Indeed, $p = 1/5$ is known to approximate the optimal situation on the sphere function where the expected one-step progress is maximized [58]. Therefore, one does not need to tune these parameters in a problem-specific manner. However, if the objective is not continuously differentiable and levelsets are non-smooth, then p^{limit} is in general smaller than 1/2. For example, it can be as low as $p^{\text{limit}} = 1/2^d$ on $f(x) = \|x\|_\infty = \max_{i=1,\ldots,n}|x_i|$. Without an appropriate adaptation of the covariance matrix the success probability will be smaller than $p = 1/5$ and one must tune $\alpha_\uparrow$ and $\alpha_\downarrow$ in order to converge to the optimum, which requires information about p^{limit}. By adapting the covariance matrix appropriately, the success probability can be increased arbitrary close to 1/2 (by elongating the variance in the direction of success domain) and $\alpha_\uparrow$ and $\alpha_\downarrow$ do not require tuning.

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

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

REFERENCES

[1] Y. Akimoto, *Analysis of a natural gradient algorithm on monotonic convex-quadratic-composite functions*, in Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation, GECCO '12, 2012, pp. 1293–1300.

[2] Y. Akimoto, A. Auger, and T. Glasmachers, *Drift theory in continuous search spaces: expected hitting time of the (1+ 1)-es with 1/5 success rule*, in Proceedings of the Genetic and Evolutionary Computation Conference, 2018, pp. 801–808.

[3] Y. Akimoto, A. Auger, and N. Hansen, *An ode method to prove the geometric convergence of adaptive stochastic algorithms*, 2018, https://arxiv.org/abs/1811.06703.

[4] Y. Akimoto, Y. Nagata, I. Ono, and S. Kobayashi, *Bidirectional relation between cma evolution strategies and natural evolution strategies*, in International Conference on Parallel Problem Solving from Nature, Springer, 2010, pp. 154–163.

[5] S. Alvernaz and J. Togelius, *Autoencoder-augmented neuroevolution for visual doom playing*, in 2017 IEEE Conference on Computational Intelligence and Games (CIG), 2017, pp. 1–8.

[6] D. V. Arnold and N. Hansen, *Active covariance matrix adaptation for the (1+1)-cma-es*, in Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation, GECCO '10, New York, NY, USA, 2010, ACM, pp. 385–392, https://doi.org/10.1145/1830483.1830556, http://doi.acm.org/10.1145/1830483.1830556.

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

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

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

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

[11] P. Bontrager, A. Roy, J. Togelius, N. Memon, and A. Ross, *Deepmasterprints: Generating masterprints for dictionary attacks via latent variable evolution**, in 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2018, pp. 1–9.

[12] S. Bubeck, *Convex optimization: Algorithms and complexity*, 2014, https://arxiv.org/abs/1405.4980.
[13] C. Cartis and K. Scheinberg, *Global convergence rate analysis of unconstrained optimization methods based on probabilistic models*, Mathematical Programming, 169 (2018), pp. 337–375.
[14] A. Chotard and A. Auger, *Verifiable Conditions for the Irreducibility and Aperiodicity of Markov Chains by Analyzing Underlying Deterministic Models*, Bernoulli, 25 (2018), pp. 112–147, https://doi.org/10.3150/17-BEJ970, https://hal.inria.fr/hal-01222222.
[15] Y. S. Chow, *On a strong law of large numbers for martingales*, Ann. Math. Statist., 38 (1967), p. 610, https://doi.org/10.1214/aoms/1177698981, https://doi.org/10.1214/aoms/1177698981.
[16] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*, SIAM, Philadelphia, PA, USA, 2009.
[17] L. Devroye, *The compound random search*, in International Symposium on Systems Engineering and Analysis, Purdue University, 1972, pp. 195–110.
[18] Y. Diouane, S. Gratton, and L. N. Vicente, *Globally convergent evolution strategies*, Mathematical Programming, 152 (2015), pp. 467–490.
[19] B. Doerr and L. A. Goldberg, *Adaptive drift analysis*, Algorithmica, 65 (2013), pp. 224–250.
[20] B. Doerr, D. Johannsen, and C. Winzen, *Multiplicative drift analysis*, Algorithmica, 64 (2012), pp. 673–697.
[21] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, *Efficient decision-based black-box adversarial attacks on face recognition*, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
[22] G. Fujii, M. Takahashi, and Y. Akimoto, *Cma-es-based structural topology optimization using a level set boundary expression—application to optical and carpet cloaks*, Computer Methods in Applied Mechanics and Engineering, 332 (2018), pp. 624 – 643.
[23] T. Geijtenbeek, M. Van De Panne, and A. F. Van Der Stappen, *Flexible muscle-based locomotion for bipedal creatures*, ACM Transactions on Graphics (TOG), 32 (2013), pp. 1–11.
[24] T. Glasmachers, *Global convergence of the (1 + 1) evolution strategy to a critical point*, Evolutionary Computation, https://doi.org/10.1162/evco_a_00248. To appear.
[25] D. Golovin, J. Karro, G. Kochanski, C. Lee, X. Song, and Q. Zhang, *Gradientless descent: High-dimensional zeroth-order optimization*, in International Conference on Learning Representations, 2020, https://openreview.net/forum?id=Skep6TVYDB.
[26] S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang, *Direct search based on probabilistic descent*, SIAM Journal on Optimization, 25 (2015), pp. 1515–1541.
[27] S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang, *Complexity and global rates of trust-region methods based on probabilistic models*, IMA Journal of Numerical Analysis, 38 (2017), pp. 1579–1597.
[28] D. Ha and J. Schmidhuber, *Recurrent world models facilitate policy evolution*, in Advances in Neural Information Processing Systems, 2018, pp. 2450–2462.
[29] D. Ha and J. Schmidhuber, *Recurrent world models facilitate policy evolution*, in Advances in Neural Information Processing Systems 31, 2018, pp. 2450–2462.
[30] B. Hajek, *Hitting-time and occupation-time bounds implied by drift analysis with applications*, Advances in Applied probability, 14 (1982), pp. 502–525.
[31] N. Hansen, *The CMA Evolution Strategy: A Comparing Review*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 75–102.
[32] N. Hansen, D. V. Arnold, and A. Auger, *Evolution strategies*, in Springer handbook of computational intelligence, Springer, 2015, pp. 871–898.
[33] N. Hansen and A. Auger, *Principled design of continuous stochastic search: From theory to practice*, in Theory and principled methods for the design of metaheuristics, Springer, 2014, pp. 145–180.
[34] N. Hansen, A. Auger, R. Ros, S. Finck, and P. Pošík, *Comparing results of 31 algorithms from the black-box optimization benchmarking bbob-2009*, in Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation, GECCO '10, 2010, pp. 1689–1696.
[35] N. Hansen and A. Ostermeier, *Completely derandomized self-adaptation in evolution strategies*, Evolutionary Computation, 9 (2001), pp. 159–195.
[36] J. He and X. Yao, *Drift analysis and average time complexity of evolutionary algorithms*, Artificial intelligence, 127 (2001), pp. 57–85.
[37] J. He and X. Yao, *A study of drift analysis for estimating computation time of evolutionary algorithms*, Natural Computing, 3 (2004), pp. 21–35.

