THÈSE DE DOCTORAT DE

L'ÉCOLE POLYTECHNIQUE DE TUNISIE & L'UNIVERSITÉ PIERRE ET MARIE CURIE

SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES

présentée par

Mohamed JEBALIA

pour obtenir les grades de

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE
DOCTEUR DE L'ÉCOLE POLYTECHNIQUE DE TUNISIE

Sujet de la thèse :

Optimisation par Stratégies d'Évolution : Convergence et vitesses de convergence pour des fonctions bruitées - Résolution d'un problème d'identification

Soutenue le 19/12/2008, devant le jury composé de :

Mlle	Anne Auger	Co-directrice de thèse (Chargée de Recherche, INRIA Saclay)
Mr	Dirk Arnold	Rapporteur (Professeur Associé, Dalhousie University, Canada)
Mr	Olivier François	Rapporteur (Professeur, INP, Grenoble)
Mr	Taïeb Hadhri	Responsable de cotutelle (Professeur, Ecole Polytechnique de Tunisie)
Mr	Frédéric Hecht	Examinateur (Professeur, Université Pierre et Marie Curie, Paris)
Mr	Pierre Liardet	Examinateur (Professeur, Université de Provence, Marseille)
Mme	Marie Postel	Examinatrice (Maître de Conférences, Univ. Pierre et Marie Curie, Paris)
Mr	Marc Schoenauer	Responsable de cotutelle (Directeur de Recherche, INRIA Saclay)

Contents

Chapter 1
Non linear continuous optimization

Theoretical and Numerical Study

Application

Chapter 5
Identification of the Isotherm Function in Chromatography Using CMA-ES

Introduction

Résumé de la thèse

Un problème d'optimisation non linéaire continu peut être formulé comme suit : Étant donné une fonction $f : \mathbb{R}^d \mapsto \mathbb{R}$, appelée fonction objectif, le but est de chercher, dans un espace contenant une ou plusieurs parties ouvertes de $\mathbb{R}^d$, le vecteur (soit d paramètres) qui maximise (ou minimise) la fonction f.
Dans cette thèse, on s'intéresse à l'optimisation non linéaire continue par des méthodes appelées Stratégies d'Évolution (SE), algorithmes évolutionnaires dédiés à l'optimisation sur un espace continu. Les SE ont montré leur efficacité pratique pour la résolution de problèmes d'optimisation réels. Cependant les SE, comme l'ensemble des algorithmes évolutionnaires, ne sont pas basés sur les premiers principes, mais adaptés d'une imitation des principes de l'évolution naturelle, la survie des individus les plus adaptés. Dans une première partie de cette thèse, on étudie théoriquement et numériquement la convergence des SE, en particulier dans le cadre de l'optimisation des fonctions objectifs bruitées. On montre par exemple que des niveaux assez élevés du bruit peuvent entraîner la non-convergence de l'algorithme. Les expressions des vitesses de convergence sont ensuite établies théoriquement. Les cas de convergence et de divergence sont distingués théoriquement et numériquement.
La seconde partie traite une application à un problème réel en génie chimique, l'identification de paramètres pour le système de la chromatographie analytique. L'approche évolutionnaire est comparée à une méthode déterministe basée sur le calcul du gradient numérique. L'approche évolutionnaire est plus robuste sur ce cas d'étude spécifique.

Introduction (In French)

Les problèmes d'optimisation sont très fréquents dans l'industrie comme dans différents domaines de la recherche. L'optimisation non linéaire continue s'intéresse aux problèmes où la fonction à optimiser, appelée fonction objectif, ou *fitness*, est définie sur un espace d'état continu de dimension d, ç.à.d., $f : \mathbb{R}^d \mapsto \mathbb{R}$, et n'est pas linéaire. Le but est donc, de chercher d paramètres réels qui maximisent (ou minimisent) une fonction f.

Pour résoudre les problèmes d'optimisation, plusieurs méthodes ont été développées. La plupart de ces méthodes sont itératives, et génèrent, à l'itération n, une (ou plusieurs) nouvelle(s) solution(s) soit de manière déterministe, soit de manière stochastique en échantillonnant une distribution de probabilité. Ces méthodes peuvent être donc classées en deux grandes familles : méthodes de recherche déterministe et méthodes de recherche stochastique.

Dans les problèmes réels d'optimisation, le processus de recherche de la (ou des) solution(s) optimale(s) peut s'avérer difficile. Les fonctions objectifs peuvent être non convexes, irrégulières, bruitées, multimodales, mal conditionnées, non séparables ... Les contraintes sur l'espace de recherche peuvent aussi rendre la recherche encore plus difficile. Enfin, la difficulté du problème d'optimisation croît également avec la dimension d de l'espace de recherche.

Certaines études empiriques [122, 55, 82, 106, 9] comparant les méthodes d'optimisation et en particulier les méthodes de recherche stochastique aux méthodes de recherche déterministe donnent un avantage aux méthodes de recherche stochastique quand les fonctions objectifs sont de plus en plus complexes à optimiser, i.e., quand les fonctions objectifs sont plutôt non-convexes, multi-modales, très mal conditionnées, non séparables, ou bruitées. En particulier, dans le cadre de l'optimisation de fonctions bruitées, qui constitue la majeure partie de cette thèse, les études empiriques [106, 9] montrent que les méthodes de recherche stochastique appelées Stratégies d'Évolution (SE) sont plus robustes face au bruit que les méthodes déterministes.

Les Stratégies d'Évolution sont des algorithmes évolutionnaires dédiés à l'optimisation continue. Ils ont montré leur efficacité pratique pour la résolution de problèmes d'optimisation réels [51, 43, 22, 104, 142]. Cependant les SE, comme l'ensemble des algorithmes évolutionnaires, ne sont pas basés sur les premiers principes, mais sont le fruit d'une imitation des principes de l'évolution naturelle (la survie et la reproduction des individus les plus adaptés). La méthode "état de l'art" en optimisation évolutionnaire continue aujourd'hui est la

méthode CMA-ES, ou *Covariance Matrix Adaptation-Evolution Strategy*, introduite par N. Hansen et A. Ostermeier au milieu des années 90 [61, 59, 56]. Des études empiriques ont montré que CMA-ES est efficace et robuste face aux problèmes non séparables et mal conditionnés [61, 59, 82], mais est également efficace pour résoudre les problèmes multimodaux [56, 82]. D'autres études empiriques [61, 56, 55, 62] comparant CMA-ES à d'autres méthodes populaires de recherche stochastique ainsi qu'à la méthode BFGS, méthode de recherche déterministe très utilisée, ont montré une grande compétitivité de CMA-ES.

Dans cette thèse, on s'intéresse à l'optimisation non linéaire continue par Stratégies d'Évolution. La thèse comprend deux parties: la première est consacrée à des études théoriques et numériques concernant la convergence de Stratégies d'Évolution plus simples que CMA-ES, algorithmes qui sont les seuls à avoir été étudiés d'un point de vue théorique pour le moment. Dans cette partie, on s'intéresse en particulier à l'optimisation des fonctions quadratiques bruitées. La seconde partie traite une application à un problème réel en génie chimique, l'identification des paramètres de la loi de comportement (ou fonction *isotherme*) pour le système de la chromatographie analytique.

1 Étude théorique et numérique

1.1 État de l'art et contexte

Les premières études théoriques des Stratégies d'Évolution ont été des études asymptotiques par rapport à la dimension de l'espace de recherche ($d \to +\infty$) [25, 114]. Les premières études théoriques établies en dimension finie sont celles de François et Bienvenue [27] et de Auger [13, 17]. Il est ainsi aujourd'hui démontré [13, 17, 27] que la convergence de Stratégies d'Évolution adaptant leur pas de recherche à chaque itération est (log-)linéaire (*i.e.* le logarithme de la distance séparant la solution de l'optimum tend linéairement vers $-\infty$ en fonction du nombre d'itérations). Ce résultat est valable pour toute fonction qui s'écrit sous la forme $g(||x||^2)$ où g est une fonction strictement croissante. Pour des classes de fonctions bruitées (que l'on écrira sous la forme $||x||^2(1+\mathcal{N})$ ou $||x||(1+\mathcal{N})$, $\mathcal{N}$ étant une variable aléatoire modélisant le bruit), les études les plus poussées sont celles de Arnold et Beyer [5, 7, 8, 24, 25], études asymptotiques ici encore par rapport à la dimension d de l'espace de recherche.
La partie théorique de cette thèse concerne l'étude de la convergence des Stratégies d'Évolution, pour l'optimisation de fonctions, non bruitées et bruitées.

1.2 Contributions

Notre apport dans cette thèse est résumé dans les points suivants :

Optimisation des fonctions non bruitées (Résultats publiés dans [77]) :

Dans le contexte décrit ci-dessus, nous démontrons :
1) Une convergence log-linéaire d'un algorithme "artificiel" de type ES[1] appelé *scale-invariant* $(1+1)$-ES, dans lequel le pas de recherche à chaque itération est proportionnel à la distance qui sépare la solution courante de l'optimum (résultat similaire à ce qui a été prouvé dans [13, 17, 27] pour l'algorithme appelé $(1, \lambda)$-ES).
2) L'optimalité en terme de vitesse de convergence du scale-invariant $(1 + 1)$-ES. Ce résultat confirme le résultat montré dans [17] pour l'algorithme $(1, \lambda)$-ES.
Cette étude est présentée dans le chapitre 2.

Optimisation des fonctions bruitées (Résultats incluant ceux publiés dans [76])

Nous étudions le comportement des stratégies scale-invariant $(1 + 1)$-ES (chapitre 3) et scale-invariant $(1, \lambda)$-ES (chapitre 4) lors de la minimisation de fonctions bruitées. Nous montrons:

- Pour l'algorithme scale-invariant $(1 + 1)$-ES : les fonctions bruitées sont ici modélisées sous la forme $||x||^2(1 + \mathcal{N})$. La convergence montrée auparavant [77] pour les fonctions non bruitées n'est plus valable lorsque le niveau de bruit est suffisamment élevé pour que des valeurs négatives de la fonction objectif puissent être générées. Si la probabilité de l'évènement $(\mathcal{N} < -1)$ est strictement positive, l'algorithme ne converge pas (si le bruit est Gaussien) et diverge (si le bruit est minoré). Pour des distributions de bruit qui ne permettent de générer que des valeurs positives de la fonction objectif, l'algorithme converge toujours.

  Pour les fonctions objectifs qui s'écrivent sous la forme $(||x||^2+\alpha)(1+\mathcal{N})$ avec $\alpha > 0$, l'algorithme converge si les valeurs des fonctions objectifs générées ne peuveut être que positives. S'il y a une probabilité strictement positive que des valeurs négatives de la fonction objectif soient générées, l'algorithme ne converge pas. Nous comparons aussi nos résultats aux résultats obtenus dans [8] qui semblent en contradiction avec les résultats que nous avons obtenus.

  Dans une autre partie de cette étude, nous établissons théoriquement les expressions des vitesses de convergence (ou divergence) de l'algorithme lors de la minimisation des fonctions objectifs de la forme $||x||^2(1 + \mathcal{N})$. Les vitesses de convergence (ou divergence) obtenues peuvent être calculées numériquement. Pour des vitesses de convergence non nulles, le comportement de l'algorithme est log-linéaire.

- Pour l'algorithme scale-invariant $(1, \lambda)$-ES : les fonctions bruitées sont ici modélisées sous la forme $||x||(1 + \mathcal{N})$. Le comportement log-linéaire (convergence/divergence) est prouvé théoriquement. Les cas de divergence ou convergence de l'algorithme, en fonction du niveau de bruit et du pas de mutation, sont distingués théoriquement (lorsque $d \to +\infty$) et numériquement (pour $d < +\infty$). Nous montrons que les vitesses de convergence varient presque linéairement avec l'inverse de la dimension

[1] l'acronyme ES se rapporte à l'appellation anglophone pour les Stratégies d'Évolution: Evolution Strategies

de l'espace de recherche. Cette étude prouve rigoureusement que certaines approximations faites (lorsque d tend vers l'infini) dans [8] sont justifiées.

1.3 Outils mathématiques utilisés

Nous avons essentiellement utilisé dans notre étude des outils de la théorie de probabilité, tels que le Lemme de Borel-Cantelli, pour prouver la convergence presque sûre des algorithmes étudiés. Nous avons aussi eu recours aux différentes lois des grands nombres relatives aux variables aléatoires orthogonales [93] ou aux chaînes de Markov [97] pour étudier la stabilité des suites associées aux algorithmes étudiés.

2 Application (Résultats publiés dans [78])

La seconde partie de la thèse est constituée du chapitre 5. Elle s'attaque à un problème d'ingénierie réel. Le but est d'identifier les paramètres de la fonction *isotherme*, loi de comportement du processus de chromatographie utilisé en génie chimique. L'approche utilisée pour résoudre ce problème d'identification est de le poser sous la forme d'un problème d'optimisation. Pour résoudre le problème d'optimisation paramétrique ainsi obtenu, nous avons utilisé l'état de l'art en Stratégies d'Évolution, l'algorithme CMA-ES. La version de l'algorithme utilisé est celle décrite dans [16]. Ce problème a déjà été traité par des méthodes à base de descente de gradient dans [73, 74]. Nous avons testé l'approche évolutionnaire sur l'ensemble des donnés réelles publiées dans [73]. La comparaison de notre approche à celle du gradient numérique [73] a révélé que 1) L'algorithme CMA-ES converge toujours vers le même point indépendamment du point de départ (contrairement au gradient). 2) Les meilleures valeurs de la fonction objectif ont été trouvées par CMA-ES pour deux configurations expérimentales. En particulier CMA-ES est capable d'optimiser les 6 paramètres simultanément, alors que l'utilisation de l'algorithme à base de gradient a nécessité de fixer certaines valeurs de 2 des paramètres à partir de données expérimentales. Une autre remarque est que les temps de calcul entre CMA-ES et la méthodes à base de gradient sont comparables, alors qu'il est en général considéré que les méthodes déterministes sont nettement plus rapides que les méthodes stochastiques à base de population de solutions.

Note : la thèse est rédigée en anglais.

Summary of contributions

Optimization problems are frequently encountered in all domains of science and engineering. They are of particular relevance in industry. They include tasks such as scheduling, shape optimization, model calibration, and parameter identification. The goal of an optimization problem is to find the optimum (or the optima) of a real-valued function f defined on some search space Ω, subset of the d-dimensional space $\mathbb{R}^d$. Many methods have been developped to solve continuous optimization problems. They can be broadly categorized in two classes: deterministic and stochastic search methods.
Among stochastic search methods, the so-called Evolution Strategies (ES) have demonstrated their efficiency in solving real-world optimization problems. This motivates the general context of this thesis, continuous optimization using ES.

The work presented in this document can be divided into two parts: The first part deals with a theoretical and numerical study of some basic ES algorithms; The second part is devoted to an application that is tackled using the CMA-ES method.

1 Part 1: Theoretical and numerical study

This part is concerned with the theoretical and numerical study of the optimization, using ES, of objective functions having a unique global optimum. Therefore, this work can be classified as belonging to the studies of local convergence. The search space Ω is supposed to be unconstrained ($\Omega = \mathbb{R}^d$). We are interested in isotropic ES, i.e., ES where no search direction is preferred. We investigate the optimization of the following objective functions, that have been widely investigated in previous theoretical studies about ES:

- the so-called spherical functions, that can be written as $g(||x||^2)$, where g is a strictly increasing function and $||x||$ denotes the norm of vector $x \in \mathbb{R}^d$, and
- noisy objective functions, that are modelized as $||x||^2(1+\mathcal{N})$ or $||x||(1+\mathcal{N})$, where $\mathcal{N}$ is a random variable representing the noise.

The unique global optimum of spherical functions is $(0, \ldots, 0) \in \mathbb{R}^d$. Note that for noisy objective functions, the goal is to reach the optimum of the non-noisy part of the objective function, i.e., $(0, \ldots, 0)$. Our theoretical contributions in this thesis lies in Chapters 1, 2 and 3 and can be summarized as follows:

1.1 Optimization of non noisy functions (Results in [77]) :

In Chapter 2, we investigate the (1+1)-ES, and in particular the (1+1)-scale-invariant-ES in which the 'radius of the search', or *step-size*, is, at each iteration, proportional to the distance between the current solution and the optimum. We rigorously prove:

1. A log-linear convergence of the simplest ES, called scale-invariant $(1+1)$-ES, when minimizing spherical functions. A log-linear convergence means that the logarithm of the distance to the optimum converges linearly to $-\infty$ as a function of the number of iterations.

2. The optimality (regarding the convergence speed) of the $(1+1)$-ES algorithm using the artificial scale-invariant rule when minimizing spherical functions. Moreover, optimal convergence rates are numerically derived as a function of the search space dimension.

1.2 Optimization of noisy functions (A part of the results has been appeared in [76])

Noisy objective functions are important to study, as real objective functions are usually noisy. Noisy spherical functions investigated here are of particular interest as the randomness of their noisy part can cover a wide range of irregular real objective functions. We investigate the scale-invariant $(1+1)$-ES (Chapter 3) and the so-called scale-invariant $(1, \lambda)$-ES (Chapter 4) for the minimization of noisy objective functions. More precisely:

- For the scale-invariant (1+1)-ES, noisy objective functions are modelized as $||x||^2(1+\mathcal{N})$. The main result is that the convergence that has been already shown in [77] for non noisy objective functions does not always hold for noisy objective functions. If the noise level is such that negative objective functions values can be sampled with a striclty positive probability, the algorithm does not converge (if the noise is Gaussian) and diverges (if the noise is lower bounded). Furthermore, for noise distributions that only sample positive fitness values, the algorithm converges. We prove also that the same results hold for a more general class of noisy objective functions that can be written as $(||x||^2 + \alpha)(1 + \mathcal{N})$ with $\alpha > 0$. Our results are compared with those in [8], with which they seem contradictory. In this study, we also theoretically derive the convergence (or divergence) rates of the algorithm minimizing noisy objective functions written as $||x||^2(1 + \mathcal{N})$. Moreover, we show that the convergence (or divergence) rates can be computed numerically. For convergence (or divergence) rates which are not equal to zero, the behavior of the algorithm is log-linear.

- For the scale-invariant $(1, \lambda)$-ES, the noisy objective functions that are investigated can be written as $||x||(1 + \mathcal{N})$. The log-linear behavior (convergence/divergence) is theoretically proven. The convergence and divergence cases are distiguished as a function of the noise level and the so-called 'normalized step-size mutation' (a parameter of the algorithm), theoretically (when d goes to infinity) and numerically

(for $d < +\infty$). We show that convergence rates vary almost linearly with the inverse of the dimension of the search space. Moreover, we theoretically prove that the approximations used in [8] for the infinite dimension study are reliable.

2 Part 2: Application (Results in [78])

The application part of this thesis is presented in Chapter 7. We investigate the resolution of a real-world problem encountered by chemical engineers. The goal is the identification of the parameters of the isotherm function governing the chromatography process. The approach used in order to solve this problem is to turn the identification problem into an optimization problem. One of the difficulties of the resulting optimization problem is that the relative search space is implicitely constrained. The resulting parametric optimization problem is tackled using the state-of-the-art in Evolution Strategies, the so-called CMA-ES (Covariance Matrix Adaptation-Evolution Strategy) introduced by N. Hansen and A. Ostermeier [57, 59, 61]. The version of this algorithm used here is that of [16]. This identification problem had already been addressed using gradient-based approaches [74, 73]. We perform the identification using the real-world data set provided in [73]: this allows us to compare our results with those of the gradient based approach. The comparison reveals that our approach is more efficient than the numerical gradient approach. More precisely, 1) The CMA-ES algorithm always converges to the same solution, independently of the starting point: this was not the case for the gradient approach. 2) Better objective values can be found by CMA-ES for two different experimental configurations. In particular, CMA-ES is able to handle the full problem and identifiy the 6 parameters, whereas the gradient approach doesn't work unless the values of 2 of the parameters are manually fixed (to experimental values). Finally, both approaches have very similar computation times, which is a rather unusual finding, as it is well known that deterministic methods are generally much more faster than population based stochastic methods.

The last part of the document is a general conclusion that summarizes the results obtained, also giving perspectives of possible future work.

Chapter 1

Non linear continuous optimization

Optimization is a recurrent task mostly investigated by Engineers, Applied Mathematicians, and Computer Scientists. We are here interested in continuous minimization [2] problems that can be formally formulated as follows:

$$\begin{cases} \text{Minimize } f(x), \\ x \in \Omega \end{cases} \tag{1.1}$$

where $f : \Omega \mapsto \mathbb{R}$ is the *objective function* defined on some open subset Ω of $\mathbb{R}^d$ and is assumed to be non linear.
Real-world continuous optimization problems are everywhere. For instance, they include shape optimization of e.g. airfoils in aeronautic industry, model calibration frequently encountered in biological or physical domains, and parameter identification in the context of inverse problems.
This work focuses more particularly on the *black box scenario*, where the only available information about the objective function is the values it takes on any input from $\mathbb{R}^d$. In particular, no gradient nor Hessian information can be obtained (except of course through numerical computation from function values). Hence we will only consider zeroth order methods, that only use function values. In order to solve real-world continuous optimization problems, many iterative methods have been developed. These methods can be broadly classified in two categories, relatively to the method they use to explore the search space: Deterministic and randomized search methods.
In the following of this Chapter, we will briefly survey both deterministic and randomized search methods operating on unconstrained search spaces (i.e., $\Omega = \mathbb{R}^d$).

1.1 Deterministic search methods for non linear continuous optimization

The most widely used deterministic search methods have been reviewed in [108, 80, 85, 29] where convergence results are given. In the following, we give a presentation of some of

[2] Without loss of generality, the minimization of a real valued function f is equivalent to the maximization of $-f$.

the most popular deterministic search methods.

1.1.1 Gradient based methods

Gradient based methods refers to methods which use the explicit value of the gradient of the objective function at a given location. These methods have been originally inspired from the approximating Taylor formula of a sufficiently smooth function. They have been designed to work well at least on the convex quadratic functions. These methods are descent methods in the sense that the newly generated point at each iteration has always a better objective function value than the previous one. More precisely, let X_n be the solution at an iteration n. The new point X_{n+1} is generated as follows:

$$\begin{cases} \mathrm{X}_{n+1} = \mathrm{X}_n + t_n d_n, \\ t_n = \arg\min_{t \geq 0}\{f(\mathrm{X}_n + t d_n)\} \end{cases} \tag{1.2}$$

where d_n is the descent direction and t_n the descent step.

The descent step t_n is determined by some line search method (such as rules of Wolfe, Goldstein and Price, Armijo [29]).

A natural idea for the choice of the descent direction d_n is to choose the opposite of the gradient at the current location, i.e., $d_n = -\nabla(f(\mathrm{X}_n))$. However, better choices can be made: In the conjugate gradient methods, the successive descent directions d_n satisfy the recurrence relation $d_{n+1} = -\nabla(f(\mathrm{X}_{n+1})) + \frac{\|\nabla(f(\mathrm{X}_{n+1}))\|^2}{\|\nabla(f(\mathrm{X}_n))\|^2} d_n$. It is shown (see for example [103]) that the conjugate gradient method theoretically converges in at most d iterations when minimizing convex quadratic functions. Another choice for the descent direction d_n is $-\tilde{H}_n^{-1}\nabla(f(\mathrm{X}_n))$ where $\tilde{H}_n$ is (an approximation of) the Hessian matrix in the current solution. Gradient methods using such a descent direction are called Quasi-Newton Methods. The state of the art of these methods is the so-called Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS).
A drawback of all gradient based methods, however, is that they are local methods: because the objective function value decreases at each iteration, the search is stuck in the first encountered local optima.

Note that even in the black box scenario where no derivative information is available, it is useful to consider gradient based methods:

- If the objective function is smooth and its values can be computed with full precision, finite-differences can be used to obtain estimates of the derivative that are accurate enough to be used as gradients in a gradient based method, such as the implicit filtering method described in Section 1.1.2.

- Many popular search software (e.g., within Matlab) used numerical gradient, and it is hence mandatory to compare the results of any newly proposed optimization method to those of gradient-based methods, even if using numerical gradient, in order to assess their performances.

1.1.2 Deterministic direct search methods

Deterministic direct search methods first appeared in the 1950's and early 1960's with the growing use of computer to fit experimental data. The name direct search was introduced in 1961 by Hooke and Jeeves. These methods do not use the explicit expression of the gradient to generate new solutions. In the following, we present some widely used deterministic direct search methods.

Derivative-free pattern search methods

The direct Pattern Search algorithm of Hooke and Jeeves [65] is one of the earliest deterministic search methods that does not make use of derivatives. The generic pattern search algorithm [39] calculates objective function values of the current pattern and tries to find a minimizer. Let X_n denotes the solution at the iteration n. The Hooke and Jeeves algorithm is a member of the so-called Generalized Pattern Search algorithms (GPS) which seeks for a lower value of the objective function by sampling points in the search space in a fixed set (or *pattern*) around the current point. Sampled points build the set $\mathcal{L}_n$ which is defined as follows

$$\mathcal{L}_n = \{x \in \mathbb{R}^d \ ; \ x = \mathrm{X}_n \pm \Delta_n s^i e^i \ , \ i \in \{1, \ldots, d\}\} \tag{1.3}$$

where $\Delta_n > 0$ is the pattern size which represents the search step, e^i is the i^{th} unit vector, and $s = (s_1, \ldots, s_d) \in \mathbb{R}^d$ is a fixed parameter that can be used to take into account the different scales of the parameters to optimize. If the algorithm finds a new minimum, then it changes the center of the pattern and iterates. If all the values on the pattern fail to produce a decrease, then the search step or pattern size is reduced by half, i.e., $\Delta_{n+1} = \frac{\Delta_n}{2}$. The search continues until the search step Δ_n gets sufficiently small, thus ensuring convergence to a local minimum. Performance is increased by reusing pattern values as the pattern center moves. Convergence analysis of GPS algorithm minimizing smooth objective functions have been performed by Torczon [139] and Audet and Dennis [11].