[38] C. Igel, N. Hansen, and S. Roth, *Covariance matrix adaptation for multi-objective optimization*, Evolutionary Computation, 15 (2007), pp. 1–28.
[39] J. Jägersküpper, *Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces*, Automata, Languages and Programming, (2003), pp. 188–188.
[40] J. Jägersküpper, *Rigorous runtime analysis of the (1+ 1) es: 1/5-rule and ellipsoidal fitness landscapes*, in International Workshop on Foundations of Genetic Algorithms, Springer, 2005, pp. 260–281.
[41] J. Jägersküpper, *How the (1+1)-ES using isotropic mutations minimizes positive definite quadratic forms*, Theoretical Computer Science, 361 (2006), pp. 38–56.
[42] J. Jägersküpper, *Algorithmic analysis of a basic evolutionary algorithm for continuous optimization*, Theoretical Computer Science, 379 (2007), pp. 329–347.
[43] S. Kern, S. D. Müller, N. Hansen, D. Büche, J. Ocenasek, and P. Koumoutsakos, *Learning probability distributions in continuous evolutionary algorithms–a comparative review*, Natural Computing, 3 (2004), pp. 77–112.
[44] J. Konečnỳ and P. Richtárik, *Simple complexity analysis of simplified direct search*, Tech. Report 1410.0390, 2014.
[45] I. Kriest, V. Sauerland, S. Khatiwala, A. Srivastav, and A. Oschlies, *Calibrating a global three-dimensional biogeochemical ocean model (mops-1.0)*, Geoscientific Model Development, 10 (2017), p. 127.
[46] J. Larson, M. Menickelly, and S. M. Wild, *Derivative-free optimization methods*, Acta Numerica, 28 (2019), pp. 287–404.
[47] P. K. Lehre and C. Witt, *General drift analysis with tail bounds*, Tech. Report arXiv:1307.2559, 2013.
[48] J. Lengler, *Drift analysis*, in Theory of Evolutionary Computation, Springer, 2020, pp. 89–131.
[49] J. Lengler and A. Steger, *Drift analysis and evolutionary algorithms revisited*, Tech. Report arXiv:1608.03226, 2016.
[50] P. MacAlpine, S. Barrett, D. Urieli, V. Vu, and P. Stone, *Design and optimization of an omnidirectional humanoid walk: A winning approach at the RoboCup 2011 3D simulation competition*, in Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI), July 2012.
[51] B. Mitavskiy, J. Rowe, and C. Cannings, *Theoretical analysis of local search strategies to optimize network communication subject to preserving the total number of links*, International Journal of Intelligent Computing and Cybernetics, 2 (2009), pp. 243–284.
[52] D. Morinaga and Y. Akimoto, *Generalized drift analysis in continuous domain: linear convergence of (1+ 1)-es on strongly convex functions with lipschitz continuous gradients*, in Proceedings of the 15th ACM/SIGEVO Conference on Foundations of Genetic Algorithms, 2019, pp. 13–24.
[53] A. Nemirovski, *Information-based complexity of convex programming*, Lecture Notes, (1995).
[54] Y. Nesterov, *Lectures on convex optimization*, vol. 137, Springer, 2018.
[55] Y. Nesterov and V. Spokoiny, *Random gradient-free minimization of convex functions*, Foundations of Computational Mathematics, 17 (2017), pp. 527–566.
[56] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen, *Information-geometric optimization algorithms: A unifying picture via invariance principles*, The Journal of Machine Learning Research, 18 (2017), pp. 564–628.
[57] C. Paquette and K. Scheinberg, *A stochastic line search method with convergence rate analysis*, Tech. Report arXiv:1807.07994, 2018.
[58] I. Rechenberg, *Evolutionsstrategie: Optimierung technisher Systeme nach Prinzipien der biologischen Evolution*, Frommann-Holzboog, 1973.
[59] I. Rechenberg, *Evolutionsstrategie'94*, frommann-holzboog, 1994.
[60] L. M. Rios and N. V. Sahinidis, *Derivative-free optimization: a review of algorithms and comparison of software implementations*, Journal of Global Optimization, 56 (2013), pp. 1247–1293.
[61] M. Schumer and K. Steiglitz, *Adaptive step size random search*, Automatic Control, IEEE Transactions on, 13 (1968), pp. 270–276.
[62] S. U. Stich, C. L. Muller, and B. Gartner, *Optimization of convex functions with random pursuit*, SIAM Journal on Optimization, 23 (2013), pp. 1284–1309.
[63] S. U. Stich, C. L. Müller, and B. Gärtner, *Variable metric random pursuit*, Mathematical Programming, 156 (2016), pp. 549–579.
[64] O. Teytaud and S. Gelly, *General lower bounds for evolutionary algorithms*, in Parallel Problem Solving from Nature-PPSN IX, Springer, 2006, pp. 21–31.
[65] H. Thorisson, *Coupling, Stationarity, and Regeneration*, Probability and Its Applications,

Springer New York, 2000.
[66] J. Uhlendorf, A. Miermont, T. Delaveau, G. Charvin, F. Fages, S. Bottani, G. Batt, and P. Hersen, *Long-term model predictive control of gene expression at the population and single-cell levels*, Proceedings of the National Academy of Sciences, 109 (2012), pp. 14271–14276.
[67] V. Volz, J. Schrum, J. Liu, S. M. Lucas, A. Smith, and S. Risi, *Evolving Mario levels in the latent space of a deep convolutional generative adversarial network*, in Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '18, 2018, pp. 221–228.
[68] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber, *Natural evolution strategies*, The Journal of Machine Learning Research, 15 (2014), pp. 949–980.

Appendix A. Some Numerical Results.

We present experiments with five algorithms on two convex quadratic functions. We compare (1+1)-ES, (1+1)-CMA-ES, simplified direction search [44], random pursuit [62], and gradientless descent [25].

All algorithms were started at the initial search point $x_0 = \frac{1}{\sqrt{d}}(1, \ldots, 1) \in \mathbb{R}^d$. We implemented the algorithms as follows, with their parameters tuned where necessary:

- The ES always uses the setting $\alpha_\uparrow = \exp(4/d)$ and $\alpha_\downarrow = \alpha_\uparrow^{-1/4}$ for step size adaptation.
- We set the constant c in the sufficient decrease condition of Simplified Direction Search to $\frac{1}{10}$, and we employed the standard basis as well as the negatives of these vectors as candidate directions. In each iteration we looped over the set of directions in random order. Randomizing the order greatly boosted performance over a fixed order.
- Random Pursuit was implemented with a golden section line search in the range $[-2\sigma, 2\sigma]$ with a rather loose target precision of $\sigma/2$, where σ is either the initial step size or the length of the previous step.
- For Gradientless Descent we used the initial step size as the maximal step size and defined a target precision of 10^{-10}. This target is reached by the ES in all cases.

The experiments are designed to demonstrate several different effects:

- We perform all experiments in $d = 10$ and $d = 50$ dimensions to investigate dimension-dependent effects.
- We investigate best-case performance by running the algorithms on the spherical function $\|x\|^2$, i.e., on the separable convex quadratic function with minimal condition number. The initial step size is set to $\sigma_0 = 1$. All algorithms have a budget of $100 \cdot d$ function evaluations.
- We investigate the dependency of the performance on initial parameter settings by repeating the same experiment as above, but with an initial step size of $\sigma_0 = \frac{1}{1000}$. All algorithms have a budget of $700 \cdot d$ function evaluations.
- We investigate the dependence on problem difficulty by running the algorithms on an ellipsoid problem with a moderate condition number of $\kappa_f = 100$. The eigenvalues of the Hessian are evenly distributed on a log-scale. We use $\sigma_0 = 1$ like in the first experiment. All algorithms have a budget of $500 \cdot d$ function evaluations.

The experimental results are presented in Figure A.1. We observe only moderate dimension-dependent effects, besides the expected linear increase of the runtime. We see robust performance of the ES, in particular with covariance matrix adaptation. The second experiment demonstrates the practical importance of the ability to grow the step size: the ES is essentially unaffected by wrong initial parameter settings while the gradientless descent and the simplified direct search are (which can be understood directly from the algorithm themselves). This property does not show up in conver-

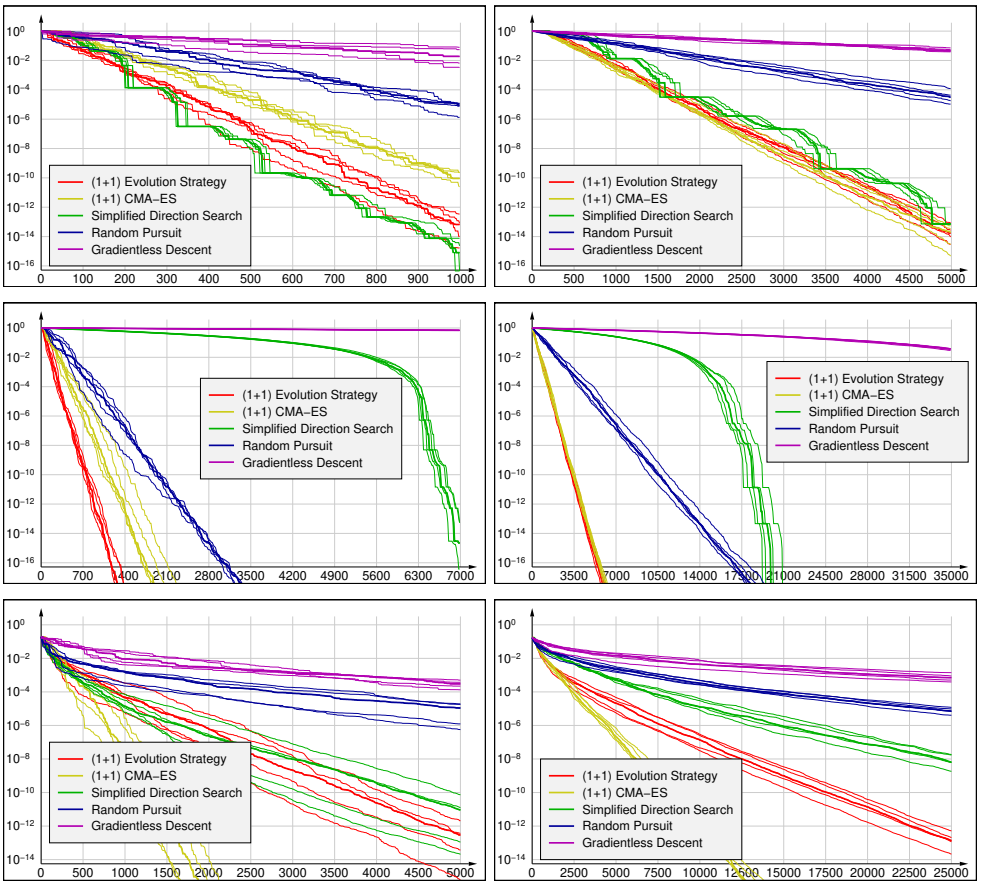


Fig. A.1: Comparison of (1+1)-ES with and without covariance matrix adaptation with three well-analyzed DFO algorithms on two convex quadratic functions. The top row of plots shows the performance on the sphere function $\|x\|^2$ in dimensions 10 (left) and 50 (right). The middle row shows the same problem,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

gence rates and is therefore often (but not always) neglected in algorithm design. The last experiment clearly demonstrates the benefit of variable-metric methods like CMA-ES. It should be noted that variable metric techniques can be implemented into most existing algorithms. This is rarely done though, with random pursuit being a notable exception [63].

Appendix B. Proofs.

B.1. Proof of Lemma 2.4.

Proof. If $f(x) = \frac{1}{2}\|x - x^*\|^2$, then we can compute the suboptimality function in x which is the d-th rooth of the volume of a sphere of radius $\|x - x^*\|$. Hence

$f_\mu(x) = \frac{\sqrt{\pi}}{\Gamma(d/2+1)^{1/d}}\|x - x^*\| = V_d\|x - x^*\|$. Then assuming w.l.g. $x^* = 0$

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

$$= \Pr_{z\sim\mathcal{N}(0,\mathrm{I})}\left[\left\|\frac{m}{\|m\|} + V_d\bar{\sigma}z\right\| \leqslant (1-r)\right] \quad . \qquad \square \tag{B.2}$$

By using the isotropy of the multivariate normal distribution we find that

$$\Pr_{z\sim\mathcal{N}(0,\mathrm{I})}\left[\left\|\frac{m}{\|m\|} + V_d\bar{\sigma}z\right\| \leqslant (1-r)\right] = \Pr_{z\sim\mathcal{N}(0,\mathrm{I})}\left[\|e_1 + V_d\bar{\sigma}z\| \leqslant (1-r)\right] \quad .$$

B.2. Proof of Proposition 2.7.