Simplex methods

The first simplex based direct search method was proposed by Spendley, Hext and Himsworth in 1962 [129]. In 1965, the original method was developed by Nelder and Mead [105]. The method evolves a convex hull of $d + 1$ points in $\mathbb{R}^d$, where the points satisfy the non-degeneracy condition that the volume of the hull is nonzero. At every iteration, the worst vertex is replaced by a new vertex using reflection, expansion or contraction. In the case where this fails, a shrink step is carried out. Thus, this method only ensures improvement of the objective function value in the sequence of worst vertexes, but it is the sequence of best vertexes that ultimately is of interest.
It has been theoretically and numerically shown that the Nelder-Mead simplex algorithm can fail in practice. Mc Kinnon [96] constructed a family of strictly convex objective functions in $\mathbb{R}^2$ for which he demonstrated that the Nelder-Mead algorithm fails to converge to a stationary point, i.e., on which the gradient equals 0. In Mckinnon examples, simplexes

converge to a straight line that it is orthogonal to the steepest descent direction. In [146], there is a discussion of the limitations, disadvantages, successes and developments of the Nelder-Mead algorithm.

To overcome the shortcomings of the Nelder-Mead algorithm, Torczon [138] proposed the so-called multi-directional search, which is also a simplex-based strategy. It has the property that shrinks occur for any number of variables, provided that the level sets of the objective function are bounded. In [138], Torczon gives a convergence proof for the mutli-directional search and performs empirical tests including the multi-directional search, Nelder and Mead algorithm and a quasi-Newton method. She showed that the multi-directional search is robust whereas the Nelder-Mead algorithm is not, and that multi-directional search can handle higher dimension problems and claimed that the multi-directional search may be useful for optimizing noisy objective functions. However, the multi-directional search also has some limits. In fact, the empirical study that has been performed in [9] demonstrates that the performance of the multi-directional search markedly degrades with increasing search space dimensions, and it is stated that, in the presence of noise, "...the multi-directional search method never stagnates but rather diverges if the noise strength is too high".

Quadratic approximation methods

These methods rely on an interpolation or an approximation of the objective function with a quadratic function Q. The approximation is supposed to be reliable on a region of the search space called the *trust region*. A quadratic function Q has $\tilde{d} = \frac{1}{2}(d+1)(d+2)$ independent coefficients, that may be defined by the interpolation conditions on $\tilde{d}$ points of $\mathbb{R}^d$:

$$Q(x^i) = f(x^i),\ i = 1, \ldots, \tilde{d} \tag{1.4}$$

The points x^i should have the property that, if Eq. 1.4 is written as a system of linear equations in terms of the coefficients Q, the matrix of the system should be non singular.

Winfield's algorithm [144] not only employs the interpolation equation Eq. 1.4 to define Q, but also includes some of earliest work on trust regions. At an iteration n, the algorithm generates the quadratic approximation Q_n using Eq. 1.4. Furthermore, the iteration computes the vector $\underline{x} \in \mathbb{R}^d$ that minimizes Q_n subject to the bound $\|x - \mathrm{X}_n\| \leq \rho_n$ where X_n is the best point among the interpolation points at iteration n, and ρ_n is the trust region radius. This algorithm presents the particularity that, an eventual degeneration of the system Eq. 1.4 is ignored and it is assumed that the calculation of Q_n is sufficiently robust to provide a quadratic function that allow the trust region sub problem to be solved and the resulting $\underline{x}$ receives no special treatment. Other methods ensure that Q_n is well defined. Powell [108] stated that Lagrange functions are highly useful for selecting the interpolations points at each iteration such that the quadratic polynomial Q_n is well defined by Eq. 1.4. Using this idea, Powell proposed in 2002 the NEW Unconstrained Optimization Algorithm (NEWUOA) algorithm as a quadratic interpolation method that uses only $\tilde{d} = 2d + 1$ to build the quadratic function Q. Therefore, the amount of work per iteration is only of order $(3d+1)^2$, which allows d to be quite large. The success of the method is, according to Powell [109], due to the use of the symmetric Brodyen method for

updating the Hessian of Q_n, $\mathrm{H}(Q_n)$, when first derivative of f are available [40]. Another claimed advantage [110] of the NEWUOA is that is suitable for the minimization of noisy objective functions.
The algorithm can be summarized as follows. First, an initial quadratic model Q_0 is created for the objective function f. An iteration n then performs the following steps:

- Compute the minimum of Q_n inside the trust region,
- Improve the model using the latest optimum,
- Stop if the latest trust region is lower than the user-defined end value,
- Stop if the distance between Q_n and f is small enough (perfect match of the model Q_n and the objective function f),
- Decrease the trust-region radius if the values computed for f stops decreasing.

A more detailed presentation of the algorithm can be found in [110].

Implicit filtering

Implicit Filtering, as devised by Gilmore and Kelley [47, 80], belongs to the so-called Stochastic Approximation methods dating back to work of Robins and Monroe [115] and Kiefer and Wolfowitz [83] and which were specifically designed to deal with noisy objective functions. In contrast with the direct deterministic search methods introduced so far, Implicit Filtering relies on the idea suggested by Kiefer and Wolfowitz of explicitly approximating the local gradient of the objective function by means of finite differencing. Because the gradient is only an approximation, the computed steepest descent direction may fail to be a descent direction and the line search may fail. In this case, the difference increment used to numerically compute the gradient is reduced. The name "implicit filtering" has been chosen because the method uses differencing to "step over" the noise at varying levels of resolution, hence implicitly filtering the objective function from the noise. The method uses the central difference gradient that we denote $\nabla_h f$ in a gradient based method. Let x a point in $\mathbb{R}^d$, and h a difference increment, a central difference gradient is defined as follows:

$$(\nabla_h f(x))^i = \frac{f(x + he^i) - f(x - he^i)}{2h}, \ i = 1, \ldots, d \tag{1.5}$$

where e^i is the i^{th} unit vector. Clearly this computation involves $2d$ evaluations. At iteration n, the algorithm computes the central difference gradient at the current solution X_n, i.e., $\nabla_h f(\mathrm{X}_n)$. As in gradient based methods, the new point X_{n+1} is generated as $\mathrm{X}_n + t_n d_n$, where t_n is determined by a standard line search in direction d_n. The descent direction d_n is usually generated as in Quasi-Newton methods. A presentation of the convergence theory of implicit filtering and of several related methods can be found in [80].

1.2 Randomized search methods for non linear continuous optimization

Randomization is an efficient research tool for seeking the optima of an objective function especially when no information about he derivative neither the Hessian of this function are provided. Randomized search methods are global search methods in the sense that the stochastic nature of the search can prevent the convergence to a local optimum[3]. Their ability to escape local optima is also due to the fact that they are usually population based. However, despite their practical ability to solve many real-world otpimization problems, the majority of these methods do not rely on a firm mathematical background: they are in general designed based on nature-inspired paradigms, and their theoretical study comes long after their effective use and successes in practical applications. This section will survey the most widely used randomized search methods.

1.2.1 Pure Random Search (PRS) and Pure Adaptive Search (PAS)

Pure Random Search (PRS) [31] is the simplest random search method. This method consists in generating the solutions X_1,...,X_n independently, using a fixed probability distribution. When the stopping criterion is met, the best point reached so far is taken as the solution proposed by the method. It has been theoretically proven [149] that PRS converges to the global minimum with probability 1 for every objective function for which the neighborhood of the optimum can be reached with a strictly positive probability. However, the search is always done around the same fixed point and the search distribution parameters, namely the radius and the covariance matrix of the search in case of continuous optimization, are kept unchanged during the run. Therefore, these parameters are not adapted, neither relatively to the history of the search, nor to the local shape of the function to optimize. This makes PRS totally inefficient in practice, with a very large convergence time that increases exponentially with the search space dimension [149].

Then Pure Adaptive Search (PAS) was introduced as a random search method having an exponentially lower complexity than that of PRS [148]. In fact, the convergence time of PAS varies linearly with the search space dimension d in the specific case of Lipschitz objective functions. This method differs from the PRS method in the fact that the new individual is uniformly generated on the set containing individuals having better objective function values than the current solution. Therefore PAS is not practical because the principal computational effort of the algorithm lies in generating points uniformly distributed in the improving region. Moreover, PAS can be seen as a particular instance of an Evolution Strategy (ES) (see Section 1.2.4 for a presentation of ES) evolving a unique solution and where no adaptation in the search parameters is done.

[3]However, the probability to escape a local optimum can be too small when using some randomized search method such as the $(1+1)$-ES for example (which will be described in Section 1.2.4).

1.2.2 Simulated Annealing (SA)

Simulated Annealing (SA) [84, 3] is a global optimization method inspired from annealing in metallurgy. The optimization method considers each point x of the search space as a state of some physical system, and the objective function value of x, $f(x)$, as the energy of the state x. The goal is then to bring the system, from an arbitrary initial state, to a state with the minimum possible energy – that is, to minimize the objective function f. The algorithm generates a sequence of solutions (X_n) as follows. Let X_n be the solution at iteration n. A new solution Y_n is generated using a search space distribution depending on X_n. The acceptance rule of the new point Y_n is the Boltzmann rule, defined as follows

$$\mathrm{X}_{n+1} = \begin{cases} Y_n \text{ if } f(Y_n) \leq f(\mathrm{X}_n), \\ Y_n \text{ if } f(Y_n) > f(\mathrm{X}_n) \text{ with probability } e^{\left(\frac{f(\mathrm{X}_n)-f(Y_n)}{T_n}\right)}, \\ \mathrm{X}_n \text{ otherwise} \end{cases} \tag{1.6}$$

where T_n, the so-called *temperature*, is a positive parameter that will be decreased to 0. The goal of the randomization in the Boltzmann acceptance rule for the new solution Y_n is to avoid getting stuck in local optima. In practice, the sequence (T_n) has to be a decreasing sequence such that the probability to accept worse solutions decreases during the run. The convergence (in probability) results [84, 102] only require that the temperature sequence (T_n) decreases to 0 and, in some cases, that this sequence decreases slowly enough in order to escape local optima. In [102] too, key concepts such as global versus local exploration and adaptability of the parameters of the search distribution and of the acceptance probability have been underlined.

In practice, however, the major inconvenient of SA methods, and especially of one of the most popular one, the so-called Adaptive Simulated Annealing (ASA) [68, 69], is the tuning of its underlying parameters. It is worth noticing that SA can be seen as a particular ES method (see paragraph 1.2.4 for a presentation of ES) evolving a single solution and using Boltzmann randomized rule for the acceptance of a new point. Along those lines, the methods discussed in [37] for the adaptation of the parameters of the search distribution are quite similar to that of the so-called derandomized ES (see Section 1.3).

1.2.3 Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) [81, 126, 127, 34] is a population-based stochastic optimization technique initially proposed by R. Eberhart and J. Kennedy in 1995, inspired by social behavior of bird flocking or fish schooling. PSO tracks a number of so-called particles (solutions vectors) in a swarm. The default swarm size is $S = 10 + [2\sqrt{d}]$. At each iteration n, let $\mathrm{X}_n = (\mathrm{X}_n{}^1, \ldots \mathrm{X}_n{}^d)$ denotes a particle of the swarm. This particle is characterized by:

- a velocity V_n (that can also be viewed as the previous displacement of this particle, i.e., $\mathrm{X}_n - \mathrm{X}_{n-1}$),
- the best solution encountered so far by that particle, denoted pbest_n i.e., $\mathrm{pbest}_n \in \{\mathrm{X}_0, \ldots, \mathrm{X}_n\}$ with $f(\mathrm{pbest}_n) \leq f(X_j),\ \forall j \in \{0, \ldots, n\}$, and

- the global best position ever visited by all particles that we will denote gbest_n.

The particle X_n is then pulled toward the best positions pbest_n and gbest_n as follows

$$\begin{aligned} V_{n+1}^j =& wV_n^j + \alpha^j \left[\text{pbest}_n^j - X_n^j\right] + \beta^j \left[\text{gbest}_n^j - X_n^j\right] , \\ \text{X}_{n+1}{}^j =& \text{X}_n{}^j + V_{n+1}^j \end{aligned} \tag{1.7}$$

for each coordinate $j = 1, \ldots, d$, where α^j and β^j are uniformly distributed in $[0, \phi]$ with $\phi = \ln(2) + \frac{1}{2}$ and the inertia weight w equals $\frac{1}{2\ln(2)}$. One of the reasons of the widespread use of PSO is that it is very easy to program (no linear algebra involved for instance), and there are very few parameters to adjust. Indeed, in the recent years, PSO has been applied in many research and application areas [35, 28, 32, 94, 95]. Unfortunately, in a recent study [62] investigating the performances of the Standard PSO 2006 [1] on ill-conditioned functions, it has been demonstrated that, whereas PSO performs very well on separable functions (even if ill-conditioned), its performance degrades dramatically on non-separable ill-conditioned functions.

1.2.4 Evolutionary Algorithms

Evolutionary Algorithms (EAs) are bio-inspired optimization methods which evolve a population of solutions. They are an iterative technique inspired by Darwin's theory of natural evolution, more precisely the idea that the emergence of species that are adapted to their environment results from the synergy between *natural selection* (survival of the fittest) and *blind variations* (random modification of the genetic material from parents to offspring, independently of any adaptation). The denomination of the ingredients of the algorithm also arise from the biological paradigm: the objective function is usually called the *fitness*, the points of the search space, possible solutions of the problem at hand, are called *individuals*, and the set of individuals that the algorithms evolves is termed a *population*. A *generation* (one iteration of the algorithm) consists in

1. Selecting among the population at current time n (also termed the *parents*) some individuals based on their fitnesses, biased toward individuals with good values with respect to the optimization problem at hand (i.e., implementing a first step of 'natural' selection);

2. Applying *variation operators* (i.e., stochastic operators independent of the objective function) to the selected parents, thus generating *offspring*. The variation operators are either unary operators (also called *mutations*), or k-ary operators (then called *recombination* or *crossover* operators);

3. Evaluating the offspring, i.e., computing the value of the objective (fitness) function at the newly generated points, the offspring;

4. Selecting among the offspring and the 'old' parents, based again on fitness values, the individuals who will survive to the next generation, thus implementing the second step of 'natural' selection.

From the description above, it is clear that Evolutionary Algorithms are zeroth order methods. Moreover, they have been applied successfully to solve many real-world problems [51, 43, 22, 104, 142]. However, their main drawback is that they are computationally costly, requiring in general a large number of generations and rather large population sizes. Moreover, another difficulty comes along with their high flexibility: when tailoring these methods to a new problem, the user has to set a high number of parameters. A promising line of research in to cope with this difficulty while maintaining the high flexibility of those algorithm is to make as many of those parameters as possible *adaptive*, i.e., automatically determined during the course of evolution. In the specific field of continuous optimization, many adaptive methods have been developed, and will be detailed in the forthcoming Section 1.2.4.

Historical roots

Before turning to the detailed description of Evolution Strategies, the Evolutionary Algorithm at the heart of this thesis, it is worth describing shortly other roots of the field that have also been applied to continuous optimization.

Genetic Algorithms (GAs) are still the most popular field of Evolutionary Algorithms. GAs has been investigated since the early sixties by J. Holland [64]. GAs were initially designed to handle bit-string representation, but were also used for continuous optimization problems by representing each real number by its 'natural' binary representation. However, such representation have some sever drawbacks. In particular, it does not respect at all the topology in $\mathbb{R}^d$, as thoroughly discussed in [135]. Today, with very few exceptions, bit-string representations are abandoned when dealing with continuous parameters, at least when accuracy matters. Hence GAs will not be discussed any more here. For more details, see the seminal book by Goldberg [49], or the more recent and comprehensive books [101, 143]. One of the earliest book about optimization by means of natural evolution is that of L. Fogel [45], introducing what has been known as **Evolutionary Programming** (EP). Initially devoted to the optimization of Finite State Automata, Evolutionary Programming was successfully applied to very diverse search spaces, including continuous ones. However, in that particular setting, EP can also be considered as a particular case of self-adaptive Evolution Strategies (see next Section), and is not an active field per se any more. It is hence only recalled here to account for the historical truth.

Modern EAs tend to forget the frontiers between those historical dialects, as advocated by Michalewicz [99] and De Jong [38], and presented in the recent textbook by Eiben and Smith [41]. The remaining differences regard the representation: Genetic Algorithms are associated with bit-strings, Genetic Programming with parse-tree, and Evolution Strategies with real-valued parameters: they are the background of this work, and will now be introduced in detail.

Evolution Strategies

Evolution Strategies (ESs) have been introduced by I. Rechenberg [114, 113] and H.P. Schweffel [123] in Germany, also in the mid-sixties. For historical reasons, specific no-

tations are used, that will be defined here. For instance, the population size is denoted $\mu \in \mathbb{N}$, and the number of generated offspring $\lambda \in \mathbb{N}$.

ESs instantiate the generic EA given above the following way:

1. There is no parent selection step per se: all μ parents are chosen with uniform probability to generate offspring
2. the main variation operator is the *Gaussian mutation* (see below); recombination, also called here *intermediate crossover*, is achieved by performing a linear combination of two or more parents (though in the original ES, no recombination was used);
3. All offspring are evaluated normally;
4. The survival selection is deterministic: the μ best individuals are chosen either among the λ offspring – and the algorithm is then called a $(\mu, \lambda) - ES$ – or among the $\mu + \lambda$ parents plus offspring, and the algorithm is then a randomized hill-climber termed a $(\mu + \lambda) - ES$.

The main operator of ESs is the Gaussian mutation: a parent X generates an offspring Y by Gaussian mutation which will be written as

$$Y = X + \sigma \mathrm{N}(0, C). \tag{1.8}$$

where $\sigma \mathrm{N}(0, C) = \mathrm{N}(0, \sigma^2 C)$ is a drawn according to the multivariate normal distribution of mean 0 and covariance matrix $\sigma^2 C$. The reason for separating the *step-size* σ from the *covariance matrix* C lies in the adaptation mechanisms that will be described later (Section 1.3): this will to separately adapt the average length of the mutation by modifying the step-size σ and the main directions of the mutation by modifying the covariance matrix C.

However, those parameters (σ and C) should be adapted along evolution to the current *fitness landscape*, that is the local characteristics of the objective function.

Adaptation in ES

As said above, parameter control (also termed on-line parameter tuning) is a general issue in Revolutionary Algorithms [42]. In the particular case of Evolution Strategies, it has received a lot of attention since the very early works in the 60's.

The 1/5 adaptation rule is the oldest known adaptation rule [121, 114]. This rule adapts a single step-size for the whole population (and used the Identity matrix as Covariance Matrix). Its mechanism is to compute the empirical success probability over the last generations and to increase the step-size mutation ($\sigma_{n+1} = \sigma_n e^{\frac{1}{3}}$) if this success probability exceeds 0.2 (or to decrease the step-size ($\sigma_{n+1} = \sigma_n / e^{\frac{1}{3}}$) if the empirical success probability is below 0.2). This rule was derived after a theoretical study on two simple objective functions (the sphere function, and the corridor function, a linear constrained function), and asymptotically when the space dimension d tends to $+\infty$. Whereas it was shown to be quite efficient on many functions, it can be totally wrong when the fitness

function does not behave like the model functions. Moreover, it does not adapt the covariance matrix of the search distribution.

The self-adaptation rules were introduced by Schwefel in the seventies [124]. *Self-Adaptive ESs* (or SA-ESs) use the evolution itself to adjust the mutation parameters. The basic idea is to associate to each individual its own mutation parameters. One mutation then amounts to first mutate the individual's mutation parameters, then to mutate the individual itself using the new values of the mutation parameters. In the long run, even though the selection only acts based on fitness values, only individuals with 'good' mutation parameters (i.e., adapted to the local characteristics of the fitness) can survive many selection steps. It is sometimes said that the mutation parameters are updated 'for free'. There are 3 variants of this technique, depending on the number of mutation parameters that evolve.

In the *isotropic SA-ES*, only one mutation step-size is considered per individual, and the covariance matrix is kept equal to I_d. The step-size undergoes a log-normal mutation (in order to keep it positive, and because it is then used multiplicatively in the Gaussian mutation):

$$\sigma := \sigma \exp \tau \tilde{\mathrm{N}}(0,1) \tag{1.9}$$

were τ is a strictly positive parameter and $\tilde{\mathrm{N}}(0,1)$ is a sampling of a normal distribution with mean 0 and standard deviation 1. The parent is then mutated using the usual Gaussian mutation with step-size σ:

$$Y = X + \sigma \mathrm{N}(0, I_d) \tag{1.10}$$

Note that considering the pairs (X, σ) of individuals together with their mutation step-size, the complete mutation can also be written as

$$(X, \sigma) \to (X + \sigma \exp \tau \tilde{\mathrm{N}}(0,1) \mathrm{N}(0, I_d), \sigma \exp \tau \tilde{\mathrm{N}}(0,1))$$

In the *non-isotropic SA-ES*, the covariance matrix is a diagonal matrix with positive coefficients denoted $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$. The mutation of the deviations σ_i's proceeds as follows

$$\sigma_i := \sigma_i \exp \tau' \mathrm{N}(0,1) \exp \tau \mathrm{N}_i(0,1) \text{ for } 1 \leq i \leq d \tag{1.11}$$

where $\mathrm{N}(0,1)$ and $\mathrm{N}_i(0,1)$ $(1 \leq i \leq d)$ are $d+1$ independent samplings of a centered reduced normal random variable. Then, each coordinate of a parent is mutated using the corresponding mutated step-size in the same direction, giving offspring Y as follows

$$Y_i = X_i + \sigma_i \mathrm{N}(0,1) \text{ for } 1 \leq i \leq d. \tag{1.12}$$

Note that there is no global step-size here, but that the log-normal mutation of all σ_i's has a first term that is common to all i's, and thus can be seen as some global update, plus a term that is specific to each coordinate i.

Finally, the *Correlated SA-ES* uses a full covariance matrix (i.e., not restricted to a diagonal matrix) in order to also adapt to the principal directions of the objective function.

In order to easily mutate this covariance matrix, it is written as the product of $d(d-1)/2$ 2D-rotation matrices $R(\alpha_{ij})$ with $1 \leq i < j \leq d$ and a diagonal matrix D with diagonal coefficients $\sigma_1^2, \ldots, \sigma_d^2$.

$$C = \left(\prod_{i=1}^{d-1} \prod_{j=i+1}^{d} R(\alpha_{ij}) \right) D \tag{1.13}$$

The mutation of the covariance matrix consists first in a log-normal mutation of the coefficients of the diagonal matrix D, as in the non-isotropic case (see Eq. 1.11). Then the angles α_{ij} $(1 \leq i < j \leq d)$ are also mutated using independent samplings of a Gaussian variable $\beta \mathrm{N}(0,1)$ (for a user-defined β). Finally, the parent X is mutated by a Gaussian mutation of covariance matrix the mutated C.

In [54], it has been shown that the different variants of SA-ES are not coordinate-independent, i.e., will behave differently if a (linear) change of coordinate is done in the search space (though the function stays the same). Moreover, the use of a randomized self-adaptation rule implies a low correlation between the mutation step-size and the distance between the new accepted offspring and its parent i.e., $\|\mathrm{X}_{n+1} - \mathrm{X}_n\|$ [124]. Those remarks have lead to different attempts to completely derandomize the SA-ES algorithm.

Recombination operator Though the initial ES algorithm didn't use any recombination, it has been shown that the performances of ESs are increased if a recombination operator is used [123, 147]. Furthermore, [25] shows a qualitative improved progress when a global intermediate recombination of μ parents is used rather than a $(1, \lambda)$-ES.

Toward completely derandomized ES These ideas has been exploited to design new ES algorithms with recombination and a derandomization of the adaptation rule of the search distribution parameters. The most advanced ES using these techniques is the so-called Covariance Matrix Adaptation Evolution Strategy (CMA-ES) introduced by N. Hansen and A. Ostermeier in 1996 [61, 59, 57, 16]. This method uses a completely derandomized self-adaptation using the cumulation of previous step-size and covariance matrix moves. The adaptation of the covariance matrix used in CMA-ES allows he algorithm to be invariant by change of coordinates. Moreover, the algorithm generates a sequence of covariance matrices C_n which is observed to converge to the inverse of the Hessian in the case of quadratic convex objective functions. Compared to other ES, CMA-ES has been shown to exhibit similar behavior on perfectly scaled objective functions, and to perform better on ill-conditioned non separable objective functions [61]. CMA-ES is also performing well on multi-modal functions [56]. The importance of CMA-ES nowadays justifies that it is be presented in detail in a stand-alone forthcoming Section 1.3.

1.2.5 Differential Evolution

Differential Evolution (DE) was introduced by Price and Storn [131, 132, 133], and can be viewed as a particular Evolutionary Algorithm for continuous optimization: DE evolves a

population of individuals $X_1, \ldots, X_\mu$ using a very specific mutation operator, that adds, at each iteration, to a given individual one (or many) *difference vector(s)* between one (or many) couple(s) of other individuals in the population (hence the name of the algorithm). A crossover operator is then performed between the mutated vector and the parent, and finally the offspring replaces its parent if it has a better fitness. There are several strategies for DE that differ in the way mutation and crossover is conducted [130, 111, 98] (the latter reference is a compartive study between some variants of DE). The variants are specified using the notation DE/x/y/z where x denotes the way the vector to mutate will be chosen (randomly or the best one for example), y denotes the number of difference vectors to add to the mutated vector and z denotes the crossover sheme (binomial or exponential for example). In the classical variant of DE, the DE/rand/1/bin, the mutation and crossover write as:

1. **Mutation** For each parent X_i, $i = 1, \ldots, \mu$, the following mutating vector is created

$$M_i = X_{r_1} + F(X_{r_2} - X_{r_3}),$$

   where r_1, r_2 and r_3 are indices that are uniformly chosen in $\{1, \ldots, \mu\}$, and where F is a user-defined amplifying factor in $[0, 2]$.

2. **Crossover** First an integer j_0 is uniformly chosen in $\{1, \ldots, d\}$. Then, a uniform crossover between X_i and M_i is performed:

$$Y_i^j = \begin{cases} M_i^{j_0} \text{ if } j = j_0 \\ M_i^j \text{with probability } CR \text{ if } j \neq j_0, \\ X_i^j \text{with probability } (1 - CR) \text{ if } j \neq j_0 \end{cases} \tag{1.14}$$

   where $CR \in [0, 1]$ is a user-defined Crossover Rate.

Then the offspring Y_i replaces its parent X_i iff it has a better fitness. In [131], Price and Storn have shown on some test functions that DE is superior to Adaptive Simulated Annealing (ASA) (see Section 1.2.2). The DE algorithm is rotationally invariant when the crossover rate CR equals 1, whereas the behavior of the algorithm is not invariant to search space rotation if $CR \neq 1$ [111, p. 98]. Note also that the performance of DE is sensitive to its control parameters [46] and that the DE is not only prone to premature convergence but also to stagnation [88] and that a successful location of the global optimum depends on choosing the correct control parameters. Finally, the recommended population size for DE is $10d$, and the performance of the algorithm hence poorly scales up with d.