Proof. We need to investigate $\liminf_{\sigma\to 0}\Pr_{z\sim\mathcal{N}(0,\Sigma)}\left(h(m+\sigma z) \leqslant h(m)\right)$. We assume first that f is continuously differentiable. From the Taylor's theorem with Lagrange's form remainder, we know that for all z, for all σ, there exists $c_{z,\sigma} \in [0,1]$ such that

$$h(m+\sigma z) = h(m) + \sigma\nabla h(m + c_{z,\sigma}\sigma z)^\top z$$

such that $1_{\{h(m+\sigma z)\leqslant h(m)\}} = 1_{\{\nabla h(m+c_{z,\sigma}\sigma z)^\top z\leqslant 0\}}$ and thus

$$\Pr_{z\sim\mathcal{N}(0,\Sigma)}\left(h(m+\sigma z) \leqslant h(m)\right) = \int 1_{\{\nabla h(m+c_{z,\sigma}\sigma z)^\top z\leqslant 0\}} p_{\mathcal{N}(0,\Sigma)}(z)dz \quad . \tag{B.3}$$

For all z, $\lim_{\sigma\to 0}\nabla h(m + c_{z,\sigma}\sigma z) = \nabla h(m)$ since $w \mapsto \nabla h(w)$ is continuous. Hence if $\nabla h(m) \neq 0$, then $1_{\{\nabla h(m+c_{z,\sigma}\sigma z)^\top z\leqslant 0\}}$ converges to $1_{\{\nabla h(m)^\top z\leqslant 0\}}$ for all $z \neq 0$ by continuity of $t \mapsto 1_{\{t\leqslant 0\}}$ for $t \neq 0$. Since $|1_{\{\nabla h(m+c_{z,\sigma}\sigma z)^\top z\leqslant 0\}} p_{\mathcal{N}(0,\Sigma)}(z)| \leqslant p_{\mathcal{N}(0,\Sigma)}(z)$, by the Lebesgue dominated convergence theorem we obtain that

$$\lim_{\sigma\to 0}\Pr_{z\sim\mathcal{N}(0,\Sigma)}\left(h(m+\sigma z) \leqslant h(m)\right) = \int 1_{\{\nabla h(m)^\top z\leqslant 0\}} p_{\mathcal{N}(0,\Sigma)}(z)dz = 1/2 \quad ,$$

where the latter equality comes from the symmetry of the multivariate normal distribution. We therefore have shown that if $\nabla h(m) \neq 0$, then m is p-improvable with $p = 1/2$.

When $\nabla h(m) = 0$ and h is twice continuously differentiable, then Taylor's theorem with Lagrange's form remainder implies that for all z, for all σ, there exists $c_{z,\sigma} \in [0,1]$ such that

$$h(m+\sigma z) = h(m) + \frac{1}{2}\sigma^2 z^\top\nabla^2 h(m + c_{z,\sigma}\sigma z)z.$$

Therefore $1_{\{h(m+\sigma z)\leqslant h(m)\}} = 1_{\{z^\top\nabla^2 h(m+c_{z,\sigma}\sigma z)z\leqslant 0\}}$ and thus

$$\Pr_{z\sim\mathcal{N}(0,\Sigma)}\left(h(m+\sigma z) \leqslant h(m)\right) = \int 1_{\{z^\top\nabla^2 h(m+c_{z,\sigma}\sigma z)z\leqslant 0\}} p_{\mathcal{N}(0,\Sigma)}(z)dz \quad . \tag{B.4}$$

By continuity of $\nabla^2 h$, $\nabla^2 h(m + c_{z,\sigma}\sigma z) \to \nabla^2 h(m)$ when σ goes to 0. By a change of variable in the previous integral, we can assume that $\nabla^2 h(m)$ is a diagonal matrix with diagonal entries $(\lambda_1, \ldots, \lambda_n)$ (this implies a change of the matrix Σ for integrating with respect to $p_{\mathcal{N}(0,\Sigma)}$. With an abuse of notation we continue to denote this matrix Σ).

We have assumed that at least one $\lambda_i \neq 0$. Hence the set $A = \{z | z^\top \nabla^2 h(m) z = 0\} = \{z | \sum_i \lambda_i z_i^2 = 0\}$ is of Lebesgue measure zero. For all $z \in \mathbb{R}^n \backslash A$, by continuity of $t \in \mathbb{R} \to 1_{t \leqslant 0}$ on $\mathbb{R} \backslash \{0\}$ we have that $\lim_{\sigma \to 0} 1_{\{z^\top \nabla^2 h(m + c_{z,\sigma} \sigma z) z \leqslant 0\}} = 1_{\{z^\top \nabla^2 h(m) z \leqslant 0\}}$. By the Lebesgue dominated convergence theorem we find that

$$\begin{aligned}
\lim_{\sigma \to 0} \int 1_{\{z^\top \nabla^2 h(m + c_{z,\sigma} \sigma z) z \leqslant 0\}} p_{\mathcal{N}(0,\Sigma)}(z) dz &= \int 1_{\{z^\top \nabla^2 h(m) z \leqslant 0\}} p_{\mathcal{N}(0,\Sigma)}(z) dz \ , \\
&= \int_{\mathbb{R}^n \backslash A} 1_{\{z^\top \nabla^2 h(m) z \leqslant 0\}} p_{\mathcal{N}(0,\Sigma)}(z) dz \\
&= \int_{\mathbb{R}^n \backslash A} 1_{\{z^\top \nabla^2 h(m) z < 0\}} p_{\mathcal{N}(0,\Sigma)}(z) dz
\end{aligned}$$

Hence, if $\nabla^2 f(m)$ has only non-positive eigenvalue $\lambda_i \leqslant 0$ (in addition to at least one $\lambda_i \neq 0$), then $1_{\{z^\top \nabla^2 h(m) z \leqslant 0\}} = 1$ and thus

$$\lim_{\sigma \to 0} \int 1_{\{z^\top \nabla^2 h(m + c_{z,\sigma} \sigma z) z \leqslant 0\}} p_{\mathcal{N}(0,\Sigma)}(z) dz = 1 \ ,$$

such that m is p-improvable with $p = 1$. If all $\lambda_i \geqslant 0$ with one $\lambda_i > 0$, then $1_{\{z^\top \nabla^2 h(m) z < 0\}} = 0$ such that m is p-critical.

If $\nabla^2 f(m)$ has one strictly negative eigenvalue, say w.l.g. $\lambda_1 < 0$ then the set $S = \{z : \sum_i \lambda_i z_i^2 \leqslant 0\}$ has a positive Lebesgue measure, because the set $[(1 - \epsilon)/\sqrt{\lambda_1}, (1 + \epsilon)/\sqrt{\lambda_1}] \times [-\epsilon, \epsilon] \times \ldots \times [-\epsilon, \epsilon]$ is included in S for small enough ϵ. Since the Lebesgue measure and the measure associated with a Gaussian distribution are equivalent, it holds

$$\int 1_{\{\sum_i \lambda_i z_i^2 < 0\}} p_{\mathcal{N}(0,\Sigma)}(z) dz > 0 \ .$$

Therefore m is p-improvable with $p = \Pr_{z \sim \mathcal{N}(0,\Sigma)} \left(z^{\mathrm{T}} \nabla^2 h(m) z \leqslant 0 \right) > 0$. ∎

B.3. Proof of Proposition 2.8.

Proof. By definition $f_\mu(x)$ is the dth root of the volume of the hyper-ellipsoid $\{y : h(y) \leqslant h(x)\}$, which is equal to

$$\begin{aligned}
\int_{\{y : (y - x^*)^{\mathrm{T}} H (y - x^*)/2 \leqslant h(x)\}} \mathrm{d}y &= \det(H)^{-1/2} \int_{\{z : \|z\|^2 \leqslant 2h(x)\}} \mathrm{d}z \\
&= V_d^d \det(H)^{-1/2} (2h(x))^{d/2} \ .
\end{aligned}$$

Hence by taking the dth rooth we find (2.12).

We will now prove the expressions for C_ℓ and C_u. We first remark that since H is symmetric positive with $\lambda_{\max}(H)$ and $\lambda_{\min}(H)$ as minimal eigenvalues then

$$\frac{1}{2} \lambda_{\min}(H) \|x - x^*\|^2 \leqslant h(x) \leqslant \frac{1}{2} \lambda_{\max}(H) \|x - x^*\|^2 \ . \tag{B.5}$$

Let $y \in \mathbb{R}^d$ such that $\|y - x^*\| < C_\ell f_\mu(x)$. Then, using the RHS of (B.5) in the expression of $f_\mu(y)$ given in (2.12) we find

$$\begin{aligned}
f_\mu(y) &\leqslant V_d \cdot \left(\lambda_{\max}(H) \|y - x^*\|^2 / \det(H)^{1/d} \right)^{1/2} \\
&< V_d \cdot \left(\lambda_{\max}(H) / \det(H)^{1/d} \right)^{1/2} C_\ell f_\mu(x) = f_\mu(x) \ .
\end{aligned}$$

Therefore, $\{y : \|y - x^*\| < C_\ell f_\mu(x)\} \subset S_0(x)$. Similarly, we can verify the expression for C_u.

The above arguments show that A1 holds with $a = 0$ and $b = \infty$. Since h is continuously differentiable, Proposition 2.7 guarantees that A2 holds with the same a and b. The measurability of f comes from the property that a strictly increasing function is measurable and h is continuous and thus measurable. ∎

B.4. Proof of Lemma 2.9.

Proof. Since f_μ is invariant to g, without loss of generality we assume $f(x) = h(x) - h(x^*)$ in this proof. Let x be such that $f(x) \leqslant \frac{L_u b^2}{2V_d^2}$. Inequality (2.15) implies that $f(y) \leqslant f(x) \Rightarrow (L_\ell/2)\|y - x^*\|^2 \leqslant f(x)$, meaning

$$\{y : f(y) \leqslant f(x)\} \subseteq \bar{\mathcal{B}}\left(x^*, \sqrt{\frac{f(x)}{L_\ell/2}}\right) . \tag{B.6}$$

Since $f_\mu(x)$ is the dth root of the volume of the left-hand side (LHS) of the above relation, we find

$$f_\mu(x) \leqslant \mu^{\frac{1}{d}} \left(\bar{\mathcal{B}}\left(x^*, \sqrt{\frac{f(x)}{L_\ell/2}}\right)\right) = V_d \sqrt{\frac{f(x)}{L_\ell/2}} . \tag{B.7}$$

Analogously, we obtain

$$\mathcal{B}\left(x^*, \sqrt{\frac{f(x)}{L_u/2}}\right) \subseteq \{y : f(y) < f(x)\} . \tag{B.8}$$

and

$$f_\mu(x) \geqslant V_d \sqrt{\frac{f(x)}{L_u/2}} . \tag{B.9}$$

Inequalities (B.6) and (B.9) imply

$$\{y : f(y) \leqslant f(x)\} \subseteq \bar{\mathcal{B}}\left(x^*, \sqrt{\frac{L_u}{L_\ell}} \frac{f_\mu(x)}{V_d}\right) . \tag{B.10}$$

Analogously, inequalities (B.8) and (B.7) imply

$$\mathcal{B}\left(x^*, \sqrt{\frac{L_\ell}{L_u}} \frac{f_\mu(x)}{V_d}\right) \subseteq \{y : f(y) < f(x)\} . \tag{B.11}$$

Since $f(x) > \frac{L_u b^2}{2V_d^2} \Rightarrow f_\mu(x) \geqslant b$ because of (B.9), we find $\{x : f_\mu(x) \in (0, b]\} \subseteq \left\{x : f(x) \in \left(0, \frac{L_u b^2}{2V_d^2}\right]\right\}$. This implies A1 for $\mathcal{X}_0^b$. This completes the proof. ∎

B.5. Proof of Proposition 2.11.

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

It is easy to see that the spatial suboptimality function $f_\mu(x)$ is proportional to $h(x) - h(x^*)$. Let $f_\mu(x) = c \cdot (h(x) - h(x^*))$ for some $c > 0$. Then, f_μ is also a

homogeneous function. Since it is homogeneous, A1 reduces to that there are open and closed balls with radius C_ℓ and C_u satisfying the conditions described in the assumption with $R = 1$. Such constants are obtained by $C_u = \sup\{\|x - x^*\| : f_\mu(x) = 1\}$ and $C_\ell = \inf\{\|x - x^*\| : f_\mu(x) = 1\}$.

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

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

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

B.6. Proof of Proposition 4.1.

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

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

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

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

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

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

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

The integral is minimized if the ball is at the opposite side of m on the ball $\bar{\mathcal{B}}_u$, see Figure 4.1. By a variable change (moving m to the origin) and letting $e_m = m/\|m\|$,

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

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

B.7. Proof of Lemma 4.2.

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

(B.12)
$$
p^{\mathrm{succ}}_r(\bar{\sigma}; m, \Sigma) \geqslant \kappa^{-d/2}\Phi\left(\mathcal{B}\left(\left(\frac{(2C_u-(1-r)C_\ell)\kappa^{1/2}}{\bar{\sigma}}\right)e_1, \frac{(1-r)C_\ell\kappa^{1/2}}{\bar{\sigma}}\right); 0, \mathrm{I}\right).
$$

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

B.8. Proof of Proposition 4.3.

Proof. First, we remark that $m_t \in \mathcal{X}_{a,b}$ is equivalent to the condition $a < f_\mu(m_t) \leqslant b$. If $f_\mu(m_t) \leqslant a$ or $f_\mu(m_t) > b$, both sides of (4.9) are zero, hence the inequality is trivial. In the following we assume that $m_t \in \mathcal{X}^b_a$.