1.2.6 Estimation of Distribution Algorithms

The first instance of an Estimation of Distribution Algorithm (EDA) is the PBIL algorithm (Population Based Incremental Learning) that has first been proposed as an alternative to Genetic Algorithms in the bit-string framework [21]. EDAs try to identify a probability distribution defined on the search space by successively sampling the current distribution, computing the fitness of the sampled points, selecting some of the

sampled point with a bias toward the best performing points, and either reconstructing a probability distribution from the selected points, or updating the current distribution using those sample points.

EDAs have been applied to continuous optimization, starting with a modified PBIL algorithm [125] that was using ... a Gaussian distribution on the real-valued search space. Several variants have then been proposed (see e.g. [91] for a survey), and all of them evolve a full multivariate normal distribution by modifying its mean and covariance matrix along evolution. This is exactly what a fully derandomized Evolution Strategy like CMA-ES is doing (see next Section). In particular, the Estimation of Multivariate Normal Algorithm (EMNA) [91] uses an update mechanisms that is very similar to that of CMA-ES, though it reconstructs the covariance matrix from the selected sample points while CMA-ES carefully updates the current covariance matrix. Experimental results [58] have demonstrated that CMA-ES takes advantage of this update and performs better than EMNA even on multi-modal test functions.

1.3 Covariance Matrix Adaptation-Evolution Strategy

Though it clearly belongs to the Evolution Strategy family of stochastic search algorithms, the Covariance Matrix Adaptation-Evolution Strategy (CMA-ES) is presented in a separate Section in order to emphasize its importance – as will be witnessed by the empirical comparisons presented in next Section.

CMA-ES was introduced by N. Hansen and A. Ostermeier in 1996 [60] and the complete almost parameter-less algorithm was published in 2001 [61]. It is a $(\mu, \lambda) - ES$ that uses a global recombination operator involving the μ parents at each iteration, and hence is referred to as a $(\mu/\mu, \lambda)$-ES. Let X_n denotes the recombination of the parents at iteration n[4]. This 'super-parent' is subject to λ independent mutations, resulting in λ offspring $Y_1, \ldots, Y_\lambda$:

$$Y_k = \mathrm{X}_n + \sigma_n \mathrm{N}_k\left(0, C_n\right) \text{ for } k = 1, \ldots, \lambda$$

The new super-parent X_{n+1} is the computed as a linear combination of the best μ offspring:

$$\mathrm{X}_{n+1} = \sum_{i=1}^{\mu} w_i Y_{i:\lambda} \ , \tag{1.15}$$

where the positive weights $w_i \in \mathbb{R}$ are set according to individual ranks and sum to one, and the index $i:\lambda$ denotes the i-th best offspring. The use of the weighted recombination of the parents as shown in Eq. 1.15 allows CMA-ES (and in general any $(\mu/\mu, \lambda)$-ES) to have a larger progress (at each iteration) than any $(1, \lambda)$-ES in the absence of noise [25].

Moreover, because it only uses an ordering of the λ offspring, CMA-ES is invariant by any monotonous transformation of the fitness function (see Section 1.4.2). In particular, (non-)convexity does not modify in any way the behavior of CMA-ES.

[4]Note that in the presentation of CMA-ES in Chapter 5, the ieration number, here n, is referred to as g. In the same chapter, the quantities X_n, Y_k, σ_n, C_n, $(\vec{p}_c)_n$ and $(\vec{p}_\sigma)_n$ are respectively referred to as $\langle \vec{x} \rangle_{\mathrm{W}}^{(g)}$, $\vec{x}_k^{(g+1)}$, $\sigma^{(g)}$, $\boldsymbol{C}^{(g)}$, $\vec{p}_c^{(g)}$ and $\vec{p}_\sigma^{(g)}$. Note also that Equations1.17 and 1.19 for the covariance matrix adaptation are more general than those of Chapter 5.

Adaptation in CMA-ES It is stated in [16] that the adaptation used in CMA-ES allows to achieve, on convex-quadratic functions, log-linear convergence (see Definition 1.1 in Section 1.5) after an adaptation time which scales between 0 and the square of the dimension of the search space. This adaptation is done deterministically and the basic idea is to increase the probability to reproduce good steps. This is done by computing the so-called evolution paths for both the step-size and the covariance matrix. Let C_n denote the covariance matrix at an iteration n and $B_n D_n D_n \left(B_n\right)^T$ its decomposition in the eigenvector basis (B_n is an orthogonal matrix and D_n a diagonal matrix whose diagonal contains the square roots of the eigenvalues of ${C_n}^5$). Let $(\vec{p}_\sigma)_n$ and $(\vec{p}_c)_n$ be the evolution paths of respectively the step-size mutation and the covariance matrix. The adaptation is done as follows: First, the cumulative path for the step-size mutation is updated:

$$(\vec{p}_\sigma)_{n+1} = (1 - c_\sigma)(\vec{p}_\sigma)_n + \frac{\sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}}}}{\sigma_n} \times B_n (D_n)^{-1} {B_n}^T \left(\mathrm{X}_{n+1} - \mathrm{X}_n\right) \tag{1.16}$$

where c_σ is a parameter in $]0, 1]$. Then, the evolution path for the covariance matrix is in turn updated as follows:

$$(\vec{p}_c)_{n+1} = (1 - c_c)(\vec{p}_c)_n + (\mathrm{H}_\sigma)_{n+1} \frac{\sqrt{c_c(2 - c_c)\mu_{\text{eff}}}}{\sigma_n} \left(\mathrm{X}_{n+1} - \mathrm{X}_n\right) \tag{1.17}$$

where $(\mathrm{H}_\sigma)_{n+1} = 1$ if $\frac{\|(\vec{p}_\sigma)_{n+1}\|}{\sqrt{1-(1-c_\sigma)^{2(n+1)}}} < (1.5 + \frac{1}{d-0.5})E(\| \mathcal{N}(0, I_d) \|)$, and 0 otherwise, $c_c \in]0, 1]$ is the cumulation coefficient and μ_{eff} is a strictly positive coefficient which denotes 'the "variance effective selection mass". It can be seen from Eq. 1.16 and Eq. 1.17 that the evolution path updates take into account the last move $(\mathrm{X}_{n+1} - \mathrm{X}_n)$ and the history of the search which is represented by $(\vec{p}_c)_n$ for the evolution of the search directions, and $(\vec{p}_\sigma)_n$ for the evolution of the radius of the search. Finally, the mutation step-size and the covariance matrix are updated using information on the whole search history as follows:

$$\sigma_{n+1} = \sigma_n \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\| (\vec{p}_\sigma)_{n+1} \|}{E(\| \mathcal{N}(0, I_d) \|)} - 1\right)\right) \tag{1.18}$$

where $d_\sigma > 0$ is a damping factor and $\mathcal{N}(0, I_d)$ is the multivariate normal distribution with covariance matrix identity. For the covariance matrix, the update takes place as follows:

$$\begin{aligned} C_{n+1} = (1 - c_{\text{cov}})C_n + c_{\text{cov}} \frac{1}{\mu_{\text{cov}}} (\vec{p}_c)_{n+1} \left((\vec{p}_c)_{n+1}\right)^T \\ + c_{\text{cov}} \left(1 - \frac{1}{\mu_{\text{cov}}}\right) \sum_{i=1}^{\mu} \frac{w_i}{\sigma_n^2} (Y_{i:\lambda} - \mathrm{X}_n)(Y_{i:\lambda} - \mathrm{X}_n)^T \end{aligned} \tag{1.19}$$

where $c_{\text{cov}}, \mu_{\text{cov}} \in]0, 1[$. This update rule is called the rank-mu update for C_n [59]. When $\mu_{\text{cov}} = 1$, this rule reduces to the so-called rank-one update [61].

[5] Such a decomposition is always possible as C_n is positive definite symmetric matrix.

On the practical side, the default parameters of CMA-ES were carefully tuned in [57]. For example, the default values for λ and μ are respectively $\lambda^{def} = \lfloor 4 + 3\log(d) \rfloor$ and $\mu^{def} = \lfloor \frac{\lambda^{def}}{2} \rfloor$. Moreover, a 'restart' version of CMA-ES has been introduced in [16] in order to increase the probability to converge towards the global optimum when minimizing multi-modal objective functions. In this method, the algorithm is restarted with an increased population size when some restart criteria are met, indicating that the search process is no more progressing. Different restart criteria are used:

1. *RestartTolFun*: Stop if the range of the best objective function values of the recent generations is below than a TolFun value.
2. *RestartTolX*: Stop if the standard deviation of the normal distribution is smaller than a TolX value and $\sigma\vec{p}_c$ is smaller than TolX in all components.
3. *RestartOnNoEffectAxis*: Stop if adding a 0.1 standard deviation vector in a principal axis direction of C_n does not change X_n.
4. *RestartCondCov*: Stop if the condition number of the covariance matrix exceeds a fixed value.

The resulting version of CMA-ES is a quasi parameter free algorithm. This version of CMA-ES performed best at the CEC 2005 Special Session on Continuous Optimization [2].

CMA-ES has also been applied to a variety of real-world optimization problems [53]. For more details about CMA-ES, we refer to [52].

1.4 Comparison of continuous optimization methods

The difficulties of real-world optimization problems can be characterized by several different features. In addition to difficulties due to the search space, such as high dimension and constraints, real-world problems difficulties are generally related to the characteristics of the objective function.

1.4.1 Objective functions

Let us first list several properties of objective functions that can be the source of difficulties for their optimization. Objective functions can be

- non-convex: The hypothesis of convexity is the basis of the gradient based methods, that were designed to have good performances at least on quadratic convex functions. The non-convexity of the objective functions is hence an obstacle for methods relying on quadratic approximation such as Conjugate Gradient, BFGS, and Implicit Filtering.
- rugged: Most convergence results that have been proved for optimization methods (especially deterministic methods) require some regularity of the objective functions. Hence those methods might fail on rugged functions.

- noisy: Noisy objective functions arise in most real-world problems, and high values of noise can totally mislead the search. For instance, numerically-computed gradients become totally unreliable in the presence of noise. But because the ranking of candidate solutions can be hindered by the noise, search methods using rank information can also be deceived by noisy functions.

- multi-modal: Some objective functions have many local optima. The performance of an optimization method can be also measured by its capacity to escape local optima and converge to the global optimum. Deterministic gradient-based methods will need some restart procedures to escape local optima, and stochastic search methods will require a careful balance between exploitation and exploration. Moreover, it is well known that population based methods can help to avoid convergence to a local optima – but how large should the population be, depending on the characteristics of the objective function?

- ill-conditioned: Ill-conditioning is well defined for quadratic functions, as the ratio between the largest and the smallest eigenvalues. More generally, an ill-conditioned problem is a problem where different variables show a very different sensitivity in their contribution to the objective function value. For this kind of objective functions, algorithms exploring all directions with a unique radius will most likely fail in their search. Algorithms have to provide some adaptation rule for the search directions, in order to gradually learn the local conditioning. Ill-conditioning also suggests the use of second order information to learn about the local curvature of the objective function. In addition, this difficulty can lead to numerical failure of some line search methods used in gradient based methods.

- non-separable: A function is separable when its global optimum can be reached by successively optimizing in each of the dimensions. Such objective functions are hence easy to optimize. However, some search algorithm do implicitly exploit the separability of the objective function [62]. On the other hand, an algorithm that is invariant by a change of coordinate will perform exactly the same on a separable function and on its (non-separable) rotated instances, thus ensuring that its performances are not the result of the separability of the objective function.

The different available optimization methods will behave differently when facing the above-mentioned possible sources of difficulty. On the other hand, knowing the characteristics of a given objective function with respect to those possible difficulties will allow the user to choose an optimization method that can cope with the corresponding difficulty. For instance, multi-modality suggests the use of population-based methods; Ruggedness, non convexity, and noise suggest the use of randomized search methods; And ill-conditioning and non-separability suggest the use of an efficient and non isotropic adaptation mechanism for the search directions.

1.4.2 Invariance properties

On the other hand, according to the No Free Lunch theorem [145], no method can outperform all other methods on all test problems. Note that the No Free Lunch Theorem of

[145], applies to finite search spaces (which is not the case here) and states that assuming a uniform distribution over all 'possible problems', no method outperforms all other methods on average. When the search space is continuous, it is impossible to define the notion of a average over all possible problems [20]. However, it is possible to find methods optimal on some class of functions [20]. For example, quadratic approximation methods (see the paragraph on quadratic approximation methods), BFGS or the conjuguate gradient method will probably be more efficient (in the sense that they will probably need less computational effort to generate solutions close to the optimum) on quadratic objective functions than other methods which do not make use of a quadratic model hypothesis of the objective functions. However, there would exists other methods that will be probably more efficient, on non-convex objective functions, than methods making use of the quadratic model hypothesis . The same reasoning holds for example for the PSO method (see Section 1.2.3) which will be probably highly competitive on separable functions, but probably not the best choice on non-separable problems. Therefore, one should look to classes of problems where a given method might outperform another method. This is where invariance properties can play an important role: when a given method is invariant with respect to a set of transformations in the space of problems, assessing its ability to solve (with some 'reasonable' computational effort) one problem immediately demonstrates similar efficiency on the set of all transformed problems. Moreover, the more invariance properties an algorithm has, the more robust it is.

Given an objective function $f : \mathbb{R}^d \mapsto \mathbb{R}$, there are different ways to transform the problem of optimizing f on $\mathbb{R}^d$. First, any transformation $T : \mathbb{R} \mapsto \mathbb{R}$ can be used to transform the objective function f to another objective function $T(f)$. Instances of such common transformations are

- Translation: Addition of a constant, i.e., $T(f) = f + a$
- Scaling: Multiplication by a positive constant, i.e., $T(f) = a * f$, $(a > 0)$
- Monotonous transformation: Composition by an order-preserving function i.e., $T(f) = g \circ f$ where $g : \mathbb{R} \mapsto \mathbb{R}$ is a strictly increasing function

Another way to transform the problem of optimizing f into another problem is to apply f to a transformation of the input parameters, i.e., optimizing $f \circ U$ where $U : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a transformation of the search space. Search space transformations include translation, parameter rescaling and any linear change of coordinate (e.g., rotations).

Two important invariance properties have already been mentioned, and will be emphasized in the remaining of this Chapter. First, *monotonous invariance* is achieved by all rank-based methods, i.e., methods that only use comparisons of possible solutions (e.g. PSO, ESs, DE, most EDAs, but not gradient-based methods). A search method with the monotonous invariance property will behave exactly the same on f and $\sqrt{\sqrt{f}}$[6]. Second, *rotation invariance* is achieved by CMA-ES and DE without crossover, but also by gradient-based methods when the gradient is computed analytically (and not numerically, coordinate by coordinate) and ensures a robust behavior of the algorithm with respect to non-separability.

[6] Obviously f is positive in this case.

The importance of those invariance properties will be empirically illustrated in the following Sections.

1.4.3 Empirical comparisons

The most widely used search methods have been presented in Sections 1.1 and 1.2. When a real-world optimization problem is encountered, the practitioner will want to know which is the most efficient method to apply to the problem at hand. From our point of view, an efficient optimization method is a method that can offer a good compromise between the 'quality' of the solution proposed and the computational effort needed to generate such a solution. There are two ways to compare the efficiencies of optimization methods: theoretical and empirical. Few theoretical studies [128] have investigated the comparison of optimization methods. Moreover, theoretical studies rely on strong assumptions on the objective functions and/or the search space that are not satisfied in practice. Furthermore, according to Powell [108] "there seems to be hardly any correlation between the algorithms that are in regular use for practical applications and the algorithms that enjoy guaranteed convergence in theory". The efficiency of an optimization method is in general 'measured' when solving real-world problems. Therefore, empirical studies seems to be an effective way for comparing optimization methods.
Empirical studies comparing efficiencies and robustness of optimization methods [122, 106, 82, 9, 55] are usually done using a set of well-known tests functions. For instance, a set of test functions were collected in [134] to compare performances of optimization methods during a Special Session at the Congress on Evolutionary Computation (CEC2005). As pointed out in [62], any set of test functions should take into account the search difficulties as described in Section 1.4.1.

Probably the most investigated objective function test is the (quadratic) *sphere function*:

$$f_{sphere}(x) = x^T x = \|x\|^2, \ x \in \mathbb{R}^d \,, \tag{1.20}$$

where $\|.\|$ denotes the euclidean norm on $\mathbb{R}^d$. This function has a unique global minimum at $(0, \ldots, 0)$ and is therefore useful for local studies where the goal is to study the convergence of uni-modal objective functions toward a local optimum. A more general class of convex quadratic functions which can be written as $f(x) = x^T H x$, where H is a symmetric positive definite matrix, is often used to compare optimization methods, as the condition number of H (the ratio between its largest and smallest eigenvalues) gives a quantified information about the conditioning of the problem. The so-called *ellipsoid function* for instance is defined, for $x = (x_1, \ldots, x_d)$ as:

$$f_{ell}(x) = \sum_{i=1}^{d} \alpha^{\frac{i-1}{d-1}} x_i^2 \,, \tag{1.21}$$

where $\alpha > 0$ is the condition number of the function. One can test the behavior of a given algorithm for different condition numbers by changing the value of α.

Another widely used function, which is not quadratic, but also allows one to study the effect of ill-conditioning on the behavior of an algorithm is the so-called *diff-powers*

function:

$$f_{diff}(x) = \sum_{i=1}^{d} x_i^{2+\alpha\frac{i-1}{d-1}}, \tag{1.22}$$

where $\alpha > 0$ controls the conditioning of the problem.

For testing algorithms on multi-modals problems, the *Rastrigin function* is often used:

$$f_{ras}(x) = 10d + \sum_{i=1}^{d} (x_i^2 - 10\cos(2\pi x_i)) . \tag{1.23}$$

However, all the above functions are separable (as sum of functions of each variables). In order to test the effects of non-separability, any rotation using an orthogonal matrix M can applied on the search space: the functions $f_{ell} \circ M$ (with $\alpha \neq 1$), $f_{diff} \circ M$ (with $\alpha \neq 0$) and $f_{ras} \circ M$ are non-separable.

Finally, in order to test the robustness relatively to noisy objective functions, one can add to these functions a random variable, as what has been done for the sphere function in [25, 9].

Performance measurement In order to quantify and compare the performance of search algorithms, one has to introduce a quantity which measures how successful an algorithm is. Arnold and Beyer [9] have used as efficiency quantity, the ratio between the expected gain at each generation and the average number of evaluations at each iteration. Another quantity estimating the success performance has been used in [82]: A successful run is a run where the algorithm solves the problem i.e., reaches a given precision of the minimal objective function value before a fixed number of evaluations. Then the success performance is defined as the average number of function evaluations for successful runs over the empirical success rate. This success performance measure is called SP1.

1.4.4 Comparison of randomized search methods

In the previous sections, we present some popular randomized search methods. The simplest ones PRS and PAS can be seen as particular ES where no adaptation is used. For this reason, they are not efficient in practice when compared to self-adaptive ES as their computation time will be (relatively) very high. Concerning Simulated Annealing, the techniques introduced for the adaptation of its search parameters in [37] are similar to those used in the derandomized CMA-ES. Therefore here again SA can be seen as a particular ES. According to empirical studies, CMA-ES is shown to perform well on ill-conditioned non-separable problems [61, 59, 82] as well as on multi-modal problems [56, 82]. The CMA-ES algorithm is also highly competitive with all of the widely used randomized search methods, as shown in [55, 62]. The latter studies include the comparison of CMA-ES with other ESs, as well as with DE, the EMNA EDA, PSO and the Matlab implementation of BFGS. Moreover, CMA-ES performed best on the set of 25 test functions proposed during the CEC05 Challenge for continuous optimization [2]. CMA-ES was competitive in all uni-modal and multi-modal objective functions. Only on

separable functions, it was significantly outperformed by other competitors. Note however that a simple modification of CMA-ES has been proposed in order to increase the performance of CMA-ES on separable functions [116], constraining the covariance matrix to be diagonal.

1.4.5 Comparison of randomized and deterministic methods

According to different empirical studies [106, 9, 56], randomized search methods, and especially CMA-ES, are highly competitive and usually more robust than deterministic search methods in solving real-world optimization problems. In [106], population based methods and especially ES are shown to outperform deterministic point-based methods in noisy environments. In [56], empirical comparisons also include the BFGS method. In [9], the efficiency of a CMA-ES like algorithm[7], Hook and Jeeves pattern search algorithm, multi-directional search simplex method and Implicit Filtering are compared for the minimization of noisy objective functions. The comparison shows that for high search space dimensions and large amounts of noise strengths, the CMA-ES like strategy is the most efficient. As a matter of fact, the multi-directional search diverges for too high noise levels and Hook and Jeeves and Implicit Filtering stagnate for sufficiently high noise levels, whereas the performance of the ES algorithm gracefully decreases for high dimensions and high noise levels. Other empirical results comparing the performances of CMA-ES, DE, PSO, NEWUOA and BFGS have been recently presented by A. Auger and N. Hansen [18] in a tutorial session during the PPSN'08 conference. The results show that CMA-ES is more robust for wide class of objective functions, thanks to its invariance to transformations such as search space rotation, composition by an order-preserving function and a less deterioration of its performance when the objective function is more and more ill-conditioned. Relying on the empirical studies surveyed above, CMA-ES clearly seems the best default choice among the different search methods presented here, when no further information about the objective function is available. In fact, it is robust, having a competitive efficiency compared to other optimization methods when dealing with difficult optimization problems, especially in the case of non separable, non convex, ill conditioned, multi-modal and noisy objective functions. Of course, in the case of convex, relatively well-conditioned functions (condition number smaller than 10^5), methods such as NEWUOA or BFGS should be preferred.

However, it is worth noticing that real-world optimization objective functions are more likely to lie in one of the difficult classes described above than in the class of convex separable functions. In any case, the application part of this work (Chapter 5) will use CMA-ES to solve a real-world optimization problem.

[7]The algorithm used is called the Cumulative Step-Size Adaptation Evolution Strategy, which is referred to as $(\mu/\mu, \lambda)$-CSA-ES. This algorithm uses the same step length adaptation as in CMA-ES, but does not attempt to adapt the covariance matrix.

1.5 Survey of theoretical studies on Evolution Strategies: Non-noisy functions

This section will rapidly survey the existing theoretical studies of search algorithms belonging to the Evolution Strategy family.

The majority of theoretical studies of ES algorithms is concerned with isotropic ES, for which no search direction is preferred (the covariance matrix is equal to the identity matrix during the whole run and is not updated). Let (X_n) be the sequence of vectors in $\mathbb{R}^d$ generated by the ES method and $(f(\mathrm{X}_n))$ be the corresponding objective function values. The goal of theoretical studies is to investigate the limit of the sequence (X_n) (respectively $(f(\mathrm{X}_n))$) to the set of optima x^* (respectively to the minimal objective function value f^*).

The behavior of ES has been empirically observed to be log-linear, and we will start by formalizing this concept:

Definition 1.1. Let A be an algorithm designed for the minimization of an objective function with a unique global optimum. Let $(d_n)_n$ be the sequence of the distances to the optimum of the best points sampled by algorithm A at iteration n. Then algorithm A (or the sequence $(d_n)_n$) is said to have a log-linear behavior if there exists $c \neq 0$ such that $\lim_n \frac{1}{n} \ln(d_n) = c$. Note that depending on c, this can mean convergence or divergence: If $c > 0$, the algorithm diverges in the sense that the logarithm of the distance to the optimum will increase linearly to $+\infty$. We shall refer to this situation as log-linear divergence. On the opposite, if $c < 0$, the algorithm converges in the sense that the logarithm of the distance to the optimum will decrease linearly to 0. We shall refer to this situation as log-linear convergence.

Existing theoretical studies can be divided into two classes: global and local convergence studies.

1.5.1 Global convergence studies

Global convergence studies refer to theoretical studies where the objective function is not subject to many hypothesis. In particular, these studies include multi-modal objective functions. In the case of the simplest ES procedure, the $(1+1)$-ES, a sufficient condition ensuring almost sure convergence of the algorithm over a compact set [150] when the sequence of step-sizes, $(\sigma_n)_{n\in\mathbb{N}}$, is deterministically updated with zero as limit is that $\sigma_n \sqrt{\ln(n)} \to +\infty$ when n goes to $+\infty$.

In the case where the step-size is not updated, Rudolph [117], and later Chonghui and Huanwen [33] prove the same result of almost sure convergence of the sequence of objective functions solutions generated by the $(1+1)$-ES to the global minimal objective function value for continuous objective functions defined on a bounded search space. For a specific ES using quasi-random mutations and a specific deterministic adaptation rule of the step-size [19], an almost sure global convergence is shown using mild assumptions on the objective function.

A negative result was shown by Rudolph [119] in the case of the $(1+1)$-ES using the one-fifth adaptation rule: there is a strictly positive probability that the algorithm gets stuck in a local optimum.

1.5.2 Local convergence studies

All theoretical studies that will be presented in this thesis belong to the local convergence studies. These studies are concerned either with objective functions that possess a unique global optimum, or with the convergence of ES to a local optimum. Without loss of generality, we can suppose that in the general case that the local (or unique) optimum x^* that we are concerned with is $(0,\dots,0) \in \mathbb{R}^d$.

Local studies can in turn be classified into studies in finite dimension and studies where the dimension is assumed very large, that we will abusively call 'infinite dimension' studies.

Infinite dimension studies

By infinite dimension studies, we refer to studies that make the approximation of a search space dimension d going to $+\infty$. The general context of this studies is the so-called progress rate theory [114, 25]. This theory investigates quantities such as the progress rate, the fitness gain, or the success probability. The progress rate is the expected progress toward the optimum of a single iteration which can be written as the conditional expectation $E\left(1-\frac{\|\mathrm{X}_{n+1}\|}{\|\mathrm{X}_n\|}|\mathrm{X}_n\right)$. The fitness gain is the expected gain in fitness at each iteration.

A class of objective functions that have been widely investigated in progress rate theory is the class of the so-called spherical functions, which are real valued functions defined on $\mathbb{R}^d$ by $f(x) = \mathrm{g}(\|x\|^2)$ where $x \in \mathbb{R}^d$, $\mathrm{g} : [0,+\infty[\mapsto \mathbb{R}$ is an increasing function and $\|.\|$ denotes the euclidean norm on $\mathbb{R}^d$. All spherical functions have a unique global minimum reached on $(0,\dots,0)$. Infinite dimension studies had also investigated other objective function models such as the corridor model, various ridge functions and other positive definite quadratic forms.

These studies use some normalizations of underlying quantities such as the step-size mutation and the progress rate. These normalizations are useful when dimension d goes to infinity. The sign of the limit of the normalized progress rate determines whether the algorithm converges or diverges when the search space dimension is sufficiently high: A strictly positive normalized progress rate implies the convergence of the relative algorithm and a strictly negative normalized progress rate implies the divergence of the algorithm. Moreover, these studies investigate isotropic ES using either realistic adaptation rules such as the one-fifth adaptation rule, the self-adaptation rule or the cumulative step length adaptation rule, or an artificial adaptation rule called scale-invariant adaptation rule. The scale-invariant adaptation rule, which assumes that the distance to the optimum of a current solution is known at each iteration (which is not the case in practice), sets the step-size mutation at a given iteration proportional to this distance.

Studies using the progress rate theory are quantitative studies, asymptotic in the dimension of the search space but that rely on some approximations. Other asymptotic

studies have been carried out by J. Jägersküpper [72, 70, 71, 75][8], the proofs are rigorous at the expense of loosing quantitative results. Most of these studies aim to determine how the runtime of ES (or more general zeroth order methods) varies as a function of the search space dimension.

Finite dimension studies

Some local studies in the finite dimensional case have also been concerned with the scale-invariant adaptation rule: It has been proved [17] that this rule is optimal, in the sense that the convergence rate that is obtained with this rule when minimizing sphere functions is optimal. However, it has also been rigorously shown that the $(1, \lambda)$-ES converges (or diverges) log-linearly when minimizing spherical functions using either the optimal scale-invariant adaptation rule [17, 27] or the true self-adaptation rule [13, 27] (see Section 1.2.4). Those results [13, 17, 27] have been established using the Laws of Large Numbers (LLN) for independent random variables or for random variables constituting a Markov chain sequence. A complete presentation of the theory investigating the stability of Markov chain sequences can be found in [97].

Other results have been obtained for more general classes than sphere functions. For a specific class of convex objective functions, Rudolph [118] investigates the $(1, \lambda)$-ES where mutations follow a uniform distribution on the sphere and the step-size is adapted proportionally to the norm of the gradient on the current solution (at iteration n, the step-size σ_n is set to $\sigma\|\nabla \mathrm{X}_n\|$, where X_n is the current solution and σ is a strictly positive constant). He proves that the sequence of objective functions $(f(\mathrm{X}_n))_n$ converges geometrically fast to the optimal value provided that σ is sufficiently small. A. Auger et al. [14] investigate a similar $(1, \lambda)$-ES algorithm using Gaussian mutations and either the scale-invariant adaptation rule (i.e., $\sigma_n = \sigma\|\mathrm{X}_n\|$) or the gradient-proportional rule (i.e., $\sigma_n = \sigma\|\nabla \mathrm{X}_n\|$, for some $\sigma > 0$). They prove that the sequence $(f(\mathrm{X}_n))_n$ converges to the optimal solution almost surely and in L^1, for a specific class of twice continuously differentiable objective functions. This result was established using the martingale theory and holds for sufficiently small values of σ.

Finally, A. Auger and N. Hansen [17] have bridged the gap between the progress rate theory and finite dimension studies. In the context of the minimization of spherical functions, they introduce the so-called *log-progress rate* as the conditional expectation $E\left(\ln(\|\mathrm{X}_n\|) - \ln(\|\mathrm{X}_{n+1}\|)|\mathrm{X}_n\right)$. They prove that the sign of this quantity gives the almost sure convergence of the algorithm for finite dimensions. Moreover, they have shown that, when using the normalizations that are used in the context of the progress rate theory, the limits of the normalized log-progress rate and of the normalized progress rate are equal when the search space dimension d goes to infinity. Another important point of their study is that, for finite dimension, the sign of the normalized progress rate determines the convergence in mean of the solutions generated by the $(1, \lambda)$-ES algorithm, and not the almost sure convergence.

[8] For the first reference, the work has been done in collaboration with Carsten Witt.

1.6 Survey of theoretical studies on Evolution Strategies: Noisy functions

1.6.1 Motivations

The most important part of the work presented in this thesis deals with the optimization of noisy objective functions. Noisy optimization is an important part of optimization, because noisy objective functions are very frequently encountered in real-world optimization problems. Several situations may lead to noisy objective functions. Objective functions can be the result of some physical measurements, and the measured values will differ due to the variability of experimental conditions at each measurement. Noise can be also the consequence of user input. Also objective functions resulting from Monte-Carlo simulations are noisy due to their stochastic nature: the precision of these methods depend on the number of iterations, but the results over different simulations will always have a positive variance.

These examples share the property that the reevaluation of these objective function with the same input data will lead to different values: we shall assume that the noise investigated here is an unknown random variable. The randomness of the noisy part of objective functions removes an important part of the information on this function. This means that ruggedness can be taken into account by the model of a noisy objective function.

Many papers have been devoted to theoretical or empirical investigations of optimization of noisy objective functions [138, 23, 106, 80, 24, 25, 7, 5, 8, 9, 10, 136]. In many empirical studies [9, 106, 138], noisy objective functions are used to assess the performances of different strategies. The work in [9] demonstrates the efficiency and the robustness of a CMA-ES-like algorithm (which is an algorithm similar to CMA-ES but wich does not use the adaptation of the covariance matrix) when dealing with noisy objective functions. Furthermore, for high noise levels, this CMA-like method outperforms the implicit filtering method, a method that was especially designed to deal with noise (see Section 1.1.2). In [106], the efficiency of population-based methods is compared to that of deterministic point-methods in noisy environments. The results favor population-based optimization, and ES in particular.

ES have thus been empirically demonstrated to be robust when minimizing noisy objective functions. However, the most investigated theoretical studies are infinite dimension studies [24, 7, 25, 8, 5] and rely on many approximations and normalizations (see Section 1.6.3).

1.6.2 Evolutionary Algorithms in noisy environments

Evolutionary Algorithms are known to be robust with respect to noise, as has been known for long in the context of discrete search spaces [44, 112, 100]. However, studies of GA in noisy environment are mostly empirical and, to the best of our knowledge, do not include any theoretical investigation.

In [24], H.-G. Beyer surveyed some studies on the behavior of different flavors of EAs (GA, ES and EP). In particular, despite the fact that GA (for discrete search spaces)

and ES (for continuous optimization) operate on different search spaces, their behaviors show some similarities when applied to noisy objective functions. In fact, the noise results in a decrease of the convergence speed, and leads to a loss of accuracy in terms of the localization of the optimum.

The critical issue when optimizing noisy objective functions is that it can make the selection process unreliable, and hence turn any search algorithm into some kind of random walk. However, because the noise is assumed to have zero mean, stochastic techniques can cope with rather high levels of noise by over-sampling the noisy fitness function: this can be achieved by assigning to each new individual an average of several evaluations of the fitness function. Another possible solution is to increase the population size: The non-zero variance of the population size in the case of the (μ, λ)-ES [7], or, the genetic repair of the $(\mu/\mu_I, \lambda)$-ES [6], lead to an increase of the performances of these strategies in noisy environments. In the same case of ES, another solution, that has been analyzed in [24], is concerned with the use of rescaled mutations: The standard ES Gaussian mutation is replaced by equation:

$$(Y_n)_j = (\mathrm{X}_n)_i + \frac{1}{k}\sigma_n \mathrm{N}_j(0, C_n), \tag{1.24}$$

where $k > 1$ is the rescaling parameter. As stated by Beyer, "the $(1, \lambda)$-ES can perform large search steps with the result of larger fitness differences which will be significant over the noise level."

There has been, however, some theoretical studies about the behavior of ESs in noisy environments, that will now be described.

1.6.3 Theoretical results for noisy optimization

Theoretical studies of optimization of noisy objective functions using ES have been mainly done in the context of of the progress rate theory in infinite dimension. However, few studies in finite dimension have been done in the context of optimization of noisy objective functions using ES [136].

The first infinite dimension studies of ES on noisy environments have been carried out by Rechenberg [114], who investigated the computation of the progress rate on the noisy instances of the sphere and the corridor functions. He succeeded in calculating the progress rate of the $(1+1)$-ES for the minimization of the noisy corridor function. Twenty years later, Beyer [23] computed the progress rate for the $(1 + \lambda)$-ES and $(1, \lambda)$-ES when minimizing the noisy sphere function. Since then, many works by Arnold and Beyer have studied the behavior of ES on noisy objective functions [10, 5, 7, 8, 24, 25, 23]. These studies cover the comma strategies [7], the comma strategies with recombination [6], and the plus strategies [8, 25, 23]. Note the plus strategies in [23, 25] in fact use a particular plus strategy in which the fitness of the parent is reevaluated at each iteration.

Noise model

Before starting a theoretical investigation of a noisy fitness function, a model has to be chosen for the noise. Let f be a fitness function with a minimal value f^* supposed to be

equal to zero (termed the 'ideal' fitness in the following). There are several possible ways to build a noisy fitness function f_{noisy} from f.

A first natural idea is to add to the ideal function some random variable, for example a Gaussian random variable: $f_{\text{noisy}}(x) = f(x) + \epsilon \mathrm{N}(0,1)$ where the noise level ϵ is a constant value. A possible defect of this model is that the noise can dominate the ideal fitness when getting close to the optimum, and consequently leads the search to behave like a random walk.

Another idea, which is often true in the case of quadratic (ideal) objective functions (which will be investigated in this thesis), is that the behavior of the algorithm depends on the ratio between the noise level and the values of the ideal objective function. This is why the noise level should be proportional to the ideal objective (quadratic) function. Note that this statement is not necessarily verified in general. In fact, for cubic or quartic ideal objective functions for example, the behavior of the algorithm really depends on the ratio between the noise level and the standard deviation of the ideal fitness values in the population. Therefore, for the specific case of quadratic ideal objective function, the idea of having a noise level proportional to the ideal objective function, should be suitable, and leads to a multiplicative noise model which writes as $f_{\text{noisy}}(x) = f(x)(1+\sigma_\epsilon \mathrm{N}(0,1))$.

In the studies cited above [10, 5, 7, 8, 24, 25, 23], the objective function is the so-called noisy sphere function: the Gaussian noise[9] has a standard deviation proportional to the ideal fitness, or, equivalently, to the distance to the optimum (for the sphere function). Moreover, the noise model takes into account an additional normalization of the noise strength with respect to the search space dimension d. In a more general context of ideal objective functions $f(x) = \|x\|^\alpha$ with $\alpha > 0$, the noise strength σ_ϵ should be written [25] $\frac{\alpha \sigma_\epsilon^*}{d}$, where $\sigma_\epsilon^* > 0$ is called normalized noise strength. Therefore, the model of noisy sphere function with a fitness-proportional Gaussian noise can be written as:

$$\mathrm{f}(x) = \|x\|^2 + \frac{2\sigma_\epsilon^*}{d}\|x\|^2 \mathrm{N}(0,1)\,. \tag{1.25}$$

In addition to the normalization of the noise strength, Arnold and Beyer use the same normalizations relative to the progress rate and the step-size mutation that had been introduced in the non-noisy case for the theoretical studies in the context of the progress rate theory (see Section 1.5.2). Using these normalizations, Arnold and Beyer [8] approximate the standard deviation of the noise at the offspring location by that at its parent location. Their argument is that, in very large dimension, the parent and its offspring are so close that the fitness has the same noise level at both locations. Mathematically speaking, if we denote y an offspring of a parent x, the expression of the fitness of the offspring which, according to Eq. 1.25, writes as $\mathrm{f}(y) = \|y\|^2 + \frac{2\sigma_\epsilon^*}{d}\|y\|^2 \mathrm{N}(0,1)$ is well approximated by $\mathrm{f}(y) = \|y\|^2 + \frac{2\sigma_\epsilon^*}{d}\|x\|^2 \mathrm{N}(0,1)$. The random part $\frac{2\sigma_\epsilon^*}{d}\|y\|^2 \mathrm{N}(0,1)$ is replaced by $\frac{2\sigma_\epsilon^*}{d}\|x\|^2 \mathrm{N}(0,1)$.

Infinite dimension results

As in the non-noisy case, the sign of the limit of the normalized progress rate is sufficient to indicate whether the algorithm converges or diverges, in the limit of infinite dimension for

[9] Note that the study in [10] does not assume a Gaussian noise.

the noisy sphere function. The first (expected) result that can be seen in the plots of the limit of the normalized progress rate as a function of the normalized step-size mutation for different normalized noise strengths (see for example [25, Fig 3.10], [8, Fig 6], [7, Fig 4]) is that the normalized progress rate decreases when the normalized noise strength increases. In particular, the best normalized progress rate corresponds to the non-noisy case (for which the noise strength σ_ϵ equals 0). For comma strategies, it is proved in [25, Fig 3.10] for the $(1, 5)$-ES that :

- For 'small' values of the normalized noise strength, the algorithm converges for small values of the normalized step-size mutation and diverges for sufficiently 'large' values of the normalized step-size mutation, and
- For 'large' values of the normalized noise strength: the algorithm diverges for any value of the normalized step-size mutation.

For plus strategies, the curves in [8, Fig 6], plotted using some normalized noise strength values, suggest that the $(1+1)$-ES which does not use reevaluation of the parent converges for any value of the normalized step-size mutation. For plus strategies, and using the reevaluation of the parent at every iteration, the plots in [25, Fig 3.12] and [8, Fig 6] suggest that for 'small' values of the normalized noise strength the algorithm converges and that it can diverge for large normalized noise strengths.

The performance of these different ES strategies (which do not use recombination) has been compared in [7, Fig 6] as a function of the normalized noise strength. It is shown that for small normalized noise strength values, plus strategies perform better than comma strategies, and that the opposite happens for large normalized noise strength values.

Moreover, some computations in the infinite dimension setting were used to decide whether re-sampling and/or increasing the population size can improve the performance of the ES in noisy environments: the $(\mu, \mu\lambda)$-ES performs slightly better than the $(1, \lambda)$-ES when using re-sampling [24]. For 'large' noise strengths, the expressions of the progress rate derived by Arnold and Beyer [25, 6] suggests that it is better to reevaluate and re-sample than to increase λ for the $(1, \lambda)$-ES, and that one should increase μ when using the $(\mu/\mu, \lambda)$-ES.

Finally, the adaptation of the mutation step-size when optimizing noisy objective functions was studied. The usefulness of the one-fifth rule was discussed in [8] and that of a self-adaptive strategy with a rescaled mutation in [24]. An interesting result was derived in [6] where the efficiency of cumulative step length adaptation when dealing with noisy environments was shown for the $(\mu/\mu, \lambda)$-ES minimizing the noisy sphere function. More precisely, the study suggests that cumulative step length adaptation generates step lengths in the vicinity of optimal ones provided that population sizes are sufficiently large. However, a limitation of this results, which has been done in the limit of infinite search space dimension, is that it requires at the same time sufficiently large population sizes and $\lambda << d$.

In our theoretical and numerical study, we investigate first (Chapter 2) the optimization using the $(1 + 1)$-ES of non noisy objective functions. Then in Chapters 3 and 4, we investigate the behavior, when minimizing noisy objective functions, of the scale-invariant $(1 + 1)$-ES (Chapter 3) and of the scale-invariant $(1, \lambda)$-ES (Chapter 4). For

the studies in noisy environments, the noisy objective function model is similar to the one investigated by Arnold and Beyer given in Eq. 1.25 but the noise distribution is not necessary supposed to be Gaussian. In fact, the distribution of the random part of noisy objective functions investigated here include lower bounded and unbounded distributions. Moreover, we uses mild assumptions on the noise distribution. Finally, we theoretically investigate the reliability of some approximations used by Arnold and Beyer.

1.7 Discussion

In previous sections (Sections 1.4.4 and 1.4.5), we have shown that ESs and in particular CMA-ES are efficient to solve difficult optimization problems. We give a particular interest to the difficulties that can be caused by noisy objective functions which are frequently encountered in practice. In particular, ESs using recombination, have been empirically shown [9, 106] to be more robust than other deterministic or randomized search methods in noisy environments. A first goal of this thesis is then to study theoretically and numerically the behavior of some simple ESs (simpler than CMA-ES) in noisy environments as they performed better in noisy environments. We are convinced that both theoretical and numerical approaches have to be investigated in a complementary approach. In fact, theoretical studies are helpful to explain the behavior of a given method but they need strong assumptions on objective functions, that are not satisfied in practice. Numerical approaches are also helpful in order to improve our understanding of the behavior of the algorithms, but one has to be careful not to hastily turn some behaviors that have been observed in very particular cases into general truths. Here again, a theoretical study can help understanding the experimental facts. For this reason, our studies are based on establishment of convergence theorems with numerical simulations that illustrate results and that helped us for the understanding of the behavior of the algorithms and wee guidelines for our theoretical results. Previous theoretical studies of ES in noisy environments (see Section 1.6) lie on the limit of infinite dimension of the search space. This hypothesis allow to use some approximations. Moreover, some normalizations have been frequently used. The noise distribution is also restricted to the Gaussian model (see Eq. 1.25). In our work, we want to investigate theoretically ESs (in particular, in noisy environments) when the dimension of the search space is finite and compare our results to to infinite dimension results. In the particular case of noisy objective functions, the noise is not always assumed to be Gaussian. Another motivation for this study is that, as pointed out in [17], infinite dimension results [17] usually provide convergence in mean results and in this work we want to give almost sure convergence results.

In the second part of this thesis, CMA-ES is applied to solve a real-world optimization problem. The problem had been previously tackled using gradient-based strategies [74, 73] and one of the goals of this study is to compare performances of randomized and deterministic search methods in this specific study and see whether it is true or not that randomized search methods seem to be more robust that deterministic search methods in solving real-world optimization problems.

Theoretical and Numerical Study

Chapter 2

Log-linear Convergence and Optimal Bounds for the $(1+1)$-ES

The material in this Chapter is mainly contained in the paper [77] that has been published in a Springer Verlag LNCS volume containing a selection of papers presented at the conference *Evolution Artificielle 2007*. This work has been done in collaboration with Pierre Liardet.

In this paper, we have studied the $(1+1)$ isotropic ES for minimizing real valued objective functions defined in $\mathbb{R}^d$ ($d \geq 1$). We have shown two main results:

- *Theorem 2.4:* The convergence of the $(1+1)$ isotropic ES is at most log-linear and the optimal convergence rate is derived.
- *Theorem 2.10:* The convergence of the specific $(1+1)$-ES using a scale-invariant adaptation rule is log-linear when the objective functions are the so-called spherical functions, $f(x) = \mathrm{g}(\|x\|^2)$ where $x \in \mathbb{R}^d$, $\mathrm{g} : [0, +\infty[\mapsto \mathbb{R}$ is an increasing function and $\|.\|$ the euclidean norm on $\mathbb{R}^d$. Moreover, the optimal convergence rate, that can be rached when the $(1+1)$ isotropic ES optimize any objective function using any adaptation rule, is obtained when the adaptation rule of the step-size is the scale-invariant adaptation rule and the objective function is the spherical function.

The log-linear behavior of the scale-invariant $(1+1)$-ES is established using the Law of Large Numbers (LLN) for orthogonal random variables (Theorem 2.9). This theorem has been derived from [93, p. 458].
Similar results had been previously proved in the case of the $(1, \lambda)$-ES: The log-linear behavior (convergence or divergence) of the scale-invariant $(1, \lambda)$-ES minimizing spherical functions have been previously shown in [27, 17]. The result has been derived using the LLN for independent random variables and suggests that the convergence results obtained hold in probability. In [12], it is stated that almost sure convergence is obtained using similar techniques as in [13] where the proof relies on the LLN for Markov chains. For specific classes of twice continuously differentiable objective functions, it has been shown in [14], that almost sure convergence holds for adaptive $(1, \lambda)$-ES with the scale-invariant adaptation rule of the step-size mutation σ_n i.e., $\sigma_n = \sigma \|\mathrm{X}_n\|$ or with different step-size mutations at each direction $(\sigma_n)^i = \sigma |\frac{\partial f(\mathrm{X}_n)}{\partial x_i}|$, where for $i \in \{1, \ldots, d\}$, $\frac{\partial f}{\partial x_i}$ is the i-th

partial derivative of f and X_n is the solution at an iteration n. Those results were derived using tools of martingale theory. For the $(1, \lambda)$-ES using a realistic self-adaptation rule, the log-linear behavior on spherical functions has been shown in [13] using the LLN for Markov chains. The optimality of the scale-invariant adaptation rule when minimizing spherical functions has been already rigorously derived for comma strategies in [17].

The contribution of this study is that it provides tight bounds for (1+1)-ES algorithms. The optimal bounds derived in this work can be used to assess the performances of a given (realistic) step-size adaptation strategy comparing the convergence rate achieved by the strategy with the optimal one, given by the (artificial) scale-invariant algorithm on sphere function.
The optimal convergence rate that can be reached by a (1+1)-ES algorithm is given by the value of σ maximizing the function F defined in Lemma 2.1. The theoretical computation of the optimal σ value is presumably impossible. However, as the convergence rate F is expressed as a function of an expectation, its computation (and then that of the optimal σ value) is investigated using Monte Carlo simulations when the search space dimension d is finite.

In the conclusion of the paper (Section 2.5), we state that the computation of the value of σ maximizing the convergence rate is equivalent to that of σ maximizing the log-progress $E(\ln \|\mathrm{X}_n\|) - E(\ln \|\mathrm{X}_{n+1}\|)$. We also state that, when the search space dimension d goes to infinity, the quantities $d(E(\ln \|\mathrm{X}_n\|) - E(\ln \|\mathrm{X}_{n+1}\|))$ and the so-called normalized progress rate $d(E(\|\mathrm{X}_n\|) - E(\|\mathrm{X}_{n+1}\|))$ are equal when replacing σ by σ^*/d ($\sigma^* > 0$), having a limit that only depends on σ^* that we can denote $l(\sigma^*)$ and whose expression is the opposite value of the one given in [25, Eq. 3.88]. The limit $l(\sigma^*)$ is also the limit, when d goes to infinity, of the normalized convergence rate $dF(\sigma^*/d)$ where F is defined in Lemma 2.1. These statements could be rigorously shown using the same technique used in Chapter 4 where a more complicate result is given in the specific case of comma strategies. Morevoer, the result will enable us to state the convergence rate varies asymptotically linearly with the inverse of the search space dimension. On the other hand, it is worth noticing, that in the similar context of a $(1+1)$-ES using isotropically distributed mutation vectors and minimizing spherical functions, an algorithmic analysis of how the runtime of the $(1+1)$-ES depends on the search space has been performed by J. Jägersküpper [71]. In particular, Jens shows for the one-fifth adaptation rule that, the time to halve the distance to the optimum is linear in the dimension. This is an other way to state the result (shown here) that the dependence of the convergence rate is inversely proportional to the dimension. However, Jens studies being asymptotic in the dimension, no convergence rates for finite dimension can be derived.

Log-linear Convergence and Optimal Bounds for the $(1+1)$-ES

Mohamed Jebalia[1], Anne Auger[1], and Pierre Liardet[2]

[1] TAO Team, INRIA Futurs
Université Paris Sud, LRI, 91405 Orsay cedex, France
{mohamed.jebalia,anne.auger}@lri.fr

[2] Université de Provence, UMR-CNRS 6632
39 rue F. Joliot-Curie - 13453 Marseille cedex 13, France
liardet@cmi.univ-mrs.fr

Proceedings of Evolution Artificielle 2007, pp 207-218.
The original publication is available at http://www.springerlink.com

Errata :
There are few errata in the published paper:

- Proof of Lemma 2.1: The surface area of the d-dimensional unit ball should read $S_d = 2\pi^{d/2}/\Gamma(\frac{d}{2})$.
- Proof of Proposition 2.7: 1) The quantities Y_n and Y'_n are random variables, not random vectors. 2) Last equation: The right hand side of the first line should be $\frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left(\ln^- \left(\left\| \frac{\mathrm{X}_m}{\|\mathrm{X}_m\|} + \sigma x \right\| \right) \right) e^{-\frac{\|x\|^2}{2}} dx - F(\sigma)$.
- In Fig 2.1, the plots are rather related to the definition of F given in Eq. 2.3 than to Eq. 2.4 which is a consequence of Eq. 2.3.
- A spelling mistake in the sentence just after Eq. 2.1, the world "euclidian" should be written as "euclidean".

- The word "independency" appears twice in the chapter (in the proofs of Lemma 2.2 and of Proposition 2.7) and should be replaced by "independence".
- In the proof of Lemma 2.2, one should have "Using the independence of $\sigma\|X\|^{-1}$ and $\mathcal{N}$ ...".
- Before Theorem 2.8, one should have " But the random vectors Y_n' are i.i.d. ..." instead of " But the random vectors Y_n are i.i.d. ...".

Log-linear Convergence and Optimal Bounds for the $(1+1)$-ES

Mohamed Jebalia[1], Anne Auger[1], and Pierre Liardet[2]

[1] TAO Team, INRIA Futurs
Université Paris Sud, LRI, 91405 Orsay cedex, France
{mohamed.jebalia,anne.auger}@lri.fr

[2] Université de Provence, UMR-CNRS 6632
39 rue F. Joliot-Curie - 13453 Marseille cedex 13, France
liardet@cmi.univ-mrs.fr

Abstract

The $(1+1)$-ES is modeled by a general stochastic process whose asymptotic behavior is investigated. Under general assumptions, it is shown that the convergence of the related algorithm is sub-log-linear, bounded below by an explicit log-linear rate. For the specific case of spherical functions and scale-invariant algorithm, it is proved using the Law of Large Numbers for orthogonal variables, that the linear convergence holds almost surely and that the best convergence rate is reached. Experimental simulations illustrate the theoretical results.

2.1 Introduction

Evolutionary algorithms (EAs) are bio-inspired stochastic search algorithms that iteratively apply operators of variation and selection to a population of candidate solutions. Among EAs, adaptive Evolution Strategies (ESs) are recognized as state of the art algorithms when dealing with continuous optimization problems. Adaptive ESs sequentially adapt the parameters of the search distribution, usually a multivariate normal distribution, based on the history of the search. Several adaptation schemes have been introduced in the past. The one-fifth success rule [114, 82] considers the adaptation of one parameter, referred as the step-size, based on the success probability. The most advanced adaptation scheme, the Covariance Matrix Adaptation (CMA), adapts the full covariance matrix of the multivariate normal distribution [61].