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

(B.13)
$$
V(\theta_t) = \log\left(f_\mu(m_t)\right) + v \cdot \log^+\left(\frac{\alpha_\uparrow \cdot \ell \cdot f_\mu(m_t)}{\sigma_t}\right) + v \cdot \log^+\left(\frac{\sigma_t}{\alpha_\downarrow \cdot u \cdot f_\mu(m_t)}\right) \quad .
$$

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

$$
\begin{aligned}
V(\theta_{t+1}) &= \log\left(f_\mu(m_{t+1})\right) && \text{(B.14a)} \\
&\quad + v \cdot \log^+\left(\frac{\ell \cdot f_\mu(m_{t+1})}{\sigma_t}\right) 1_{\{\sigma_{t+1} > \sigma_t\}} && \text{(B.14b)} \\
&\quad + v \cdot \log^+\left(\frac{\alpha_\uparrow \cdot \ell \cdot f_\mu(m_t)}{\alpha_\downarrow \cdot \sigma_t}\right) 1_{\{\sigma_{t+1} < \sigma_t\}} && \text{(B.14c)} \\
&\quad + v \cdot \log^+\left(\frac{\alpha_\uparrow \cdot \sigma_t}{\alpha_\downarrow \cdot u \cdot f_\mu(m_{t+1})}\right) 1_{\{\sigma_{t+1} > \sigma_t\}} && \text{(B.14d)} \\
&\quad + v \cdot \log^+\left(\frac{\sigma_t}{u \cdot f_\mu(m_t)}\right) 1_{\{\sigma_{t+1} < \sigma_t\}} \quad . && \text{(B.14e)}
\end{aligned}
$$

We want to estimate the conditional expectation

$$\mathbb{E}\left[\max\{V(\theta_{t+1}) - V(\theta_t)\,,\, -A\} \mid \theta_t\right]. \tag{B.15}$$

We partition the possible values of θ_t into three sets: first the set of θ_t such that $\sigma_t < \ell \cdot f_\mu(m_t)$ (σ_t is small), second the set of θ_t such that $\sigma_t > u \cdot f_\mu(m_t)$ (σ_t is large), and last the set of θ_t such that $\ell \cdot f_\mu(m_t) \leqslant \sigma_t \leqslant u \cdot f_\mu(m_t)$ (reasonable σ_t). In the following, we bound (B.15) for each of the three cases and in the end our bound B will equal the minimum of the three bounds obtained for each case.

Reasonable σ_t case: $\frac{f_\mu(m_t)}{\sigma_t} \in \left[\frac{1}{u}, \frac{1}{\ell}\right]$. In case of success, where $1_{\{\sigma_{t+1}>\sigma_t\}} = 1$, we have $f_\mu(m_{t+1})/\sigma_{t+1} \leqslant f_\mu(m_t)/(\alpha_\uparrow \sigma_t) \leqslant 1/(\alpha_\uparrow \ell)$, implying that (B.14b) is always 0. Similarly, in case of failure, $f_\mu(m_{t+1})/\sigma_{t+1} = f_\mu(m_t)/(\alpha_\downarrow \sigma_t) \geqslant 1/(\alpha_\downarrow u)$ and we find that (B.14e) is always zero. We rearrange (B.14c) and (B.14d) into

$$\text{(B.14c)} = v \cdot \log^+\left(\frac{\alpha_\uparrow \cdot \ell \cdot f_\mu(m_t)}{\alpha_\downarrow \cdot \sigma_t}\right) 1_{\{\sigma_{t+1}<\sigma_t\}}\ ,$$

$$\text{(B.14d)} = v \cdot \left[\log\left(\frac{\alpha_\uparrow \cdot \sigma_t}{\alpha_\downarrow \cdot u \cdot f_\mu(m_t)}\right) - \log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)\right] 1_{\left\{\frac{\alpha_\downarrow u f_\mu(m_{t+1})}{\alpha_\uparrow \sigma_t}<1\right\}} 1_{\{\sigma_{t+1}>\sigma_t\}}\ .$$

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

$$\begin{aligned}
\Delta_t \leqslant & \left(1 - v \cdot 1_{\left\{\frac{\alpha_\downarrow u f_\mu(m_t)}{\alpha_\uparrow \sigma_t}<1\right\}} \cdot 1_{\{\sigma_{t+1}>\sigma_t\}}\right) \log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right) \\
& + v \cdot \log^+\left(\frac{\alpha_\uparrow \cdot \ell \cdot f_\mu(m_t)}{\alpha_\downarrow \cdot \sigma_t}\right) \cdot 1_{\{\sigma_{t+1}<\sigma_t\}} \\
& + v \cdot \log^+\left(\frac{\alpha_\uparrow \cdot \sigma_t}{\alpha_\downarrow \cdot u \cdot f_\mu(m_t)}\right) \cdot 1_{\{\sigma_{t+1}>\sigma_t\}} \\
\leqslant & (1-v) \log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right) \\
& + v \cdot \log^+\left(\frac{\alpha_\uparrow \cdot \ell \cdot f_\mu(m_t)}{\alpha_\downarrow \cdot \sigma_t}\right) \cdot 1_{\{\sigma_{t+1}<\sigma_t\}} \\
& + v \cdot \log^+\left(\frac{\alpha_\uparrow \cdot \sigma_t}{\alpha_\downarrow \cdot u \cdot f_\mu(m_t)}\right) \cdot 1_{\{\sigma_{t+1}>\sigma_t\}}\ .
\end{aligned}$$

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

$$\begin{aligned}
\max\{\Delta_t\,,\, -A\} \leqslant & (1-v) \max\left\{\log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)\,,\, -\frac{A}{1-v}\right\} \\
& + v \cdot \log^+\left(\frac{\alpha_\uparrow \cdot \ell \cdot f_\mu(m_t)}{\alpha_\downarrow \cdot \sigma_t}\right) \cdot 1_{\{\sigma_{t+1}<\sigma_t\}} \\
& + v \cdot \log^+\left(\frac{\alpha_\uparrow \cdot \sigma_t}{\alpha_\downarrow \cdot u \cdot f_\mu(m_t)}\right) \cdot 1_{\{\sigma_{t+1}>\sigma_t\}}\ .
\end{aligned}$$

To consider the expectation of the above upper bound, we need to compute the expectation of the maximum of $\log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)$ and $-\frac{A}{1-v}$. Let $a \leqslant 0$ and $b \in \mathbb{R}$ then $\max(a,b) = a \cdot 1_{\{a>b\}} + b \cdot 1_{\{a\leqslant b\}} \leqslant b \cdot 1_{\{a\leqslant b\}}$. Applying this and taking the conditional expectation, a trivial upper bound for the conditional expectation of $\max\left\{\log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)\,,\, -\frac{A}{1-v}\right\}$ is $-\frac{A}{1-v}$ times the probability of $\log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right)$ being

no greater than $-\frac{A}{1-v}$. The latter condition is equivalent to $f_\mu(m_{t+1}) \leqslant (1-r)\cdot f_\mu(m_t)$ corresponding to successes with rate $r = 1 - \exp\left(-\frac{A}{1-v}\right)$ or better. That is,

(B.16)
$$(1-v)\cdot \mathbb{E}\left[\max\left\{\log\left(\frac{f_\mu(m_{t+1})}{f_\mu(m_t)}\right),\, -\frac{A}{1-v}\right\}\right] \leqslant -A\cdot p_r^{\mathrm{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right) .$$