The first theoretical works carried out in the context of Evolution Strategies focused on the so-called progress rate defined as a one-step expected progress towards the optimum [114, 25]. The progress rate approach consists in looking for step-sizes maximizing the

expected progress. This amounts to investigating an artificial step-size adaptation scheme called scale-invariant, in which, at each iteration, the step-size is proportional to the distance to the optimum. The results derived in the context of the progress rate theory hold asymptotically in the dimension of the search space and the techniques used do not allow to obtain finite dimension estimations.

Finite dimension results were obtained in the context of 'comma' strategies on the class of the so-called sphere functions, mapping $\mathbb{R}^d$ into $\mathbb{R}$ (d being the dimension of the search space) and defined as

$$f(x) = \mathrm{g}(\|x\|^2)\,, \tag{2.1}$$

where $\mathrm{g} : [0, +\infty[\mapsto \mathbb{R}$ is an increasing function and $\|.\|$ denotes the usual euclidian norm on $\mathbb{R}^d$. On this class of functions, scale-invariant ESs [27] and self-adaptive ESs (which use a real adaptation rule) [27, 13] do converge (or diverge) with order one, or log-linearly[1].

In this paper, finite dimension results are investigated and the focus is on the simplest ES, namely the $(1+1)$-ES. Section 2.2 introduces the mathematical model associated to the algorithm in a general framework and provides preliminary results. In Section 2.3, a sharp lower bound of the log-convergence rate is proved. In Section 2.4, it is shown that this lower bound is reached for a scaled-invariant algorithm on the class of sphere functions. The proof of convergence on the class of sphere functions uses the Law of Large Numbers for orthogonal random variables. A central limit theorem is also derived from this analysis. In Section 2.5 our results are discussed and related to previous works. Some numerical experiments illustrating the theoretical results are presented.

2.2 Mathematical model for the $(1+1)$-ES

Let $\mathbb{R}^d$ be equipped with the Borel σ-algebra and the Lebesgue measure. In the sequel we always assume that $(\mathcal{N}_n)_n$ denotes a sequence of random vectors (r.vec.) independent and identically distributed (i.i.d.), defined on a suitable probability space (Ω, P), with common law the multivariate isotropic normal distribution on $\mathbb{R}^d$ denoted by $\mathrm{N}(0, I_d)$ [(2)]. Let $(\sigma_n)_n$ be a given sequence of positive random variables (r.var.). We also assume that for each index n, σ_n is defined on Ω and is independent of $\mathcal{N}_n$; further we will also require that the sequences $(\sigma_n)_n$ and $(\mathcal{N}_n)_n$ are mutually independent. Finally, let $f : \mathbb{R}^d \to \mathbb{R}$ be an objective function (which is always assumed to be Lebesgue measurable) and let $\delta_n : \mathbb{R}^d \times \Omega \to \{0, 1\}$ $(n \geq 0)$ be the measurable function defined by $\delta_n(x, \omega) := \mathbf{1}_{\{f(x+\sigma_n(\omega)\mathcal{N}_n(\omega)) \leqslant f(x)\}}$. In this paper, $(1+1)$-ES algorithms are modeled by the $\mathbb{R}^d$-valued random process $(X_n)_{n \geqslant 0}$ defined on Ω by the recurrence relation

$$\mathrm{X}_{n+1} = \mathrm{X}_n + \delta_n(\mathrm{X}_n, I_\Omega)\sigma_n\mathcal{N}_n\,, \tag{2.2}$$

where I_Ω is the identity function $\omega \mapsto \omega$ on Ω and X_0 is given.

[1] We say that the sequence $(\mathrm{X}_n)_n$ converges log-linearly to zero (resp. diverges log-linearly) if there exists $c < 0$ (resp. $c > 0$) such that $\lim_n \frac{1}{n} \ln \|\mathrm{X}_n\| = c$.

[2] $\mathrm{N}(0, I_d)$ is the multivariate normal distribution with mean $(0, \ldots, 0) \in \mathbb{R}^d$ and covariance matrix the identity I_d.

The classical terminology used for algorithms defined by (2.2) stresses the parallel with the biology: the iteration index n is referred as generation, the random vector X_n is called the parent, the perturbed random vector $\tilde{\mathrm{X}}_n = \mathrm{X}_n + \sigma_n \mathrm{N}_n$ is the n-th offspring. The scalar r.var. σ_n is called step-size. The r.var. δ_n translates the plus selection "+" in the $(1+1)$-ES: the offspring is accepted if and only if its fitness value is smaller than the fitness of the parent. Several heuristics have been introduced for the adaptation of the step-size σ_n, the most popular being the one-fifth success rule [114, 82].

Notations and preliminary results

For a real valued function $x \mapsto h(x)$ we introduce its positive part $h^+(x) := \max\{0, h(x)\}$ and negative part $h^- = (-h)^+$. In other words $h = h^+ - h^-$ and $|h| = h^+ + h^-$. In the sequel, we denote by e_1 a unitary vector in $\mathbb{R}^d$. The following technical lemmas will be useful in the sequel.

Lemma 2.1. Let $\mathcal{N}$ be a r.vec. of distribution $\mathrm{N}(0, I_d)$. The map $F : [0, \infty] \to [0, +\infty]$ defined by $F(+\infty) := 0$ and

$$F(\sigma) := E\left[\ln^-\left(\|e_1 + \sigma\mathcal{N}\|\right)\right] = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \ln^-(\|e_1 + \sigma x\|) e^{-\frac{\|x\|^2}{2}} dx \tag{2.3}$$

otherwise, is continuous on $[0, +\infty]$ (endowed with the usual compact topology), finite valued and strictly positive on $]0, \infty[$.

Proof :
The integral (2.3) always exists but could be infinite. In any case, $F(\sigma)$ is independent of the choice of e_1 due to the invariance of $\mathcal{N}$ under rotations. For convenience we choose $e_1 = (1, 0, \ldots, 0)$ so that $\ln^-(\|e_1 + \sigma x\|) = 0$ if $x = (x_1, \ldots, x_d)$ with $x_1 \geqslant 0$. Let $f_1 : \mathbb{R}^d \times [0, \infty] \to [0, +\infty]$ be defined by

$$f_1(x, \sigma) = \ln^-(\|e_1 + \sigma x\|^2) e^{-\frac{\|x\|^2}{2}}$$

for $x \neq (-1/\sigma, 0, \ldots, 0)$ and $f_1((-1/\sigma, 0, \ldots, 0), \sigma) = +\infty$ (with $\sigma > 0$) and finally $f_1(x, +\infty) = 0$ $(= \lim_{\sigma \to +\infty} f_1(x, \sigma))$. Notice that $f_1(x, \sigma) = 0$ if $x_1 \geqslant 0$ and readily $f_1((x_1, x_2, \ldots, x_d), \sigma) = f_1((x_1, \epsilon_2 x_2, \ldots, \epsilon_d x_d), \sigma)$ for any $(\epsilon_2, \ldots, \epsilon_d)$ in $\{-1, +1\}^{d-1}$ so that we can restrict the integration giving $F(\sigma)$ to the domain $\mathcal{D} :=]-\infty, 0[\times]0, \infty[^{d-1}$, more precisely one has

$$F(\sigma) = \frac{1}{4}\Big(\frac{2}{\pi}\Big)^{d/2} \int_{\mathcal{D}} f_1(x, \sigma) dx \tag{2.4}$$

with in addition f_1 is finite everywhere in $\mathcal{D}$. From the definition of $F(+\infty)$ and f_1 one has $\frac{1}{4}(2/\pi)^{d/2} \int_{\mathcal{D}} f_1(x, +\infty) dx = 0 = F(+\infty)$ so that (2.4) holds also for $\sigma = +\infty$. Now, for any real number $\sigma > 0$ fixed, the inequality $f_1(x, \sigma) > 0$ holds on $B_\sigma := \{x \in \mathcal{D}\,;\ \|e_1 + \sigma x\| < 1\}$ which is a nonempty open set, therefore $F(\sigma) > 0$. In addition,

$f_1(x,0) = 0$ for all x and so, $F(0) = 0$. Passing to spherical coordinates (with $d \geqslant 2$)we obtain after partial integration

$$\int_{\mathcal{D}} f_1(x)dx \;=\; 2c_d \int_0^{+\infty} \int_0^{\pi/2} \ln^-(|\sigma r - e^{i\theta_1}|) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}\theta_1 dr\; d\theta_1$$

where

$$c_d = \int_0^{\pi/2} \cdots \int_0^{\pi/2} \sin^{d-3}(\theta_2) \ldots \sin(\theta_{d-2}) d\theta_2 \ldots d\theta_{d-1}$$

for $d \geqslant 3$ and $c_2 = 1$. With the classical Wallis integral $W_{d-2} = \int_0^{\pi/2} \sin^{d-2}\theta \; d\theta$ and the surface area of the d-dimensional unit ball $S_d = 2\pi^{d/2}/\Gamma(\frac{n}{2})$ we have $S_d = 2^d c_d W_{d-2}$ and after collecting the above results we get

$$F(\sigma) = \Big(\frac{1}{2\pi}\Big)^{d/2} \frac{1}{W_{d-2}\Gamma(\frac{d}{2})} \int_0^{+\infty}\int_0^{\pi/2} \ln^-(|\sigma r - e^{i\theta}|) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta)\, dr\, d\theta\,.$$

The integrand $g : (r,\theta,\sigma) \mapsto \ln^-(|\sigma r - e^{i\theta}|) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta)$ defined on the set $]0,+\infty[\times[0,\pi/2] \times [0,\infty]$ (with $g(r,\theta,+\infty) = 0$) is continuous. In fact, the continuity is clear at each point (r,θ,σ) with $\sigma \neq +\infty$ and for the points $(r,\theta,+\infty)$, one has $g(\rho,\alpha,\sigma) = 0$ on $]r/2,+\infty[\times[0,\pi/2]\times]\frac{4}{r},+\infty]$. Moreover, g is dominated by $g_1 : (r,\theta) \mapsto \ln^-(\sin\theta) r^{d-1} e^{-r^2/2}$ *i.e.*, $g(r,\theta,\sigma) \leqslant g_1(r,\theta)$ for all (r,θ,σ) in $]0,+\infty[\times[0,\pi/2] \times [0,+\infty]$. Since g_1 is integrable, the continuity of F on $[0,+\infty]$ follows from the Lebesgue dominated convergence theorem. For the remaining case $d = 1$ the conclusions of the lemma follow easily from (2.4) that gives $F(\sigma) = \frac{1}{2\sqrt{2\pi}} \int_0^{\infty} \ln^-(|1-\sigma r|) e^{-\frac{r^2}{2}} dr$. □

Corollary 1. The supremum $\tau := \sup F([0,+\infty])$ is reached and $\sigma_F := \min F^{-1}(\tau)$ exists. Moreover $0 < \sigma_F < +\infty$ and $0 < \tau < +\infty$.

Proof :
This corollary is a straightforward consequence of the continuity of F according to Lemma 2.1 which implies that $F^{-1}(\tau)$ is nonempty and compact. □

Lemma 2.2. Let X denote a r.vec. in $\mathbb{R}^d$ such that $\|X\|^{-1}$ is finite almost surely. Let σ be a non negative random variable and let $\mathcal{N}$ be a random vector in $\mathbb{R}^d$ with distribution $\mathcal{N}(0, I_d)$ and independent of $\sigma\|X\|^{-1}$. Assume that

$$E\Big(\ln\Big(1 + r\frac{\sigma}{\|X\|}\Big)\Big) \in O(e^{cr})$$

with a constant $c \geqslant 0$, then the expectation of $\ln^+(\|X\|^{-1}\|X + \sigma\mathcal{N}\|)$ is finite.

Proof :
Obviously $E(\ln^+(\|X\|^{-1}\|X + \sigma\mathcal{N}\|)) \leqslant E(\ln(1 + \frac{\sigma}{\|X\|}\|\mathcal{N}\|))$. Using the independency of

$\sigma\|X\|$ and $\mathcal{N}$, and passing to the spherical coordinates, one gets

$$\begin{aligned} E\big(\ln\big(1+\frac{\sigma}{\|X\|}\|\mathcal{N}\|\big)\big) &\leqslant E\Big(\int_{\mathbb{R}^d}\ln(1+\frac{\sigma}{\|X\|}\|x\|)e^{-\frac{\|x\|^2}{2}}dx\Big) \\ &= S_d E\Big(\int_0^{+\infty}\ln(1+r\frac{\sigma}{\|X\|})r^{d-1}e^{-\frac{r^2}{2}}dr\Big) \\ &= S_d\int_0^{+\infty}E(\ln(1+r\frac{\sigma}{\|X\|}))r^{d-1}e^{-\frac{r^2}{2}}dr \\ &\ll \int_0^{+\infty}r^{d-1}e^{cr-\frac{r^2}{2}}dr<+\infty\,. \end{aligned}$$

□

Remark 2.2.1. The assumption $E(\ln(1+r\frac{\sigma}{\|X\|}))\in O(e^{cr})$ (with $c=0$) is verified if there exists $\alpha>0$ such that the expectation of the r.var. $(\sigma/\|X\|)^\alpha$ is finite.

2.3 Lower bounds for the $(1+1)$-ES

In this section, we consider a general measurable objective function $f:\mathbb{R}^d\to\mathbb{R}$. We prove that the $(1+1)$-ES defined by (2.2) for minimizing f, under suitable assumptions, satisfies for all x^* in $\mathbb{R}^d$ and all indices $n\geqslant 0$:

$$-\infty<E(\ln\|\mathrm{X}_n-x^*\|)-\tau\leqslant E(\ln\|\mathrm{X}_{n+1}-x^*\|)<+\infty \tag{2.5}$$

where τ is defined in Corollary 1.

If x^* is a limit point of (X_n) (that could be a local optimum of f), (2.5) means that the expected log-distance to x^* cannot decrease more than τ, in other words, $-\tau$ is a lower bound for the convergence rate of $(1+1)$-ES. The proof of this result uses the following easy Lemma whose proof is left to the reader.

Lemma 2.3. Let Z and V be r.vec. and let Θ be any r.var. valued in $\{0,1\}$. Assume that the r.var. $\ln(\|Z\|)$ is finite almost surely. Then the following inequalities

$$\begin{aligned} \ln(\|Z\|)-\ln^-(\|Z\|^{-1}\|Z+V\|) &\leq \ln(\|Z+\Theta V\|) \\ &\leq \ln(\|Z\|)+\ln^+(\|Z\|^{-1}\|Z+V\|) \end{aligned} \tag{2.6}$$

hold almost surely.

We are ready to prove the following general theorem.

Theorem 2.4 (Lower bounds for the $(1+1)$-ES)**.** Let $(\mathrm{X}_n)_n$ be the sequence of random vectors verifying (2.2) with a given objective function f as above. Assume that for each step $n=0,1,2,\ldots$ the random vector $\mathcal{N}_n$ is independent of both the random variable σ_n

and the random vector X_n. Let x^* be any vector in $\mathbb{R}^d$ and suppose that $E(\big|\ln(\|\mathrm{X}_0 - x^*\|)\big|) < +\infty$ and for all $n \geqslant 0$,

$$E\Big(\ln(1 + r\frac{\sigma_n}{\|X_n - x^*\|})\Big) \in O(e^{c_n r})$$

with a constant $c_n \geqslant 0$. Then

$$E\left(|\ln\left(\|\mathrm{X}_n - x^*\|\right)|\right) < +\infty \ ,$$

and

$$E(\ln(\|\mathrm{X}_n - x^*\|)) - \tau \leq E(\ln(\|\mathrm{X}_{n+1} - x^*\|))\,, \tag{2.7}$$

for all $n \geqslant 0$, where τ is defined in Corollary 1. In particular, the convergence of the $(1+1)$-ES is at most linear, in the sense that

$$\inf_{n\in\mathbb{N}} \frac{1}{n} E\big(\ln\left(\|\mathrm{X}_n - x^*\|/\|\mathrm{X}_0 - x^*\|\right)\big) \geq -\tau\,. \tag{2.8}$$

Proof :
Set $Z_n = \mathrm{X}_n - x^*$, $\tilde{\mathrm{X}}_n = \mathrm{X}_n + \sigma_n\mathcal{N}_n$ and $\tilde{\mathrm{Z}}_n = \tilde{\mathrm{X}}_n - x^*$. We prove the integrability of $\ln\left(\|Z_n\|\right)$ by induction. By assumption $E\big(\ln(\|Z_0\|)\big)$ is finite. Suppose that $E\big(\ln\|Z_n\|\big)$ is finite, then $0 < \|Z_n\| < +\infty$ almost surely, hence $\ln\left(\|Z_{n+1}\|\right)$ is also finite almost surely. We claim that $E\big(\ln(\|Z_{n+1}\|)\big)$ is finite. By applying Lemma 2.3 we get (2.6) and derive

$$\ln^+\left(\|Z_{n+1}\|\right) \leq \ln^+\left(\|Z_n\|\right) + \ln^+\left(\|Z_n\|^{-1}(\|Z_n + \sigma_n\mathcal{N}_n\|)\right). \tag{2.9}$$

By Lemma 2.2 the expectation of $\ln^+\left(\|Z_n\|^{-1}(\|Z_n + \sigma_n\mathcal{N}_n\|)\right)$ is finite and using (2.9) we conclude that $E\big(\ln^+\left(\|Z_{n+1}\|\right)\big) < +\infty$. It remains to show that $E\big(\ln^-(\|Z_{n+1}\|)\big)$ is also finite. Using the first inequality in (2.6) we obtain

$$\ln^-\left(\|Z_{n+1}\|\right) \leq -\ln\left(\|Z_n\|\right) + \ln^-\left(\Big\|\frac{Z_n}{\|Z_n\|} + \frac{\sigma_n}{\|Z_n\|}\mathcal{N}_n\Big\|\right) + \ln^+\left(\|Z_{n+1}\|\right)\,. \tag{2.10}$$

For each $n \geq 0$, let $\mathcal{F}_n$ denote the σ-algebra generated by the r.vec. X_n and the r.var. σ_n. Taking the conditional expectation we obtain

$$\begin{aligned} &E[\ln^-(\|Z_{n+1}\|)\,|\,\mathcal{F}_n] \\ &\qquad \leq -\ln(\|Z_n\|) + E\Big[\ln^-\left(\Big\|\frac{Z_n}{\|Z_n\|} + \frac{\sigma_n}{\|Z_n\|}\mathcal{N}_n\Big\|\right)|\,\mathcal{F}_n\Big] + E\big[\ln^+\left(\|Z_{n+1}\|\right)|\,\mathcal{F}_n\big]\,. \end{aligned}$$

Since the distribution $\mathcal{N}_n$ is invariant under rotation and independent of $\mathcal{F}_n$,

$$\begin{aligned} E\big(\ln^-\left(\Big\|\frac{Z_n}{\|Z_n\|} + \frac{\sigma_n}{\|Z_n\|}\mathcal{N}_n\Big\|\right)|\,\mathcal{F}_n\big) &= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} \ln^-(\|e_1 + t_n x\|)e^{-\frac{\|x\|^2}{2}}dx \\ &= F(t_n) \end{aligned}$$

where e_1 is any unit vector on $\mathbb{R}^d$, $t_n = \sigma_n/\|Z_n\|$ (and F is the map introduced in Lemma 2.1). Using Lemma 2.1, we get

$E\left[\ln^-(\|Z_{n+1}\|) \mid \mathcal{F}_n\right] \leq -\ln(\|Z_n\|)+\tau+E\left[\ln^+(\|Z_{n+1}\|) \mid \mathcal{F}_n\right]$ (recall that $\tau = \max F([0,+\infty])$). Passing to the expectation we get

$$E\left[\ln^-(\|Z_{n+1}\|)\right] \leq -E\left[\ln(\|Z_n\|)\right] + \tau + E\left[\ln^+(\|Z_{n+1}\|)\right] < +\infty\,.$$

Hence $E[|\ln(\|Z_{n+1}\|)|]$ is finite for all $n \geqslant 0$. Moreover, we also get

$$E(\ln\|Z_{n+1}\|) \;\geqslant\; E(\ln\|Z_n\|) - \tau$$

and after summing such inequalities we obtain

$$E\left(\ln\left(\|Z_n\|/\|Z_0\|\right)\right) \geq -\tau n$$

and (2.8) follows. □

When x^* is a local minimum of the objective function, $E(\ln\|\mathrm{X}_n - x^*\|) - E(\ln\|\mathrm{X}_{n+1} - x^*\|)$ represents the expected log-distance reduction towards x^* at the n-th step of iteration, called *log-progress* in [17]. Theorem 2.4 shows that the log-progress is bounded above by $\tau = F(\sigma_F)$.

2.4 Spherical functions and the scale-invariant algorithm

In this section we prove that the lower bound $-\tau$ obtained in Theorem 2.4 is reached for spherical objective functions when $\sigma_n = \sigma_F\|\mathrm{X}_n\|$ $(n \geqslant 0)$. Recall that sphere objective functions are defined by $f(x) = g(\|x\|^2)$ where g is any increasing map, so that the acceptance condition $f(\mathrm{X}_{n+1}) \leqslant f(\mathrm{X}_n)$ is equivalent to $\|\mathrm{X}_{n+1}\| \leqslant \|\mathrm{X}_n\|$. It follows that $(\|\mathrm{X}_n\|)_{n\geqslant 0}$ is a non-increasing sequence of positive random variables (finite almost surely), hence converges pointwise almost surely. For spherical functions, Lemma 2.3 becomes:

Lemma 2.5. Let X and W be any random vectors and let $\Theta = \mathbf{1}_{\{f(X+W)\leqslant f(X)\}}$ and assume that the random variable $\ln(\|X\|)$ is finite almost surely. Then the equality

$$\ln(\|X+\Theta W\|) - \ln(\|X\|) = -\ln^+(\|X\|^{-1}\|X+W\|) \tag{2.11}$$

holds almost surely.

Proof :
The equality (2.11) emphasizes the fact that $\|X+\Theta\| \leqslant \|X\|$ with equality on the event $\{\Theta = 0\}$ $(= \{\|X+W\| > \|X\|\})$. □

Proposition 2.6. Let $(\mathrm{X}_n)_n$ be the sequence of random vectors valued in $\mathbb{R}^d$ satisfying the recurrence relation (2.2) involving spherical function $f(x) = g(\|x\|^2)$ where $g : [0,\infty[\to \mathbb{R}$ is an increasing map. Assume that $E(\ln(\|X_0\|)$ is finite and that, at each step n, the random vector N_n is independent of both the random variable σ_n and the random vector X_n. Then $E(\ln(\|X_n\|)$ is finite for all indices n, the inequalities

$$E(\ln(\|X_n\|) - \tau \leqslant E(\ln(\|X_{n+1}\|)$$

hold, where τ is defined above in Corollary 1, and

$$\ln(\|X_n\|) - \ln(\|X_{n+1}\|) = \ln^-(\|X_n\|^{-1}\|X_n + \sigma_n\mathrm{N}_n\|) < +\infty \ a.s. \tag{2.12}$$

Proof :
By construction $\|X_{n+1}\| \leqslant \|X_n\| \leqslant \|X_0\|$ so that $E(\ln^+(\|X_{n+1}\|)) \leqslant E(\ln^+(\|X_0\|)) < +\infty$. Now assume that $\ln(\|X_n\|)$ is integrable, hence $0 < \|X_n\| < +\infty$ a.s. and so, by Lemma 2.5, to obtain the inequalities and equality asserted in the proposition it is enough to prove that $E(\ln^-(\|X_n\|^{-1}\|X_n+\sigma_n \mathrm{N}_n\|)) \leqslant \tau$. But similarly to the end part of the proof of Theorem 2.4 we have $E(\ln^-(\|X_n\|^{-1}\|X_n + \sigma_n \mathrm{N}_n\|)) = E(F(\sigma_n/\|X_n\|)) \leqslant \tau$. □

Now we pay attention to the particular case where $\sigma_n = \sigma\|\mathrm{X}_n\|$ with $\sigma > 0$ fixed. The resulting $(1+1)$-ES is said to be *scale-invariant*, and is modeled by the d-dimensional random process

$$\mathrm{X}_{n+1} = \mathrm{X}_n + \delta_n(\mathrm{X}_n, I_\Omega)\sigma\|\mathrm{X}_n\|\mathcal{N}_n \qquad (n \geqslant 0)\,. \tag{2.13}$$

For convenience of the reader we collect the hypothesis that govern the scale-invariant random process (2.13):

> **(HSI)** *The sequence of random vectors* $(\mathrm{N}_n)_n$ *in* $\mathbb{R}^d$ *is i.i.d. with common law* $\mathrm{N}(0, I_d)$*, is independent of the initial random vector* X_0 *and* $\ln(\|X_0\|)$ *has a finite expectation.*

Notice that Assumption (HSI) implies in particular that for $m \geqslant n \geqslant 0$, N_m is independent of X_n and by Proposition 2.6, $\ln(\|X_n\|)$ has a finite expectation. The update rule (2.13) is not so realistic because in practice, at each step n, the distance of X_n to the optimum is unknown. Nevertheless, we will show that the stochastic process defined by (2.13) converges log-linearly for sphere functions and that for $\sigma = \sigma_F$ the convergence rate in log is equal to $-F(\sigma_F)$ $(= -\tau)$. In other words, the choice $\sigma_n = \sigma_F\|\mathrm{X}_n\|$ correspond to the adaptation scheme that gives the optimal convergence rate for isotropic Evolution Strategies.

It is usual for studying stochastic search algorithms to consider log-linear convergence of X_n by investigating the stability of $\ln(\|\mathrm{X}_{n+1}\|/\|\mathrm{X}_n\|)$. This idea was introduced in the context of ESs by Bienvenüe and François [27] and exploited in [13]. The process X_n given by (2.13) has a remarkable property expressed in terms of orthogonality of the random sequences $Y_n = \ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathcal{N}_n\right\|\right) - F(\sigma)$:

Proposition 2.7. Consider the random variables

$$Y_n := \ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathcal{N}_n\right\|\right) - F(\sigma)$$

where F is defined by (2.4) and let $\sigma > 0$. Under the hypothesis (HSI) the followings hold:

1. For $n \geq 0$, $E(Y_n) = 0$ and $E(|Y_n|^2) < +\infty$.
2. Let $(Y'_n)_{n\geq 0}$ be the sequence of random variables

$$Y'_n := \ln^-(\|e_1 + \sigma\mathcal{N}_n\|) - F(\sigma).$$

The random variables Y_n $(n \geq 0)$ are identically distributed and for every $n \geq 0$, Y_n and Y'_n follow the same distribution.

3. The sequence of random variables $(Y_n)_{n\geq 0}$ is orthogonal, i.e. for all indices i, j, with $i \neq j$ one has $E(Y_i) = 0$, $E(Y_i^2) < +\infty$ and $E(Y_i Y_j) = 0$.

Proof :
The isotropy of the standard d-dimensional normal distribution gives

$$\begin{aligned} E\Big(\ln^- \Big(\Big\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathcal{N}_n \Big\| \Big) \,|\, \mathrm{X}_n \Big) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \ln^-(\|e_1 + \sigma x\|) e^{-\frac{\|x\|^2}{2}} dx \\ &= F(\sigma) \end{aligned}$$

hence $E\left[\ln^- \left(\left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathcal{N}_n \right\| \right)\right] = E\left[F(\sigma)\right]$ and so, $E(Y_n) = 0$. Let $F_2 : [0,\infty] \to [0, +\infty[$ be defined by $F_2(\infty) = 0$ and, for $t \in [0, +\infty[$,

$$F_2(t) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left[\ln^-(\|e_1 + tx\|)\right]^2 e^{-\frac{\|x\|^2}{2}} dx \,. \tag{2.14}$$

Similarly to the proof of Lemma 2.1, we prove that F_2 is continuous, hence bounded. Now, from the definitions of F and F_2 one has

$$E(|Y_n|^2) = F_2(\sigma) - (F(\sigma))^2 < +\infty \,. \tag{2.15}$$

This ends the proof of the first point.

The random vectors Y_n and Y'_n have the same distribution if their characteristic functions are identical. But successively

$$\begin{aligned} E(e^{itY_n} \,|\, \mathrm{X}_n) &= e^{-itF(\sigma)} E\Big(e^{it \ln^- \left(\left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathcal{N}_n \right\| \right)} \,|\, \mathrm{X}_n \Big) \\ &= \frac{e^{-itF(\sigma)}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{it \ln^- \left(\left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma x \right\| \right)} e^{-\|x\|^2/2} dx \\ &= \frac{e^{-itF(\sigma)}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{it \ln^- \left(\left\| e_1 + \sigma x \right\| \right)} e^{-\|x\|^2/2} dx \\ &= E(e^{itY'_n}) \,. \end{aligned}$$

Therefore $E(e^{itY_n}) = E(E(e^{itY_n} \,|\, \mathrm{X}_n)) = E(e^{itY'_n})$. To finish the proof we show the orthogonality property of the Y_n $(n \geq 0)$. Let n and m be indices such that $n < m$. The random vector Y_n is $\sigma(\mathrm{X}_n, \mathcal{N}_n)$-measurable, so that

$$E(Y_m Y_n \,|\, \mathrm{X}_n, \mathrm{X}_m, \mathcal{N}_n) = Y_n E(Y_m | \mathrm{X}_n, \mathrm{X}_m, \mathcal{N}_n) \,.$$

Using the independency of $\mathcal{N}_m$ with the random vectors. X_n, $\mathcal{N}_n$ and X_m, we get

$$\begin{aligned} E(Y_m | \mathrm{X}_n, \mathrm{X}_m, \mathcal{N}_n) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \Big(\ln^- \Big(\Big\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma x \Big\| \Big) \Big) e^{-\frac{\|x\|^2}{2}} dx - F(\sigma) \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \Big(\ln^-(\|e_1 + \sigma x\|) \Big) e^{-\frac{\|x\|^2}{2}} dx - F(\sigma) = 0 \,, \end{aligned}$$

that implies $E(Y_m Y_n) = 0$. □

With the above notations define the random vectors $S_n = Y_0 + \cdots + Y_n$ and $S'_n = Y'_0 + \cdots + Y'_n$. Under the hypothesis (HSI), the characteristic function of S_n can be written as $E(itS_n) = E(E(itS_n \,|\, X_0, \mathcal{N}_0, \ldots, \mathcal{N}_{n-1}))$ and so, $E(itS_n) = E(itS'_n) = (E(itY'_0))^{n+1}$. But the random vectors Y_n are i.i.d. with expectation 0 and variance $F_2(\sigma) - F(\sigma)^2$ (see (2.15)). As a consequence, the central limit theorem holds for both $(Y_n)_n$ and $(Y'_n)_n$:

Theorem 2.8. Under the hypothesis (HSI) one has

$$\lim_{n\to+\infty} P\left(\frac{\ln(\|X_n\|) - \ln(\|X_0\|) + F(\sigma)n}{\sqrt{(F_2(\sigma) - F(\sigma)^2)n}} \leqslant t\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{u^2}{2}} du\,.$$

The pointwise stability of $\ln(\|\mathrm{X}_{n+1}\|/\|\mathrm{X}_n\|)$ is obtained by applying the following Law of Large Numbers (LLN) for orthogonal random variables (see [93, p. 458] where a more general statement is given).

Theorem 2.9 (LLN for Orthogonal Random Variables)**.** Let $(Y_n)_{n\geq 0}$ be a sequence of identically distributed real random variables with finite variance and orthogonal, *i.e.*, for all indices i, j, with $i \neq j$ one has $E(Y_i) = 0$, $E(Y_i^2) < +\infty$ and $E(Y_iY_j) = 0$. Then

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} Y_k = 0 \quad a.s.$$

We are now ready to prove the following main result

Theorem 2.10. Let $\sigma > 0$ and let $(\mathrm{X}_n)_n$ be the sequence of random vectors satisfying the recurrence relation (2.13) with $f(x) = g(\|x\|^2)$ where g is an increasing map. Assume that the hypothesis (HSI) holds. Then $(\mathrm{X}_n)_n$ converges log-linearly to the minimum, in the sense that

$$\lim_n \frac{1}{n} \ln\left(\frac{\|\mathrm{X}_n\|}{\|\mathrm{X}_0\|}\right) = -F(\sigma)(< 0) \quad \text{a.s.} \tag{2.16}$$

where F is defined by (2.4). The optimal convergence rate is obtained for $\sigma = \sigma_F := \min F^{-1}(\max F)$ (see Corollary 1).

Proof :
In case $\sigma_n = \sigma\|\mathrm{X}_n\|$ for all indices n the equality (2.12) becomes

$$\ln\|\mathrm{X}_{n+1}\| - \ln\|\mathrm{X}_n\| = -\ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathcal{N}_n\right\|\right).$$

and after summing the equations for $k = 0, \ldots, n-1$, we obtain

$$\frac{1}{n}\left(\ln\|\mathrm{X}_n\| - \ln\|\mathrm{X}_0\|\right) = -\frac{1}{n}\sum_{k=0}^{n-1} \ln^-\left(\left\|\frac{X_k}{\|X_k\|} + \sigma\mathcal{N}_k\right\|\right).$$

Proposition 2.7 and Theorem 2.9 end the proof. □

2.5 Discussion and conclusion

Theorems 2.4 and 2.10 show that optimal bounds for the convergence rate of an isotropic $(1+1)$-ES with multivariate normal distribution are reached for the scale-invariant algorithm with $\sigma_n = \sigma_F \|\mathrm{X}_n\|$ for the sphere function, where σ_F maximizes

$$F(\sigma) = E(\ln^- \|e_1 + \sigma\mathcal{N}\|) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \ln^-(\|e_1 + \sigma x\|) e^{-\frac{\|x\|^2}{2}} dx \ .$$

From (2.12) and from the isotropy of the multivariate normal distribution $\mathcal{N}$, it follows that finding σ maximizing F amounts to finding σ maximizing the log-progress $E(\ln \|\mathrm{X}_n\|) - E(\ln \|\mathrm{X}_{n+1}\|)$.

Most of the works based on the progress rate, consist in finding σ maximizing estimations of the expected progress $E(\|\mathrm{X}_n\|) - E(\|\mathrm{X}_{n+1}\|)$ (when d goes to infinity) [114, 25]. Note that the definition of progress in those works does not consider $\ln \|\mathrm{X}_n\|$ and so is different from the one underlying our study. Assuming that both definitions matches[3], our results give an interpretation of this approach in terms of lower bounds for convergence of ESs.

The lower bounds derived in this paper are tight. Consequently they can be used in practice to assess the performances of a given step-size adaptation strategy comparing the convergence rate achieved by the strategy with the optimal one, given by the scale-invariant algorithm.

The numerical estimation of the optimal convergence rate $-\tau$ can be achieved with a Monte Carlo integration: for different σ, $F(\sigma)$ equals the expectation $E(\ln^- \|e_1 + \sigma\mathcal{N}\|)$. This expectation can be estimated by summing independent samplings of the random variable $\ln^- \|e_1 + \sigma\mathcal{N}\|$. This is illustrated in Fig 2.1.

The analysis of the log-linear convergence carried out in this paper relies on the application of the Strong Law of Large Numbers for orthogonal random variables. This study uses deeply the invariance under rotations of the standard d-dimensional multivariate normal distribution and does not cover directly the usual case of stable Markov chains that will be investigated in future works.

Acknowledgments

The authors thank the referees for their constructive remarks on the previous version that lead to this new version and are very grateful to Nicolas Monmarché for his encouragements. This work receives partial supports from the ANR/RNTL project Optimisation Multidisciplinaire (OMD) and from the ACI CHROMALGEMA.

[3]This will be true asymptotically in the dimension d, though we do not prove it rigorously in this paper.

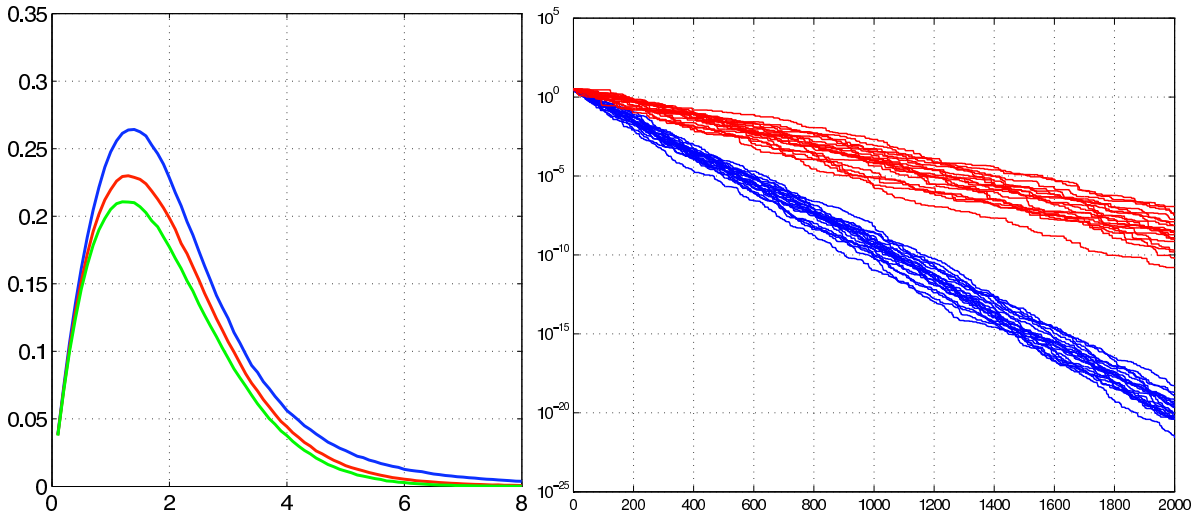


Figure 2.1: Left: Plot of the function $\sigma \mapsto dF(\sigma/d)$ (Eq. (2.4)) versus σ for $d = 5$ (resp. 10, 30) and $0 \leq \sigma \leq 8$. The upper curve corresponds to $d = 5$, the middle one to $d = 10$ and the lower one to $d = 30$. Note that the function F defined in (2.4) implicitly depends on d. Using the more explicit notation F_d instead of F, the plots represent actually $\sigma \mapsto dF_d(\sigma/d)$. For $d = 10$, we see that σ_F maximizing F (defined in Corollary 1) approximately equals 0.13. The plots were obtained doing Monte Carlo estimations of F using 10^6 samples.
Right: Twenty realizations of the scale-invariant algorithm on the sphere function for $d = 10$. The y-axis shows the distance to the optimum (in log-scale) and the x-axis the number of iterations n. The twenty curves below correspond to the optimal algorithm, *ie.* $\sigma_n = \sigma_F \|\mathrm{X}_n\|$ for all n where σ_F equals to 0.13 (value maximizing the curve of F on the left for $d = 10$). The twenty curves above correspond to 20 realizations of the scale-invariant algorithm for $\sigma_n = 0.3\|\mathrm{X}_n\|$. Observed, the log-linear convergence as well as the optimality of the scale-invariant algorithm for $\sigma = \sigma_F$.

Chapter 3

Study of the Scale-invariant $(1+1)$-ES in Noisy Spherical Environments

In real-world optimization problems, objective functions are noisy. The noise can stem from physical measurement limitations or Monte Carlo simulations .... In Chapter 2, we have established that the scale-invariant $(1+1)$-ES converges log-linearly when minimizing sphere functions. The goal of this chapter is to see how the behavior of the $(1+1)$-ES is affected when the sphere function is disturbed by noise. We investigate the $(1+1)$-ES with the artificial scale-invariant adaptation rule because this rule is optimal in the case of non-noisy spherical functions, as shown in Chapter 2. The noise model investigated here is multiplicative, i.e., the noisy objective function result from the multiplication of the non-noisy objective function by the random variable $1+\mathcal{N}$ where $\mathcal{N}$ is the noise random variable. Theoretical studies of minimization of noisy objective functions using ES have been mainly performed by Arnold and Beyer [23, 24, 7, 25, 5, 8, 10]. These studies rely on the approximation of an infinite dimension of the search space and use classical normalizations previously used in the field of progress rate theory for the step-size of the mutation and the progress rate. Moreover, Arnold and Beyer used an additional normalization for the noise strength which represents the variance of the random variable $\mathcal{N}$.

The chapter is composed of three parts:

The first part (Section 3.1) is the paper [76], that has been published in the proceedings of the conference *Parallel Problem Solving From Nature (PPSN 2008)*. The noisy sphere function model used in this part is similar to the one studied by Arnold and Beyer in [23, 24, 7, 25, 5, 8, 10]. It can be written as

$$\mathcal{F}_s(x) = \|x\|^2(1+\mathcal{N}) \tag{3.1}$$

where we assume that the random variable $\mathcal{N}$ has a finite expectation such that $E(\mathcal{N}) > -1$ and admits a density function $p_{\mathcal{N}}$ which lies in the range $[m_{\mathcal{N}}, M_{\mathcal{N}}[$ ($-\infty < m_{\mathcal{N}} < M_{\mathcal{N}} \leq +\infty$, $M_{\mathcal{N}} > -1$ and $m_{\mathcal{N}} \neq -1$). Arnold and Beyer's model is similar (see Eq. 3.2) except that they used a normalization for the variance of the noise. Moreover, most of the studies of Arnold and Beyer use the assumption of Gaussian noise. A notable exception is the study in [10] which investigates the behavior of a class of ES using recombination,

under the assumption of a general noise distribution. In this paper, we prove (Theorem 3.1), that the behavior of the scale-invariant $(1+1)$-ES minimizing the noisy sphere (Eq. 3.1) depends on the infimum of the (support of the) noise $m_{\mathcal{N}}$. More precisely, we prove that the sequence of solutions generated by the algorithm converges almost surely to zero if $m_{\mathcal{N}} > -1$ and diverges to infinity when $-\infty < m_{\mathcal{N}} < -1$. The result is demonstrated using the Borel-Cantelli Lemma (Lemma 3.2). The study does not include the case $m_{\mathcal{N}} = -1$.
The second part (Section 3.2) is the main material for a paper that we intend to submit soon. The goal of this part is to see if the log-linear behavior that we have proved in Chapter 2 for (non-noisy) sphere functions also holds for noisy spherical functions. This second part uses the same context as the first part (Section 3.1), i.e., the same noisy objective function given by Eq. 3.1 and a scale-invariant $(1+1)$-ES. Therefore, the main result derived in Theorem 3.1 is also used in this part. It is shown (Theorem 3.18) that the convergence (if $m_{\mathcal{N}} > -1$) or divergence (if $-\infty < m_{\mathcal{N}} < -1$) of the (1+1)-ES minimizing the noisy sphere (Eq. 3.1) holds in the sense: $\frac{1}{n}\ln\|\mathrm{X}_n\|$ converges in probability to γ (see Eq. 3.20) where γ is finite and $(\mathrm{X}_n)_n$ is the solution of the algorithm at an iteration n defined in Eq. 3.5. However, according to the definition of the log-linear behavior given in Eq. 3.7, one has to show that $\gamma \neq 0$ which is not proven in our study.

The main result of this part (Theorem 3.18) has been established using the Law of Large Numbers (LLN) for Markov chains (Theorem 3.12).
The third part (Section 3.3) is made of some additional theoretical results that were not included in paper [76] that they generalize. They are related to 'spatial' convergence (or divergence) of the scale-invariant $(1+1)$-ES for the shifted noisy sphere function defined by $\mathcal{F}_\alpha(x) = (\|x\|^2 + \alpha)(1+\mathcal{N})$ where α is a positive constant [4]. Moreover, in these studies, non lower-bounded noise distributions, i.e., $m_{\mathcal{N}} = -\infty$, are also investigated (In [76], only lower bounded noise distributions had been investigated). Therefore our study includes the particular case of Gaussian noise that has been investigated by Arnold and Beyer. It is shown in Section 3.3.1 that if $m_{\mathcal{N}} > -1$ the algorithm converges. However, if $-\infty \leq m_{\mathcal{N}} < -1$, it is shown in Section 3.3.2 that the algorithm cannot converge (in the sense that the L^2-norm of the distance to the optimum of the noiseless part of the objective function cannot converge to zero), as negative objective function values are sampled after a finite number of iterations.

Comparison with results in [8] In [8], the scale-invariant $(1+1)$-ES has been investigated using the following model of noisy sphere function:

$$\mathrm{f}(x) = \|x\|^2 + \frac{2\sigma_\epsilon^*}{d}\|x\|^2\mathrm{N}(0,1) \tag{3.2}$$

where d is the search space dimension, σ_ϵ^* is a strictly positive constant called the normalized noise strength and $\mathrm{N}(0,1)$ is the Gaussian random variable with mean 0 and variance 1. The expected progress rate computed in [8] is positive and convergence occurs for all σ_ϵ^* values. On the other hand, our theoretical study shows (see Section 3.3.2) that for noise distributions with $m_{\mathcal{N}} = -\infty$, which is the case of a Gaussian noise, no convergence

[4] For $\alpha = 0$, $\mathcal{F}_\alpha(x)$ simplifies to $\mathcal{F}_s(x)$ defined in Eq. 3.1.

occurs. This result is also illustrated by experimental observations: In Fig. 3.3, it can be seen that divergence happens for sufficiently large noise strength values. Therefore our results may seem in contradiction with Arnold and Beyer's results. The reason for this apparent contradiction is that, in [8], the expression $\frac{2\sigma_\epsilon^*}{d}$ for the noise level implies a small noise strength for large search space dimensions. For example, in [8, Fig 8], and for $\sigma_\epsilon^* = 2$ and $d = 80$, the noisy sphere function can be written as $\mathrm{f}(x) = \|x\|^2(1 + 0.05\mathrm{N}(0, 1))$. Therefore, the probability to sample a negative fitness, which is the event that leads to non convergence, is upper bounded by 10^{-88}. Sampling a negative fitness value is then an event that will 'never' happens in practical simulations as it has a probability less that 10^{-88} to happen, and, the algorithm is observed to converge.

Future work Our study can be completed by investigating the case $m_{\mathcal{N}} = -1$ which was not solved here. Moreover, in the second part of the study, we have only shown that converge rates (for $m_{\mathcal{N}} > -1$) and divergence rates (for $-\infty < m_{\mathcal{N}} < -1$) is positive or negative without excluding the case of null convergence or divergence rate to prove the log-linear behavior as defined in Eq. 3.7. Fortunately, the convergence rate given in Eq. 3.7 can be easily computed using Monte Carlo simulations. Therefore, one has to compute numerically this convergence rate. It seems that the case $m_{\mathcal{N}} = -1$ is equivalent to having $\frac{1}{n} \ln \|\mathrm{X}_n\| \to 0$ in Eq. 3.20. Furthermore, the convergence established in Eq. 3.20 holds in probability and one has to investigate almost sure convergence in this equation. Another issue to clarify is the reliability of an approximation that has been done in [8], stating that an offspring and its parent have similar noise levels in large dimensions. For comma strategies, we confirm in Chapter 4 that such an approximation is reliable, but in the limit of infinite dimension of the search space.

The (1+1)-ES with reevaluation of the parent, and link with Chapter 4 In this chapter, we investigate the behavior of the $(1 + 1)$-ES when minimizing noisy objective functions with positive ideal function values. The $(1 + 1)$-ES does not converge for noise distributions allowing the sampling of negative fitness values and for the specific scale-invariant adaptation rule. In fact, after a certain number of iterations, a strictly negative objective function value will happens almost surely. Then, as the selection scheme used in the $(1 + 1)$-ES is elitist, the sequence of (negative) fitness functions decreases and will probably have as a limit $-\infty$. The same reasoning applies for a $(1 + \lambda)$-ES where $\lambda \geq 1$. This means that increasing the number of offspring λ is not a solution to avoid divergence. To avoid divergence cases, an alternative is to use the $(1 + 1)$-ES with a reevaluation of the parent in the selection step [25, 8]. Another possible solution is to use a non elitist ES such as the $(1, \lambda)$-ES which will be investigated in the next chapter. Note that the behavior of the $(1 + 1)$-ES with reevaluation will be very similar to a $(1, 2)$-ES especially for high dimensions of the search space, as suggested by relative progress rates computed in [25]. Moreover, the study that we present in the next chapter (Chapter 4) uses the LLN for orthogonal random variables and the same techniques can also be applied for the variant of $(1 + 1)$-ES reevaluating the parent.

Mohamed Jebalia[1] and Anne Auger[1,2]

[1] TAO Team, INRIA Saclay
Université Paris Sud, LRI, 91405 Orsay cedex, France
[2] Microsoft Research-INRIA Joint Centre
28, rue Jean Rostand, 91893 Orsay Cedex, France
mohamed.jebalia@lri.fr, anne.auger@inria.fr

On Multiplicative Noise Models for Stochastic Search

Proceedings of Parallel Problem Solving from Nature 2008, pp 52-61.
The original publication will be available at
http://www.springerlink.com

Errata :
All over Section 3.1 of the chapter, the quantity $m_{\mathcal{N}}$ should be referred to as the infimum of the support of the noise and not the lower bound of the noise (even if $m_{\mathcal{N}} = -\infty$). This implies, in the abstract of the first part of the chapter for example, that the sentence "... the $(1+1)$-ES diverges when the lower bound allows to sample negative fitness ..." should write "... the $(1+1)$-ES diverges when the infimum of the support of the noise distribution allows to sample negative fitness ...". Similarly, the quantity $M_{\mathcal{N}}$ should be referred to as the supremum (which can be infinite) of the support of the noise instead of upper bound of the noise.
In Lemma 3.6, an additional hypothesis is necessary to establish the result. we have to suppose that : for all $n \geq 0$, the random vectors U_n and N_n are independent. Moreover, there are two errata in the second paragraph of the conclusion of the published paper:

- In the second paragraph of the conclusion, in the sentence "... the normalization of the standard deviation of the noise implies a so small probability to sample $1 + \mathcal{N}$

below $-1 \ldots$", one should have "$1+\mathcal{N}$ below 0" instead of $1+\mathcal{N}$ below -1

- In the second paragraph of the conclusion, in the sentence "$\ldots$ where the standard deviation of 0.1 corresponds to a probability to have $(1+0.1\,\mathcal{N}) < 0$ lower bounded by 10^{-23}.", one should have "upper bounded" instead of "lower bounded".
- In the abstract of the paper "dimensionality" should be replaced by "dimension".
- A spelling mistake in Section 3.1.2 (paragraph Experimental observations): "respecitvely" should be written as "respectively".
- A spelling mistake in the sketch of the proof of Proposition 3.4: "striclty" should be written as "strictly".
- In the beginning of the Section "Mathematical model for the $(1+1)$-ES", "perturbated" should be written as "perturbed".
- The word "independency" at the end of the proof of Lemma 3.5 in the Section Appendix and the word "independance" in the proof of Lemma 3.6 in the Section Appendix should be written as 'independence".