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

(B.17)
$$\begin{aligned}
&\mathbb{E}\left[\max\{\Delta_t\,,\, -A\}|\theta_t\right] \\
&\leqslant -A\cdot p_r^{\mathrm{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right) \\
&\quad + \left(\log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) + \underbrace{\log\left(\frac{\ell\cdot f_\mu(m_t)}{\sigma_t}\right)}_{\leqslant 0}\right)\cdot v\cdot\left(1 - p_0^{\mathrm{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right)\right) \\
&\quad + \left(\log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) + \underbrace{\log\left(\frac{\sigma_t}{u\cdot f_\mu(m_t)}\right)}_{\leqslant 0}\right)\cdot v\cdot p_0^{\mathrm{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right) \\
&\leqslant -A\cdot p_r^* + v\cdot\log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) .
\end{aligned}$$

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.13) is positive. Moreover, if $\sigma_{t+1} < \sigma_t$, we have $\ell f_\mu(m_{t+1}) = \ell f_\mu(m_t) > \sigma_t > \sigma_{t+1}$ and hence the 2nd summand in (B.13) is positive for $V(\theta_{t+1})$ as well. If $\sigma_{t+1} > \sigma_t$, any regime can happen. Then,

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

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

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

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

$$\begin{aligned}
\mathbb{E}\left[\max\{\Delta_t\,,\, -A\} | \mathcal{F}_t\right] &\leqslant -v \cdot \left(\log(\alpha_\uparrow/\alpha_\downarrow) p_0^{\mathrm{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right) + \log(\alpha_\downarrow)\right) \\
&\leqslant -v \cdot \left(\log(\alpha_\uparrow/\alpha_\downarrow) p_\ell + \log(\alpha_\downarrow)\right) \\
&\leqslant -v \cdot \frac{p_\ell - p_u}{2} \cdot \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) \ .
\end{aligned} \tag{B.18}$$

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

Large σ_t case: $\frac{f_\mu(m_t)}{\sigma_t} < \frac{1}{u}$. Since $\frac{f_\mu(m_{t+1})}{\sigma_{t+1}} \leqslant \frac{f_\mu(m_t)}{\alpha_\downarrow \sigma_t} < \frac{1}{\alpha_\downarrow u}$, the 3rd summand in (B.13) is positive in both $V(\theta_t)$ and $V(\theta_{t+1})$. For the 2nd summand in (B.13), recall that $\alpha_\uparrow \ell f_\mu(m_t)/\sigma_t < \alpha_\uparrow \ell/u \leqslant \alpha_\downarrow < 1$ since we have assumed that $u/\ell \geqslant \alpha_\uparrow/\alpha_\downarrow$. Hence, for $V(\theta_t)$ the 2nd summand in (B.13) is zero. Also, $\alpha_\uparrow \ell \|m_{t+1}\|/\sigma_{t+1} \leqslant \alpha_\uparrow \ell/(\alpha_\downarrow u) = (\alpha_\uparrow/\alpha_\downarrow)\ell/u \geqslant 1$ and thus for $V(\theta_{t+1})$ the 2nd summand in (B.13) also equals 0. We obtain

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

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

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

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

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

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

$$\begin{aligned}
\mathbb{E}\left[\max\{\Delta_t\,,\, -A\} | \theta_t\right] &\leqslant v\left(\log(\alpha_\downarrow) + \log(\alpha_\uparrow/\alpha_\downarrow) p_0^{\mathrm{succ}}\left(\frac{\sigma_t}{f_\mu(m_t)}; m_t, \Sigma_t\right)\right) \\
&\leqslant v\left(\log(\alpha_\downarrow) + \log(\alpha_\uparrow/\alpha_\downarrow) p_u\right) \\
&\leqslant -v \cdot \frac{p_\ell - p_u}{2} \cdot \log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) \ .
\end{aligned} \tag{B.19}$$

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

Conclusion. Inequalities (B.17)–(B.19) together cover all possible cases and we hence obtain (4.9).

Finally, we prove the positivity of B for an arbitrary $A > 0$. Lemma 4.2 guarantees the positivity of p_r^* for any choice of A since $r = 1 - \exp(-A/(1-v)) \in (0,1)$ for any $A > 0$ and $v < 1$. Therefore, $A \cdot p_r^* > 0$ for any A and $v \leqslant \min(1,\ A/\log(1/\alpha_\downarrow),\ A/\log(\alpha_\uparrow))$. Moreover, for a sufficiently small v, p_r^* is strictly positive for any $A > 0$. Therefore, one can take a sufficiently small v that satisfies $A \cdot p_r^* > v \log(\alpha_\uparrow/\alpha_\downarrow)$. The first term in the minimum in (4.9) is positive. The second term therein is clearly positive for $v > 0$. Hence, their minimum is positive. This completes the proof. ∎

B.9. Proof of Lemma 4.11.

Proof. In (4.14), for increasing values of $\bar\sigma$ the ball-shaped integration area shrinks, and in case of $r > 0$ it also moves away from the origin. Together with the monotonicity of φ w.r.t. $\|z\|$ this proves that p_r is strictly decreasing. Continuity of p_r follows from the boundedness of φ, and positivity from the fact that the integral area in (4.14) is non-empty and φ is positive. This proves the first claim. For $r = 0$ the balls are nested. This immediately proves the second claim. The last claim follows from Propositions 2.7 and 4.1. ∎

B.10. Proof of Lemma 4.12.

Proof. Fix $\bar\sigma > 0$. Let $Z \sim \mathcal{N}(0,1)$ and $Y_{d-1} \sim \chi^2_{d-1}$ (χ^2 random variable with $d-1$ degrees of freedom) be independent. Define $X_d = Z + \frac{\hat\sigma}{2d} Z^2 + \frac{\hat\sigma}{2d} Y_{d-1} + \frac{r\cdot d}{\hat\sigma} - \frac{r^2\cdot d}{2\hat\sigma}$ and $X = Z + \frac{\hat\sigma}{2} + \frac{\rho}{\hat\sigma}$. Then, by a simple derivation, we find that $p_r(\hat\sigma) = \Pr[X_d \leqslant 0]$ and $\Psi\left(-\frac{\rho}{\hat\sigma} - \frac{\hat\sigma}{2}\right) = \Pr[X \leqslant 0]$, i.e., they are the cumulative densities evaluated at 0. For the sake of simplicity, we denote these cumulative density functions by F_d and F. The convergence in distribution of a collection of random variables generally implies that $\lim_{d\to\infty} F_d(x) = F(x)$ for all x where F is continuous. In our case, F is continuous everywhere. Therefore, we have $\lim_{d\to\infty} F_d(0) = F(0)$ if X_d converges in distribution to X. The convergence in distribution is implied by the almost sure convergence, which is proved by the strong law of large numbers of X_d, where $Z^2/d \to 0$, $Y_{d-1}/d \to 1$, $r\cdot d \to \rho$, and $r^2 \cdot d \to 0$ almost surely. This argument holds for any $\hat\sigma > 0$, resulting in the pointwise convergence $\lim_{d\to\infty} p_r(\hat\sigma) = p_\rho^{\lim}(\hat\sigma) = \Psi\left(-\frac{\rho}{\hat\sigma} - \frac{\hat\sigma}{2}\right)$. The rest of the claims are trivial from the properties of Ψ. ∎

B.11. Proof of Proposition 4.13.