3.1 On Multiplicative Noise Models for Stochastic Search

On Multiplicative Noise Models for Stochastic Search

Mohamed Jebalia[1] and Anne Auger[1,2]

[1] TAO Team, INRIA Saclay
Université Paris Sud, LRI, 91405 Orsay cedex, France
[2] Microsoft Research-INRIA Joint Centre
28, rue Jean Rostand, 91893 Orsay Cedex, France
mohamed.jebalia@lri.fr, anne.auger@inria.fr

Abstract

In this paper we investigate multiplicative noise models in the context of continuous optimization. We illustrate how some intrinsic properties of the noise model imply the failure of reasonable search algorithms for locating the optimum of the noiseless part of the objective function. Those findings are rigorously investigated on the $(1+1)$*-ES for the minimization of the noisy sphere function. Assuming a lower bound on the support of the noise distribution, we prove that the* $(1+1)$*-ES diverges when the lower bound allows to sample negative fitness with positive probability and converges in the opposite case. We provide a discussion on the practical applications and non applications of those outcomes and explain the differences with previous results obtained in the limit of infinite search-space dimensionality.*

3.1.1 Introduction

In many real-world optimization problems, objective functions are perturbed by noise. Evolutionary Algorithms (EAs) have been proposed as effective search methods in such contexts [9, 79]. A noisy optimization problem is a rather general optimization problem where for each point x of the search space, we can observe $f(x)$ perturbed by a random variable or in other words for a given x we can observe a distribution of possible objective values. The goal is in general to converge to the minimum of the averaged value of the

observed random variable. One type of noise encountered in real-world problems is the so-called multiplicative noise where the noiseless objective function $f(x)$ is perturbed by the addition of a noise term proportional to f, ie. the noisy objective function $\mathcal{F}$ reads

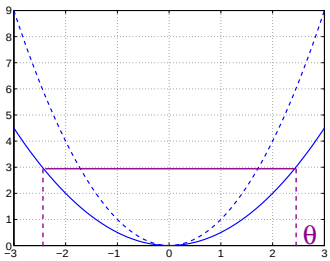
$$\mathcal{F}(x) = f(x)(1+\mathcal{N}) \tag{3.3}$$

where $\mathcal{N}$ is the noise random variable, sampled independently at each new evaluation of a point. Such noise models are in particular used to benchmark robustness of EAs with respect to noise [134]. The focus here is continuous optimization (that will be minimization) where f maps a continuous search space, ie. a subset of $\mathbb{R}^d$, into $\mathbb{R}$. The EAs specifically designed for continuous optimization are usually referred as Evolution Strategies (ES), where a set of candidate solutions evolves by first applying Gaussian perturbations (mutations) to the current solutions then selection. ES in noisy environments have been studied by Arnold and Beyer [25, 7, 5]. Multiplicative noise has been investigated in the case of $\mathcal{N}$ being normally distributed with a standard deviation scaled by $1/d$ for a $(1+1)$-ES [8], (μ, λ)-ES [7, 24], $(\mu/\mu_{\text{I}}, \lambda)$-ES [6] and f being the sphere function $f(x) = \|x\|^2$. Under the assumption that d goes to infinity, Arnold and Beyer show, for $f(x) = \|x\|^2$, positive expected fitness gain for the elitist $(1+1)$-ES (if the fitness of the parent is not reevaluated in the selection step which is the case of our study). This implies a decrease of the expectation of the square distance to the optimum (here zero). However, convergence of the $(1+1)$-ES to the optimum of the noiseless part of the noisy objective function seems to be unlikely if the noise random variable takes values smaller than -1 as we illustrate now on a simple example. Assume indeed that $\mathcal{N}$ takes three distinct values (each with probability 1/3) $+\gamma$, 0 and $-\gamma$ where γ satisfies $\gamma > 1$. For a given $x \in \mathbb{R}^d$, the objective function $\mathcal{F}(x)$ takes 3 different values (each with probability 1/3) $(1+\gamma)\|x\|^2$, $\|x\|^2$, $(1-\gamma)\|x\|^2$. The last term is strictly negative for x non equal to zero. Therefore, if one negative objective function value is reached, the $(1+1)$-ES that can only accept solutions having a lower objective function value will never accept solutions closer to the optimum since they have higher objective function values[1]. On the contrary the $(1+1)$-ES will diverge log-linearly [2], i.e. the logarithm of the distance to the optimum will increase linearly.

Starting from this observation, we investigate how the properties of the support of the noise distribution relate to convergence or divergence of stochastic search algorithms and can make the convergence to the optimum of the noiseless part of the objective function hopeless for reasonable search algorithms. Compared to previous approaches, we do not make use of asymptotic assumptions, trying to capture effects that were not observed before [8]. In Section 3.1.2, we detail the noise model considered and show experimentally on a $(1+1)$-ES that divergence and convergence is determined by the probability to sample noise values smaller than -1. In Section 3.1.3, we provide some simple proofs of convergence and divergence for the $(1+1)$-ES. In Section 3.1.4 we discuss the results and explain where the difference with the results in [8] stems from.

[1] Their absolute value is smaller though. However, trying to minimize the absolute value of $\mathcal{F}$ instead is not a solution in general, consider for instance the function $f(x) = (\|x\|^2 + 1)(1+\mathcal{N})$.

[2] We will say that a sequence $(d_n)_n$ diverges (resp. converges) log-linearly if there exists $c > 0$ (resp. $c < 0$) such that $\lim_n \frac{1}{n} \ln(d_n) = c$.



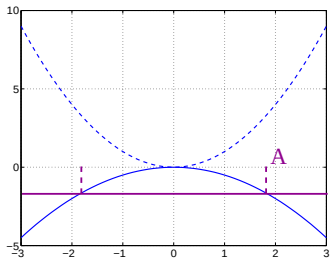


Figure 3.1: [Dashed Line] One dimensional cut of $f(x) = \|x\|^2$ along one arbitrary unit vector. [Straight line] Left: One dimensional cut of $g_{-0.5}(x) = \|x\|^2(1-0.5)$. Right: One dimensional cut of $g_{-1.5}(x) = \|x\|^2(1-1.5)$. For a given x, the noisy-objective function can, in particular, take any value between the dashed curve and the straight curve.

3.1.2 Motivations

Elementary remarks on the noise model We investigate multiplicative noise models as defined in Eq. 3.3 where $\mathcal{N}$ is a random variable with finite mean and $f(x)$ is the noiseless function that we assume positive in the sequel. We also assume that $1+E(\mathcal{N}) > 0$ such that the argmin[3] of the expected value of $\mathcal{F}(x)$ is the argmin of $f(x)$. Often, the distribution of $\mathcal{N}$ is assumed symmetric, implying then that $1 + E(\mathcal{N}) = 1 > 0$. Though one might think that this condition is sufficient such that minimizing $\mathcal{F}(x)$ amounts to minimizing $f(x)$, we sketch now, why divergence to ∞ of the distance to the optimum happens if $1 + \mathcal{N}$ can take negative values.

Assume that $f(x)$ converges to infinity when $\|x\|$ goes to ∞; typically $f(x)$ can be the famous sphere function $f(x) = \|x\|^2$ and assume that the random variable $\mathcal{N}$ admits a density function $p_{\mathcal{N}}(t), t \in \mathbb{R}$ whose support is an interval $[m_{\mathcal{N}}, M_{\mathcal{N}}[$, i.e. $\mathcal{N} \in [m_{\mathcal{N}}, M_{\mathcal{N}}[$ and the probability that $\mathcal{N} \in [a, b]$ for any $m_{\mathcal{N}} \leq a < b \leq M_{\mathcal{N}}$ is strictly positive. The function $g_{m_{\mathcal{N}}}(x) = f(x)(1+m_{\mathcal{N}})$ gives a lower bound of the values that can be reached by the noisy fitness function for different instantiations of the random variable $\mathcal{N}$ (because f is positive). For a given x, $\mathcal{F}(x)$ can take values with positive probability in any open interval of $]g_{m_{\mathcal{N}}}(x), f(x)[$ [(4)].

In Fig. 3.1 are depicted a cut of $f(x) = \|x\|^2$ and $g_{m_{\mathcal{N}}}(x) = f(x)(1 + m_{\mathcal{N}})$ for $m_{\mathcal{N}}$ equals -0.5 and -1.5. The position of $m_{\mathcal{N}}$ with respect to -1 determines whether $g_{m_{\mathcal{N}}}(x)$ is convex or concave: for $m_{\mathcal{N}} > -1$, $g_{m_{\mathcal{N}}}(x)$ is convex, converging to infinity when $\|x\|$ goes to ∞ and for $m_{\mathcal{N}} < -1$, $g_{m_{\mathcal{N}}}(x)$ is concave, converging to minus infinity when $\|x\|$ goes to ∞. Minimizing $g_{m_{\mathcal{N}}}(x)$ in the case of $m_{\mathcal{N}} < -1$ means that $\|x\|$ is diverging to $+\infty$ and $g_{m_{\mathcal{N}}}(x)$ is diverging to $-\infty$ which is the opposite of the behavior one would like since we are aiming at minimizing the non-noisy function $f(x) = \|x\|^2$. Note that in the example sketched in the introduction with $\mathcal{N}$ taking the values γ, $-\gamma$ and 0, the plot of $\|x\|^2$ and $(1-\gamma)\|x\|^2$ for $\gamma = 1.5$ are the curves represented in Fig 3.1 (right).

Experimental observations We investigate now numerically how the "shape" of the lower bound might affect the convergence. For this purpose we use a $(1, 5)$-ES and a

[3]The argmin of an objective function $x \mapsto h(x)$ are defined as $h(\arg\min_x h) = \min_x h(x)$

[4]Note that $g_{m_{\mathcal{N}}}(x) < f(x)$ iff $m_{\mathcal{N}} < 0$.

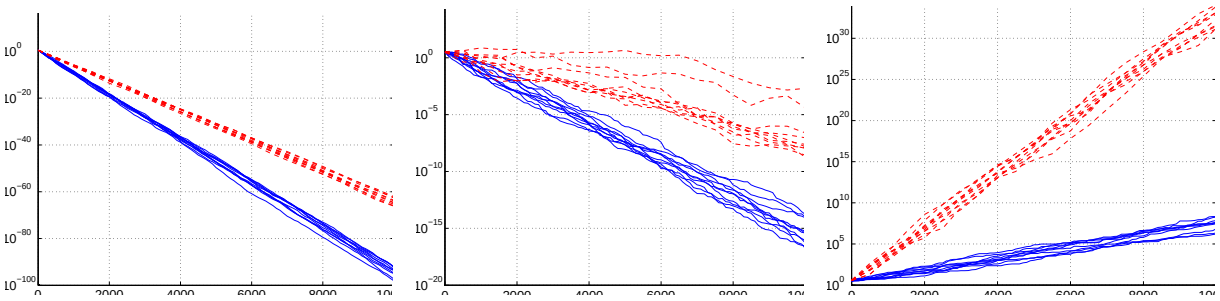


Figure 3.2: Distance to the optimum (in log-scale) versus number of evaluations. Ten independent runs for the scale-invariant $(1,5)$-ES (10 upper curves of each graph) and $(1+1)$-ES (10 lower curves of each graphs) with $d = 10$ and $\sigma = 1/d$. Left: $f(x) = \|x\|^2$. Middle: $f(x) = \|x\|^2(1 + U_{[-0.5,0.5]})$. Right: $f(x) = \|x\|^2(1 + U_{[-1.5,1.5]})$.

$(1+1)$-ES using scale-invariant adaptation scheme for the step-size[5].

We investigate the function $\mathcal{F}_s(x) = \|x\|^2(1+\mathcal{N})$ when the noise $\mathcal{N}$ is uniformly distributed in the ranges $[-0.5, 0.5]$ and $[-1.5, 1.5]$ respecitvely denoted $U_{[-0.5,0.5]}$ and $U_{[-1.5,1.5]}$. This latter noise corresponds to the concave lower bound $g_{-1.5}(x) = -0.5\|x\|^2$ plotted in Fig. 3.1. In Figure 3.2, the result of 10 independent runs of the $(1,5)$-ES (10 upper curves of each graph) in dimension $d = 10$ are plotted for the non-noisy sphere (left), $f(x) = \|x\|^2(1 + U_{[-0.5,0.5]})$ (middle) and $f(x) = \|x\|^2(1 + U_{[-1.5,1.5]})$ (right). Not too surprisingly, we observe a drastic difference in the last two cases: the algorithm converges to the optimum for the noise $U_{[-0.5,0.5]}$ whereas the distance to the optimum increases (log)-linearly for the noise having a lower bound smaller than -1 [6]. Comparing the left and middle graphs we also observe, as expected, that the presence of noise slows down the convergence. On the same figure (lower curves of the graphs), the results of 10 independent runs of the $(1+1)$-ES are plotted for the three same functions. As in the case of the comma strategy we observe that the $(1+1)$-ES diverges in the case of the noise $U_{[-1.5,1.5]}$ and that, when convergence occurs, the convergence rate is slower in presence of noise. Last, we investigate numerically the $(1+1)$-ES where $\mathcal{N}$ is normally distributed and in particular unbounded. This corresponds to the case investigated in [8]. We carry out tests for a standard deviation of the Gaussian noise equals 0.1, 2 and 10. Results are presented in Fig. 3.3. We observe convergence when the standard deviation of the noise equals 0.1 and divergence in the last two cases.

[5] In a scale-invariant ES, the step-size is set at each iteration as a (strictly positive) constant σ times the distance to the optimum. This artificial adaption scheme (since in practice one does not know the distance to the optimum!) allows to achieve optimal convergence rate for ES and is therefore very interesting from a theoretical point of view. The algorithm is mathematically defined in Section 3.1.3.

[6] However, contrary to what we will see for the $(1+1)$-ES, we do not state that "-1" is a limit value between convergence and divergence in the case of $(1,\lambda)$-ES. Indeed convergence and divergence depends on the intrinsic properties of the noise and on λ and σ as well (see [25]).

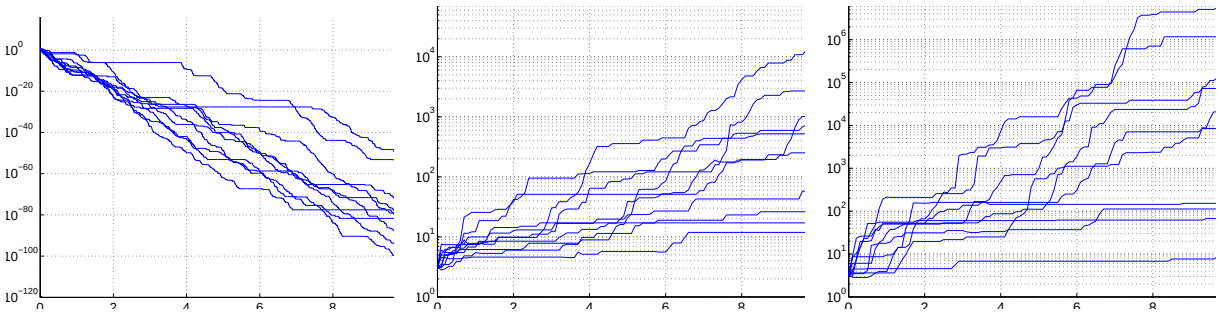


Figure 3.3: Ten independent runs for the scale-invariant $(1+1)$-ES with a normally distributed noise: on $f(x) = \|x\|^2(1+\sigma_\epsilon \mathcal{N}(0,1))$ with σ_ϵ equals 0.1 (left), 2 (middle) and 10 (right) for $d = 10$ and $\sigma = 1/d$.

3.1.3 Convergence and divergence of the $(1+1)$-ES

In this section, we provide a simple mathematical analysis of the convergence and divergence of the $(1+1)$-ES experimentally observed in the previous section. We focus for the sake of simplicity on lower bounded noise, i.e. the support of the noise is included in $[m_{\mathcal{N}}, +\infty[$. We prove that the $(1+1)$-ES minimizing the noisy sphere converges if $m_{\mathcal{N}} > -1$ and diverges if $m_{\mathcal{N}} < -1$. The proofs are rather simple and rely on the Borel-Cantelli Lemma. For the sake of readability we provide here a sketch of the demonstrations and send the proofs with the technical details in the Appendix of the paper.

Mathematical model for the $(1+1)$-ES

The $(1+1)$-ES is a simple ES which evolves a single solution. At an iteration n, this solution denoted X_n, is called parent. The minimization of a given function f mapping $\mathbb{R}^d$ $(d \geq 1)$ into $\mathbb{R}$ using the $(1+1)$-ES algorithm is as follows: At every iteration n, the parent X_n is perturbated by a Gaussian random variable $\sigma_n \mathrm{N}_n$, where σ_n is a strictly positive value called step-size and $(\mathrm{N}_n)_n \in \mathbb{R}^d$ are independent realizations of a multivariate isotropic normal distribution on $\mathbb{R}^d$ denoted by $\mathrm{N}(0, I_d)$ [7]. The resulting offspring $\mathrm{X}_n + \sigma_n \mathrm{N}_n$ is accepted if and only if its fitness value is smaller than the one of its parent X_n. One of the key points in minimization using isotropic ES[8] is how to adapt the sequence of step-sizes (σ_n). Convergence of the $(1+1)$-ES is sub-log-linear bounded below by an explicit log-linear rate. This lower bound for the convergence rate is attained for the specific case of the sphere function and scale-invariant algorithm where the step-size is chosen proportional to the distance to the optimum, i.e. $\sigma_n = \sigma\|\mathrm{X}_n\|$ where σ is a strictly positive constant [17, 77]. The scale-invariant algorithm has a major place in the theory of ES since it corresponds to the dynamic algorithm implicitly studied in the one-step analysis computing progress rate or fitness gain [113, 25]. Using this adaptation scheme, the algorithm is referred to as the scale-invariant $(1+1)$-ES and the offspring writes as

[7] $\mathrm{N}(0, I_d)$ is the multivariate normal distribution with mean $(0, \ldots, 0) \in \mathbb{R}^d$ and covariance matrix the identity I_d.

[8] ES are called isotropic when the covariance matrix of the distribution of the random vectors $(\mathrm{N}_n)_n$ is I_d.

$\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n$. The noisy sphere function is denoted

$$\mathcal{F}_s(x) = \|x\|^2(1+\mathcal{N}) \tag{3.4}$$

where we assume that the random variable $\mathcal{N}$ has a finite expectation such that $E(\mathcal{N}) > -1$ and admits a density function $p_{\mathcal{N}}$ which lies in the range $[m_{\mathcal{N}}, M_{\mathcal{N}}[$ where $-\infty < m_{\mathcal{N}} < M_{\mathcal{N}} \leq +\infty$, $M_{\mathcal{N}} > -1$ and $m_{\mathcal{N}} \neq -1$. The normalized noisy part $\mathcal{N}$ of the noisy sphere function will be called normalized overvaluation of x. The term normalized overvaluation was already defined in [8] where it corresponds to the opposite of the quantity considered here up to a factor $d/2$. The minimization of this function using the scale-invariant (1+1)-ES is mathematically modeled by the sequence of parents (X_n) with their relative noisy objective functions $(\mathcal{F}_s(\mathrm{X}_n))$ and normalized overvaluations (O_n). At an iteration n, the fitness of the parent is $\mathcal{F}_s(\mathrm{X}_n) = \|\mathrm{X}_n\|^2(1+O_n)$ and the fitness of an offspring equals $\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n\|^2(1+\mathcal{N}_n)$ where $(\mathcal{N}_n)_n$ is a sequence of independent random variables with $\mathcal{N}$ as a common law. Let $\mathrm{X}_0 \in \mathbb{R}^d$ be the first parent with a normalized overvaluation O_0 sampled from the distribution of $\mathcal{N}$. Then the update of X_n for $n \geq 0$ writes as:

$$\begin{aligned} \mathrm{X}_{n+1} &= \mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n \text{ if } \|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n\|^2(1+\mathcal{N}_n) < \|\mathrm{X}_n\|^2(1+O_n)\ , \\ &= \mathrm{X}_n \text{ otherwise}\,, \end{aligned} \tag{3.5}$$

and the new normalized overvaluation O_{n+1} is then:

$$\begin{aligned} O_{n+1} &= \mathcal{N}_n \text{ if } \|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n\|^2(1+\mathcal{N}_n) < \|\mathrm{X}_n\|^2(1+O_n)\ , \\ &= O_n \text{ otherwise}\,. \end{aligned} \tag{3.6}$$

The $(1+1)$-ES algorithm ensures that the sequence relative to the function to minimize (which is $(\mathcal{F}_s(\mathrm{X}_n))$ in our case) decreases. This property makes the theoretical study of the $(1+1)$-ES easier than that of comma strategies. Our study shows that the behavior of the scale-invariant $(1+1)$-ES on the noisy sphere function (3.4) depends on the lower bound of the noise $m_{\mathcal{N}}$.

Theorem 3.1. The $(1+1)$-ES minimizing the noisy sphere (Eq. 3.4) defined in Eq. 3.5 converges to zero if $m_{\mathcal{N}} > -1$ and diverges to infinity when $m_{\mathcal{N}} < -1$.

Proof :

The proof of this theorem is split in two cases $m_{\mathcal{N}} > -1$ and $m_{\mathcal{N}} < -1$ respectively investigated in Proposition 3.3 and Proposition 3.4. □

The proofs heavily rely on the second Borel-Cantelli Lemma that we recall below. But first, we need a formal definition of 'infinitely often (i.o.)': Let q_n be some statement, eg. $|a_n - a| > \epsilon$. We say (q_n i.o.) if for all n, $\exists\ m \geq n$ such that q_m is true. Similarly, for a sequence of events A_n in a probability space, (A_n i.o.) equals $\{w | w \in A_n \text{ i.o.}\} = \cap_{n\geq 0} \cup_{m \geq n} A_m := \overline{\lim}\ A_n$. The second Borel-Cantelli Lemma (BCL) states that:

Lemma 3.2. Let $(A_n)_{n\geq 0}$ be a sequence of events in some probability space. If the events A_n are independent and verify $\sum_{n\geq 0} P(A_n) = +\infty$ then $P(\overline{\lim}\ A_n) = 1$.

Proposition 3.3 (Convergence for $m_{\mathcal{N}} > -1$)**.** If $m_{\mathcal{N}} > -1$, the sequences $(\mathcal{F}_s(\mathrm{X}_n))$ and $(\|\mathrm{X}_n\|)$ converge to zero almost surely.

Sketch of the proof (see detailed proof in Appendix) The condition $m_{\mathcal{N}} > -1$ ensures that the decreasing sequence $(\mathcal{F}_s(\mathrm{X}_n))$ is positive. Therefore it converges. Besides the sequence $(\|\mathrm{X}_n\|)$ is upper bounded by $\theta := \mathcal{F}_s(X_0)/(1 + m_{\mathcal{N}})$ as shown in Fig. 3.1 (left). Consequently, the probability to hit, at each iteration n, a fixed neighborhood of 0 is lower bounded by a strictly positive constant. Applying BCL we deduce the convergence of the sequence $(\mathcal{F}_s(\mathrm{X}_n))$ (and then that of $(\|\mathrm{X}_n\|)$) to zero. □

Proposition 3.4 (Divergence for $m_{\mathcal{N}} < -1$)**.** If $m_{\mathcal{N}} < -1$, the sequence $(\mathcal{F}_s(\mathrm{X}_n))$ diverges to $-\infty$ almost surely and the sequence $(\|\mathrm{X}_n\|)$ diverges to $+\infty$ almost surely.

Sketch of the proof (see detailed proof in Appendix) As $1 + m_{\mathcal{N}} < 0$, the probability to sample a noise $\mathcal{N}_n$ such that $1 + \mathcal{N}_n < 0$ is striclty positive. Therefore there exists an integer n_1 such that for all $n \geq n_1$, $\mathcal{F}_s(\mathrm{X}_n) < 0$. Consequently $(\|\mathrm{X}_n\|)$ is lower bounded by A as illustrated in Fig. 3.1 (right) where the straight horizontal line represents the slope $y = \mathcal{F}_s(X_{n_1})$. Besides, the probability to have $\mathcal{F}_s(\mathrm{X}_n)$ as small as we want is lower bounded by a strictly positive constant which gives with BCL the divergence of the sequence $(\mathcal{F}_s(\mathrm{X}_n))$ to $-\infty$, i.e. the sequence $(\|\mathrm{X}_n\|)$ diverges to $+\infty$. □
Remark that for the example sketched in the introduction where $\mathcal{N}$ takes the 3 different values γ, 0 and $-\gamma$ and under the condition $\gamma > 1$ the proof of divergence will follow the same lines.

3.1.4 Discussion and conclusion

We conclude from Theorem 3.1 that what matters for convergence or divergence of the (1+1)-ES in the case of noisy objective function with positive noiseless part is the position of the lower bound $m_{\mathcal{N}}$ of the noise distribution $\mathcal{N}$ with respect to -1 or in other words the existence or not of possible negative fitness values. This result applies in particular when $\mathcal{N}$ equals a truncated normal distribution, i.e. $\mathcal{N} = \sigma_\epsilon \mathcal{N}(0,1)1_{[-a,a]}$[9] for any a and σ_ϵ positive. Whenever $\sigma_\epsilon a > 1$, Proposition 3.4 applies and the $(1+1)$-ES diverges.

Those results might appear in contradiction with those of Arnold and Beyer [8] proving that the expected fitness gain is positive−and therefore convergence in mean holds for the scale-invariant ES−for a noise distributed according to a normal distribution. In their model, Arnold and Beyer scale the standard deviation of the noise σ_ϵ with $1/d$, i.e. when $d \to \infty$, σ_ϵ converges to 0. The largest value for the normalized σ^*_ϵ in [8, Fig 5, 6, 8], for $d = 80$ corresponds to a standard deviation of 0.05 for which the probability to have $(1 + 0.05\mathcal{N}) < 0$ is upper bounded by 10^{-88} [(10)], i.e. relatively unlikely! Therefore though they consider some unbounded noise having a support in $\mathbb{R}$, the normalization of the standard deviation of the noise implies a so small probability to sample $1 + \mathcal{N}$ below -1 that the unbounded noise reduces to the case of convergence where $m_{\mathcal{N}} > -1$. The same conclusion holds for the numerical example given in Section 3.1.2, Fig. 3.3 (left) where

[9] The indicator function $1_{[-a,a]}(x)$ equals 1 if $x \in [-a, a]$ and 0 otherwise.

[10] For computing the lower bound we use the fact that $P(\mathcal{N}(0,1) < x) \leq \exp(-x^2/2)/|x|\sqrt{(2\pi)}$ for $x < 0$.

the standard deviation of 0.1 corresponds to a probability to have $(1+0.1\mathcal{N})<0$ lower bounded by 10^{-23}. Therefore though the theory predicts divergence as soon as $m_{\mathcal{N}}<-1$, what matters in practice is how likely the probability to sample $\mathcal{N}<-1$ is.

In conclusion, we have illustrated that convergence but also divergence can happen for the multiplicative noise model. Those results are due to the probability to sample $1+\mathcal{N}$ smaller than 0 and are therefore intrinsic to the noise model and not to the '+' strategy. The probability that $1+\mathcal{N}$ can be very small, in which case theory predicts divergence that will not be observed in simulations. We decided to present simple proofs relying on Borel-Cantelli Lemma. As a consequence, those proofs do not show the log-linear convergence and divergence observed in Section 3.1.2. Obtaining the log-linear behavior can be achieved using the theory of Markov chain on continuous state space. Last, we did not include results concerning a translated sphere $f(x)=\|x\|^2+\alpha$ with $\alpha\geq 0$ for which our proofs of convergence can be extended but where linear convergence does not hold anymore due to the fact that the variance of the noise distribution does not reduce to zero close to the optimum.

Acknowledgments

The authors would like to thank Nikolaus Hansen for many valuable discussions. This work receives partial supports from the ANR/RNTL project Optimisation Multidisciplinaire (OMD).