Proof. Consider $d \geqslant 2$. We set $A = 1/d$. We bound B from below by taking a specific value for $v \in (0,\ \min(1, A/\log(1/\alpha_\downarrow),\ A/\log(\alpha_\uparrow))$ instead of considering sup for v.

Our candidate is $v = \frac{A\cdot p'}{\log(\alpha_\uparrow/\alpha_\downarrow)} \cdot \frac{2}{(2+p_\ell - p_u)}$, where $p' = \inf_{\bar\sigma\in[\ell,u]} p_{r'}(\bar\sigma)$ and $r' = 1 - \exp\left(-\frac{A}{1-\frac{1}{d\cdot\log(\alpha_\uparrow/\alpha_\downarrow)}}\right)$. It holds $v < \frac{1}{d\cdot\log(\alpha_\uparrow/\alpha_\downarrow)}$ and hence $r' > r$, from which we obtain $p' < p^*$.

Now we consider the terms in (4.9). The first term is lower bounded as

$$A\cdot p^* - v\log(\alpha_\uparrow/\alpha_\downarrow) = \frac{p'}{d}\left(\frac{p^*}{p'} - \frac{2}{2+p_\ell - p_u}\right) \geqslant \frac{p'}{d}\left(\frac{p_\ell - p_u}{2+p_\ell - p_u}\right) \ ,$$

whereas the second term is

$$v\frac{p_\ell - p_u}{2}\log\left(\frac{\alpha_\uparrow}{\alpha_\downarrow}\right) = \frac{p'}{d}\frac{p_\ell - p_u}{2+p_\ell - p_u} \ .$$

Therefore, we have

$$B \geqslant \frac{p'}{d} \frac{p_\ell - p_u}{2 + p_\ell - p_u} \quad . \tag{B.20}$$

Note that one can take $p_\ell - p_u \in \Theta(1)$ since the only condition is $p_{\text{target}} = (p_\ell + p_u)/2 \in \Theta(1)$. To obtain $B \in \Omega(1/d)$, it is sufficient to show $p' \in \Theta(1)$ for $d \to \infty$.

Fix p_ℓ and p_u independently of d. In the light of Lemma 4.11, we have that $p_0 : \mathbb{R}_> \to (0, 1/2)$ is continuous and strictly decreasing from $1/2$ to 0 for all $d \in \mathbb{N}$. Therefore, for each $d \in \mathbb{N}$ there exists an inverse map $p_0^{-1} : (0, 1/2) \to \mathbb{R}_>$. Define $\hat{\sigma}_\ell^d = d \cdot V_d \cdot p_0^{-1}(p_\ell)$ and $\hat{\sigma}_u^d = d \cdot V_d \cdot p_0^{-1}(p_u)$ for each $d \in \mathbb{N}$. It follows from Lemma 4.12 that $p_0^{\lim}$ is also strictly decreasing, hence invertible. We let $\hat{\sigma}_\ell^\infty = (p_0^{\lim})^{-1}(p_\ell)$ and $\hat{\sigma}_u^\infty = (p_0^{\lim})^{-1}(p_u)$. Because of the pointwise convergence of $p_0(\bar{\sigma} = \hat{\sigma}/(dV_d))$ to $p_0^{\lim}(\hat{\sigma})$, we have $\hat{\sigma}_\ell^d \to \hat{\sigma}_\ell^\infty$ and $\hat{\sigma}_u^d \to \hat{\sigma}_u^\infty$ for $d \to \infty$. Hence, for any $\hat{u} > \hat{\sigma}_u^\infty$ and $\hat{\ell} < \hat{\sigma}_\ell^\infty$ with $u/\ell \geqslant \alpha_\uparrow/\alpha_\downarrow$, there exists $D \in \mathbb{N}$ such that for all $d \geqslant D$ we have $\hat{u} > \hat{\sigma}_u^d$ and $\hat{\ell} < \hat{\sigma}_\ell^d$. Now we fix $\hat{u}$ and $\hat{\ell}$ in this way. This amounts to selecting $u = d \cdot V_d \cdot \hat{u}$ and $\ell = d \cdot V_d \cdot \hat{\ell}$.

We have $\lim_{d\to\infty} d \cdot r' = 1$ since $\lim_{d\to\infty} d \cdot \log(\alpha_\uparrow/\alpha_\downarrow) = \infty$ and hence according to Lemma 4.12 we have

$$\begin{aligned}
\liminf_{d\to\infty} p' &= \liminf_{d\to\infty} \left(\min_{\bar{\sigma}\in[\ell,u]} \{p_{r'}(\bar{\sigma})\} \right) \\
&= \liminf_{d\to\infty} \left(\min_{\hat{\sigma}\in[\hat{\ell},\hat{u}]} \left\{ p_{r'} \left(\frac{\hat{\sigma}}{d \cdot V_d} \right) \right\} \right) \\
&\overset{(\star)}{=} \min_{\hat{\sigma}\in[\hat{\ell},\hat{u}]} \left\{ \lim_{d\to\infty} \left(p_{r'} \left(\frac{\hat{\sigma}}{d \cdot V_d} \right) \right) \right\} = \min_{\hat{\sigma}\in[\hat{\ell},\hat{u}]} \Psi \left(-\frac{1}{\hat{\sigma}} - \frac{\hat{\sigma}}{2} \right) \quad ,
\end{aligned}$$

where the equality $(\star)$ follows from the pointwise convergence of $p_{r'}$ to $\lim_{d\to\infty} p_{r'}$ and the continuity of $p_{r'}$ and $\lim_{d\to\infty} p_{r'}$.[5] This completes the proof. ∎

Appendix C. Technical Lemma.

LEMMA C.1. *For all $\Sigma \in \mathcal{S}_\kappa$ the multivariate Gaussian density fulfills*

$$\kappa^{-d/2} \varphi\left(x; 0, \kappa^{-1}\mathrm{I}\right) \leqslant \varphi\big(x; 0, \Sigma\big) \leqslant \kappa^{d/2} \varphi\left(x; 0, \kappa\mathrm{I}\right).$$

Moreover, it holds

$$\kappa^{-d/2} \Phi\left(\mathcal{B}(\kappa^{1/2}c, \kappa^{1/2}r); 0, \mathrm{I}\right) \leqslant \Phi\big(\mathcal{B}(c, r); 0, \Sigma\big) \leqslant \kappa^{d/2} \Phi\left(\mathcal{B}(c/\kappa^{1/2}, r/\kappa^{1/2}); 0, \mathrm{I}\right).$$

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

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

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

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