Appendix

Proof of Proposition 3.3 The sequence $(\mathcal{F}_s(\mathrm{X}_n))$ is decreasing and is lower bounded by 0 as $\mathcal{F}_s(\mathrm{X}_n)\geq\|\mathrm{X}_n\|^2(1+m_{\mathcal{N}})\geq 0$. Therefore it converges to a limit $l\geq 0$. Let us show that $l=0$. Let $\epsilon>0$, we have to show that $\exists\, n_0\geq 0$ such that $\mathcal{F}_s(\mathrm{X}_n)\leq\epsilon$ for $n\geq n_0$. Since the sequence $(\mathcal{F}_s(\mathrm{X}_n))$ is decreasing, we only have to show that $\exists\, n_0\geq 0$ such that $\mathcal{F}_s(X_{n_0})\leq\epsilon$. Let $\beta>1$ and such that $[1+m_{\mathcal{N}},\beta(1+m_{\mathcal{N}})[\subset \mathrm{supp}(1+\mathcal{N})$. In Lemma 3.5, we have defined the event $A_{n,\epsilon,\beta}$, shown that it is included in the event $\{\mathcal{F}_s(\mathrm{X}_{n+1})\leq\epsilon\}$ and proved that the events $(A_{n,\epsilon,\beta})_n$ are independent. Moreover, $P(A_{n,\epsilon,\beta})=P(\|\mathrm{e}_1+\sigma\mathrm{N}\|^2\leq\frac{\epsilon}{(1+\beta)\theta^2(1+m_{\mathcal{N}})})P(1+\mathcal{N}\leq\beta(1+m_{\mathcal{N}}))$ (where θ is defined in Lemma 3.5) is a strictly positive constant for all n. Then $\sum_{n=0}^{+\infty}P(A_n)=+\infty$. This gives by BCL that $P(\overline{\lim}\ An)=1$. Therefore $P(\overline{\lim}\ \{\mathcal{F}_s(\mathrm{X}_{n+1})\leq\epsilon\})=1$, i.e. $\exists n_0$ such that $\forall n\geq n_0$, $\mathcal{F}_s(\mathrm{X}_n)\leq\epsilon$. Therefore $\mathcal{F}_s(\mathrm{X}_n)$ converges to 0. The sequence $(\|\mathrm{X}_n\|)$ converges also to 0 as $\|\mathrm{X}_n\|^2\leq\frac{\mathcal{F}_s(\mathrm{X}_n)}{1+m_{\mathcal{N}}}$. □

Lemma 3.5. If $m_{\mathcal{N}}+1>0$, the following points hold:

1. The sequence $(\|\mathrm{X}_n\|)$ is upper bounded by $\theta:=\sqrt{\frac{\mathcal{F}_s(\mathrm{X}_0)}{1+m_{\mathcal{N}}}}>0$.
2. Let $\epsilon>0$ and $\beta>1$ such that $\beta(1+m_{\mathcal{N}})\in\mathrm{supp}(1+\mathcal{N})$. For $n\geq 0$, the event $A_{n,\epsilon,\beta}:=\left(\left\{\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\right\|^2\leq\frac{\epsilon}{(1+\beta)\theta^2(1+m_{\mathcal{N}})}\right\}\cap\{1+\mathcal{N}_n\leq\beta(1+m_{\mathcal{N}})\}\right)$ [11] verifies $A_{n,\epsilon,\beta}\subset\{\mathcal{F}_s(\mathrm{X}_{n+1})\leq\epsilon\}$. Moreover, the events $(A_{n,\epsilon,\beta})_n$ are independent.

[11] The multivariate Gaussian distribution is absolutely continuous with respect to the Lebesgue measure such that $P(\|\mathrm{X}_n\|=0)=0$ and then we can divide by $\|\mathrm{X}_n\|$ almost surely.

Proof :
1. For $n \geq 0$, $\mathcal{F}_s(\mathrm{X}_n) = \|\mathrm{X}_n\|^2 (1 + O_n) = \|\mathrm{X}_n\|^2 \left(1 + \mathcal{N}_{\phi(n)}\right)$ where $\phi(n)$ is the index of the last acceptance (obviously $\phi(n) \leq n$). Then, for $n \geq 0$
$\mathcal{F}_s(\mathrm{X}_n) \geq \|\mathrm{X}_n\|^2 (1 + m_{\mathcal{N}}) \geq 0$ and consequently $\|\mathrm{X}_n\|^2 \leq \frac{\mathcal{F}_s(\mathrm{X}_n)}{1+m_{\mathcal{N}}} \leq \frac{\mathcal{F}_s(\mathrm{X}_0)}{1+m_{\mathcal{N}}}$.
2. Let $\epsilon > 0$ and $\beta > 1$ such that $[1 + m_{\mathcal{N}}, \beta(1 + m_{\mathcal{N}})[\subset \mathrm{supp}(1 + \mathcal{N})$ (with $\beta m_{\mathcal{N}} < M_{\mathcal{N}}$ if $M_{\mathcal{N}} < +\infty$). For $n \geq 0$, the event
$\left\{ \left(\left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}_n \right\|^2 < \frac{\epsilon}{(1+\beta)\theta^2(1+m_{\mathcal{N}})} \right) \cap (1 + \mathcal{N}_n < \beta(1 + m_{\mathcal{N}})) \right\}$ implies for the offspring $\tilde{\mathrm{X}}_n := \mathrm{X}_n + \sigma \|\mathrm{X}_n\| \mathrm{N}_n$ created at the iteration n that
$\mathcal{F}_s(\tilde{\mathrm{X}}_n) = \|\mathrm{X}_n\|^2 \left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}_n \right\|^2 (1 + \mathcal{N}_n) \leq \theta^2 \frac{\epsilon}{(1+\beta)(1+m_{\mathcal{N}})\theta^2} \beta(1 + m_{\mathcal{N}})$.
Then $\mathcal{F}_s(\tilde{\mathrm{X}}_n) \leq \frac{\beta}{\beta+1}\epsilon < \epsilon$. If this offspring is accepted then $\mathcal{F}_s(\mathrm{X}_{n+1}) < \epsilon$, otherwise the fitness is already less than ϵ and we have also $\mathcal{F}_s(\mathrm{X}_{n+1}) < \epsilon$. Finally, the independency of the events $(A_{n,\epsilon,\beta})_n$ result from Lemma 3.6 applied to the sequence (X_n). □

Lemma 3.6. Let (U_n) be a sequence of random vectors in $\mathbb{R}^d$ such that $P(\|U_n\| = 0) = 0$ and N_n independent random vectors distributed as $\mathrm{N}(0, I_d)$. Then the variables $Y_n := \left\| \frac{U_n}{\|U_n\|} + \sigma \mathrm{N}_n \right\|$ are independent.

Proof :
The independance of the random variables Y_n is due to the fact that the multivariate Gaussian variable $\mathrm{N}(0, I_d)$ is isotropic and is therefore invariant by rotation. The length of the vector $\frac{U_n}{\|U_n\|} + \sigma \mathrm{N}_n$ will therefore be independent of where we start on the unit hypersphere, i.e., independent of the vector $\frac{U_n}{\|U_n\|}$. □

Proof of Proposition 3.4 Let $n \geq n_1$ (n_1 defined in Lemma 3.7). We have to show that for any $m < \mathcal{F}_s(X_{n_1}) < 0$, $\exists\ n \geq n_1$ such that $\mathcal{F}_s(\mathrm{X}_n) \leq m$, or equivalently $|\mathcal{F}_s(\mathrm{X}_n)| \geq |m|$. Similarly to the proof of Proposition 3.3, by BCL we have ($B_{n,m,\beta}$ i.o.) (($B_{n,m,\beta}$ being defined in Lemma 3.7) therefore Lemma 3.7 gives that ($\mathcal{F}_s(\mathrm{X}_{n+1}) \leq m$ i.o.). Then $\mathcal{F}_s(\mathrm{X}_n) = \|\mathrm{X}_n\|^2 (1 + O_n)$ tends to $-\infty$. For all $n \geq n_1$, $0 \geq 1 + O_n \geq 1 + m_{\mathcal{N}}$, then $\frac{|\mathcal{F}_s(\mathrm{X}_n)|}{|1+m_{\mathcal{N}}|} \leq \|\mathrm{X}_n\|^2$ for $n \geq n_1$. Consequently $(\|\mathrm{X}_n\|)$ converges to $+\infty$ almost surely. □

Lemma 3.7. Assume that $m_{\mathcal{N}} + 1 < 0$. The following points hold:

1. There exists $n_1 \geq 0$ and $A := \sqrt{\frac{|\mathcal{F}_s(X_{n_1})|}{|1+m_{\mathcal{N}}|}} > 0$ such that $\mathcal{F}_s(\mathrm{X}_n) < 0$ and $\|\mathrm{X}_n\| \geq A$ for $n \geq n_1$ almost surely.

2. Let $m < \mathcal{F}_s(X_{n_1}) < 0$ and $\beta > 1$. For $n \geq n_1$, the event $B_{n,m,\beta}$ defined by $B_{n,m,\beta} := \left(\left\{ |1 - \sigma\|\mathrm{N}_n\||^2 \geq \frac{|m|}{|m_{\mathcal{N}}+1|} \frac{\beta+1}{A^2} \right\} \cap \left\{ 1 + \mathcal{N}_n \leq \frac{1+m_{\mathcal{N}}}{\beta} \right\} \right)$ verifies $B_{n,\epsilon,\beta} \subset (\mathcal{F}_s(\mathrm{X}_{n+1}) \leq m)$.

Proof :
1. We first prove that the event $\mathcal{A} := \{\ \exists\ n_1 \geq 0$ such that $\forall\ n \geq n_1$,
$\mathcal{F}_s(\mathrm{X}_n) < 0\}$ is equivalent to the event $\mathcal{B} := \{\ \exists\ p_0 \geq 0$ such that $\mathcal{N}_{p_0} < -1\ \}$.
Proving that $\mathcal{A} \subset \mathcal{B}$ is equivalent to show that $\mathcal{B}^c \subset \mathcal{A}^c$. Suppose that $\forall p \geq 0$, $\mathcal{N}_p \geq -1$. Then $\forall p \geq 0$, $O_p \geq -1$. Therefore $\forall p \geq 0$, $\mathcal{F}_s(X_p) = \|X_p\|^2 (1 + O_p) \geq 0$. Now we have to show that $\mathcal{B} \subset \mathcal{A}$: Suppose that $\exists\ p_0 \geq 0$ such that $\mathcal{N}_{p_0} < -1$. We denote

$p_1 \geq 0$ the integer defined by $p_1 = \min\{p \in \mathbb{N} \text{ such that } \mathcal{N}_p < -1\}$. Then $\mathcal{F}_s(X_{p_1}) < 0$ and $\mathcal{F}_s(X_p) \geq 0$ for all $0 \leq p \leq p_1 - 1$. Since $(\mathcal{F}_s(\mathrm{X}_n))$ is a decreasing sequence, $\mathcal{F}_s(\mathrm{X}_n) < 0 \ \forall\ n \geq p_1$. This implies that $P(\mathcal{A}) = P(\mathcal{B})$. Now, we have for all $n \geq 0$, $P(\mathcal{B}^c) = P(\cap_{p=0}^{+\infty}(\mathcal{N}_p \geq -1)) \leq \Pi_{p=0}^{n} P(\mathcal{N}_p \geq -1) = (P(\mathcal{N} \geq -1))^n$.
Let $a := P(\mathcal{N} \geq -1)$[12]. As $m_{\mathcal{N}} < -1$, then $a < 1$ which gives $P(\mathcal{B}^c) = 0$ and therefore $P(\mathcal{A}) = 1$. Then $\exists\ n_1 \geq 0$ such that $\mathcal{F}_s(\mathrm{X}_n) < 0$ for $n \geq n_1$ almost surely. The sequence $(\mathcal{F}_s(\mathrm{X}_n))_n$ is decreasing (because of the elitist selection). Then for $n \geq n_1$, $\mathcal{F}_s(\mathrm{X}_n) \leq \mathcal{F}_s(X_{n_1}) < 0$. This gives $|\mathcal{F}_s(\mathrm{X}_n)| \geq |\mathcal{F}_s(X_{n_1})| > 0$. It is easy to see (from Eq. 3.6) that for all $n \in \mathbb{N}$, $O_n = \mathcal{N}_{\psi(n)}$ where $\psi(n)$ is the last acceptance index before the iteration n. Combining this with the fact if $1 + m_{\mathcal{N}} \leq 1 + \mathcal{N}_{\psi(n)} < 0$ one gets $0 < |\mathcal{F}_s(X_{n_1})| \leq |\mathcal{F}_s(\mathrm{X}_n)| = \|\mathrm{X}_n\|^2 |1 + \mathcal{N}_{\psi(n)}| \leq \|\mathrm{X}_n\|^2 |1 + m_{\mathcal{N}}|$. Then $\|\mathrm{X}_n\|^2 \geq \frac{|\mathcal{F}_s(X_{n_1})|}{|1+m_{\mathcal{N}}|} > 0$.
2. By the first result of the Lemma, $\exists\ n_1 \geq 0$, $A > 0$ such that $\mathcal{F}_s(\mathrm{X}_n) < 0$ and $\|\mathrm{X}_n\| \geq A\ \forall n \geq n_1$. We consider $n \geq n_1$, then $\|\mathrm{X}_n\| > A$. We notice that $\forall\ y \in \mathbb{R}^d \backslash \{(0,0)\}$, $\left\| \frac{y}{\|y\|} + \sigma \mathrm{N} \right\| \geq |1 - \sigma \|\mathrm{N}\||$. Let $\beta > 1$. As the upper bound $M_{\mathcal{N}}$ verifies $1 + M_{\mathcal{N}} > 0$, $\frac{1+m_{\mathcal{N}}}{\beta} \in \mathrm{supp}(1+\mathcal{N}) \cap \mathbb{R}^-$. Suppose that we have $|1 - \sigma\|\mathrm{N}_n\||^2 \geq \frac{(\beta+1)|m|}{A^2|1+m_{\mathcal{N}}|}$ and $|1 + \mathcal{N}_n| \geq \frac{|1+m_{\mathcal{N}}|}{\beta}$, then the offspring $\tilde{\mathrm{X}}_n := \mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n$ is such that
$|\mathcal{F}_s(\tilde{\mathrm{X}}_n)| = \|\mathrm{X}_n\|^2 \left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}_n \right\|^2 |1 + \mathcal{N}_n| \geq \|\mathrm{X}_n\|^2 |1 - \sigma\|\mathrm{N}_n\||^2 |1 + \mathcal{N}_n|$. Then $|\mathcal{F}_s(\tilde{\mathrm{X}}_n)| \geq \frac{\beta+1}{\beta}|m| > |m|$ which gives $\mathcal{F}_s(\mathrm{X}_{n+1}) \leq \mathcal{F}_s(\tilde{\mathrm{X}}_n) \leq m$. Consequently, for $n \geq n_0$, the event $B_{n,m,\beta} := \left\{ |1 - \sigma\|\mathrm{N}_n\|^2| \geq \frac{(\beta+1)|m|}{A^2|1+m_{\mathcal{N}}|} \right\} \cap \left\{ |1 + \mathcal{N}_n| \geq \frac{|1+m_{\mathcal{N}}|}{\beta} \right\}$ is included in $\{\mathcal{F}_s(\mathrm{X}_{n+1}) \leq m\}$. □

[12] We apply the same reasoning with $a = 2/3$ for the example given in the introduction where $\mathcal{N}$ take values in $\{-\gamma, 0, \gamma\}$ (with $\gamma > 1$) .

3.2 Convergence and divergence rates of the $(1+1)$-ES under multiplicative noise

It is generally observed in the case of optimization with Evolution Strategies (ES) and theoretically proven, in the case of minimization of non-noisy sphere functions, using either the artificial scale-invariant adaptation rule [1] [27, 17, 77] or the real Self-Adaptation rule [27, 13] that ESs converge (or diverge) log-linearly. This means that, after an adaptation time, the logarithm of the distance to the optimum decreases (or increases) linearly with the number of iterations. Let d_n denote the distance, at the iteration n, of the current solution to the optimum. The log-linear behavior of the algorithm is here mathematically expressed as:

$$\exists\, c \in \mathbb{R}^* \text{ such that } \lim_{n\to\infty} \frac{1}{n} \ln\left(d_n\right) = c\,. \tag{3.7}$$

The limit c is called convergence rate. The term "convergence" has to be considered in the mathematical sense relative to the convergence of the sequence $\frac{1}{n}\ln\left(d_n\right)$. In fact, if $c > 0$, the algorithm diverges. If $c < 0$, the algorithm converges.
Specific results have been derived in the case of minimization using the simplest ES, the $(1+1)$-ES. When minimizing the non-noisy sphere function, the scale-invariant $(1+1)$-ES converges log-linearly with a strictly negative convergence rate [77]. When the objective function is the sphere function with multiplicative noise lower bounded, the $(1+1)$ scale-invariant ES converges or diverges (according to the infimum of the noise) as we have shown in Section 3.1. Moreover, a log-linear behavior has been observed in Figure 3.2 but only the convergence or divergence have been proven and not the log-linear behavior.
The aim of this section is to generalize the theoretical result of log-linear behavior of the $(1+1)$-ES minimizing sphere functions to noisy sphere functions First, in Section 3.2.1, we recall the mathematical definition of the algorithm, of the objective function model and previous results of convergence and divergence obtained (This section may seem redundant with defintions and results of Section 3.1, but it will be useful for the paper that we intend to submit and which will be constituted of the whole Section 3.2.). Then, in Section 3.2.2, we investigate the log-linear behavior of the algorithm and show that there exists $c \in \mathbb{R}$ such that $\lim_{n\to\infty} \frac{1}{n}\ln\left(d_n\right) = c$ where c is given in terms of the expectation, with respect to the probability measure relative to the stationary state of the algorithm, of the difference of the logarithms of two consecutvie distances to the optimum. The proofs of all results of this section are in the appendix section

[1] The scale-invariant rule is not realistic as it assumes the knowledge of the optimum location.

3.2.1 Mathematical formulation of the problem and (spatial) convergence and divergence of the $(1+1)$-ES

In this section, we present the model of the noisy sphere function and the mathematical model of the $(1+1)$-ES used for finding the optimum of the noiseless part of this function. Then we recall the results derived in [76] relative to the scale-invariant $(1+1)$-ES minimizing this noisy function: The scale-invariant $(1+1)$-ES converges or diverges relatively to the infimum of the noise distribution support.

Noisy objective function model : Sphere function with multiplicative noise

The noisy sphere function mapping $\mathbb{R}^d$ into $\mathbb{R}$ is defined as:

$$\mathcal{F}_s(x) = \|x\|^2(1+\mathcal{N}) \tag{3.8}$$

where $\mathcal{N}$ is the noise random variable, sampled independently at each new evaluation of a point. The noisy part of $\mathcal{F}_s(x)$ is $\|x\|^2\mathcal{N}$. Therefore, the term $\mathcal{N}$ represents the normalized noisy part of the noisy sphere function which will be called normalized overvaluation of x. The term normalized overvaluation has been introduced in [8] where it corresponds to the normalized difference between the ideal and the noisy objective function. We assume that $\mathcal{N}$ has a finite expectation and that $E(\mathcal{N}) > -1$. Therefore, our study includes the particular case of white noise where $E(\mathcal{N}) = 0$. We also assume that $\mathcal{N}$ admits a density function $p_{\mathcal{N}}$ with support $[m_{\mathcal{N}}, M_{\mathcal{N}}[$ where $-\infty < m_{\mathcal{N}} < M_{\mathcal{N}} \leq +\infty$, $M_{\mathcal{N}} > -1$ and $m_{\mathcal{N}} \neq -1$.

Mathematical model for the scale-invariant $(1+1)$-ES minimizing $\mathcal{F}_s$ (Eq. 3.8)

The (1+1)-ES is a simple ES evolving a unique solution. At every iteration n, this solution denoted X_n and called parent is perturbed by the addition of a centered multivariate normal distribution to create a new candidate solution called offspring. The offspring writes as $\mathrm{X}_n + \sigma_n \mathrm{N}_n$, where σ_n is a strictly positive real number called step-size and $(\mathrm{N}_n)_n \in \mathbb{R}^d$ are independent realizations of a multivariate isotropic normal distribution in $\mathbb{R}^d$ denoted by $\mathrm{N}(0, I_d)$ [2]. The density of $\mathrm{N}(0, I_d)$ is denoted p_{N}. In the specific case of random variables $(\mathrm{N}_n)_n \in \mathbb{R}^d$ following the spherical multivariate normal distribution $\mathrm{N}(0, I_d)$, the algorithm is called isotropic ES. The efficiency of an isotropic ES is closely related to the adaptation rule of the sequence $(\sigma_n)_n$. The best adaptation scheme is the so-called scale-invariant adaptation rule for which the step-size is set proportionally to the distance to the optimum, i.e., $\sigma_n = \sigma\|\mathrm{X}_n\|$ where σ is a strictly positive constant. The optimality of this artificial rule in spherical environments has been derived in [17, 77]. The algorithm using this adaptation rule is referred to as the scale-invariant $(1+1)$-ES for which the offspring writes as $\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n$.

Let $\mathrm{X}_0 \in \mathbb{R}^d$ be the first parent randomly chosen such that $\|\mathrm{X}_0\| > 0$ almost surely and with a normalized overvaluation O_0 sampled from the distribution of $\mathcal{N}$. At an iteration n, and for the objective function investigated here (Eq. 3.8), the fitness of a parent X_n with a

[2] $\mathrm{N}(0, I_d)$ is the multivariate normal distribution with mean $(0, \ldots, 0) \in \mathbb{R}^d$ and covariance matrix identity I_d.

normalized overvaluation O_n equals $\|\mathrm{X}_n\|^2(1+O_n)$ and the fitness of an offspring equals $\|\mathrm{X}_n+\sigma\|\mathrm{X}_n\|\mathrm{N}_n\|^2(1+\mathcal{N}_n)$ where $(\mathcal{N}_n)_n$ is a sequence of random variables independent with $\mathcal{N}$ as a common law. The new parent X_{n+1} is the offspring $\mathrm{X}_n+\sigma\|\mathrm{X}_n\|\mathrm{N}_n$ iff its fitness value is smaller than the one of its parent X_n, otherwise X_{n+1} equals X_n. Therefore, this acceptance condition implies, for $n \geq 0$, that:

$$\begin{aligned}\mathrm{X}_{n+1} &= \mathrm{X}_n+\sigma\|\mathrm{X}_n\|\mathrm{N}_n \text{ if } \left[\left\|\mathrm{X}_n+\sigma\|\mathrm{X}_n\|\mathrm{N}_n\right\|^2\right](1+\mathcal{N}_n) < \|\mathrm{X}_n\|^2(1+O_n)\,,\\ &= \mathrm{X}_n \text{ otherwise}\,,\end{aligned} \tag{3.9}$$

and the normalized overvaluation O_{n+1} of the new parent X_{n+1} is then:

$$\begin{aligned}O_{n+1} &= \mathcal{N}_n \text{ if } \left[\left\|\mathrm{X}_n+\sigma\|\mathrm{X}_n\|\mathrm{N}_n\right\|^2\right](1+\mathcal{N}_n) < \|\mathrm{X}_n\|^2(1+O_n)\,,\\ &= O_n \text{ otherwise}\,.\end{aligned} \tag{3.10}$$

Convergence and divergence of the $(1+1)$-ES

The behavior of the algorithm defined by Eq. 3.9 and Eq. 3.10 designed for the minimization of the objective function (Eq. 3.8) was established in [76]. The result is recalled in the following theorem.

Theorem 3.8 ([76])**.** The $(1+1)$-ES defined in Eq. 3.9 minimizing the noisy sphere (Eq. 3.8) converges to zero if $m_{\mathcal{N}} > -1$ and diverges to infinity when $m_{\mathcal{N}} < -1$.

This theorem states that the behavior of the algorithm depends on the infimum $m_{\mathcal{N}}$ of the noise $\mathcal{N}$. If $m_{\mathcal{N}} < -1$, there is a strictly positive probability to sample negative fitness values and the algorithm diverges since the best fitness, which becomes negative after some iterations, is decreasing. If $m_{\mathcal{N}} > -1$, the algorithm converges. In the following section, we theoretically investigate the log-linear behavior of the algorithm defined by Eq. 3.9 and Eq. 3.10.

3.2.2 Convergence and divergence rates of the $(1+1)$-ES

Theoretical results of convergence of stochastic search algorithms can be obtained using mathematical tools such as Law of Large Numbers (LLN) for independent or orthogonal random variables or LLN for Markov chains. In the specific case of the noisy sphere function, Eq. 3.9 and Eq. 3.10 show that the variables are correlated and suggest the use of Markov chains to investigate the stability of these dynamics.

Motivations

The log-linear behavior means that, after an adaptation time, the sequence $\left(\ln\left(\|\mathrm{X}_n\|\right)\right)_n$ –where $\left(\|\mathrm{X}_n\|\right)_n$ is defined in Eq. 3.9– increases or decreases linearly with the number of iterations. This means that one has to investigate the sequence $\left(\ln\left(\|\mathrm{X}_n\|\right)\right)_n$. The following proposition is a basic step for proving the log-linear behavior expressing $\frac{1}{n}\ln\left(\frac{\|X_n\|}{\|X_0\|}\right)$

as the sum of n random variables divided by n. The same idea has been previously used in [27, 13, 17, 77].

Proposition 3.9. Let $(\mathrm{X}_n)_n$ be the sequence of random vectors valued in $\mathbb{R}^d$ satisfying the recurrence relation (3.9). Then for all indices n, we have

$$\frac{1}{n}\ln\left(\frac{\|X_n\|}{\|X_0\|}\right)=\frac{1}{n}\sum_{k=0}^{n-1}\ln\left(\left\|\frac{\mathrm{X}_k}{\|\mathrm{X}_k\|}+\sigma\mathrm{N}_k\mathbb{1}_{\left\{\left\|\frac{\mathrm{X}_k}{\|\mathrm{X}_k\|}+\sigma\mathrm{N}_k\right\|^2(1+\mathcal{N}_k)<1+O_k\right\}}\right\|\right)a.s. \tag{3.11}$$

Proposition 3.9 states that the limit of $\frac{1}{n}\ln\left(\frac{\|X_n\|}{\|X_0\|}\right)$ is given by the limit of the right hand side of Eq. 3.11. The right hand side of Eq. 3.11 can be simplified using the invariance by rotation of the multivariate normal distribution. For this purpose, we will introduce the sequences $(Z_n)_n$ and $(F(Z_n))_n$:

Definition 3.10. Consider a sequence of independent identically distributed (i.i.d.) random vectors $(\mathrm{N}'_n)_n$ in $\mathbb{R}^d$ with common law $\mathrm{N}(0, I_d)$ and a sequence of random variables $(\mathcal{N}'_n)_n$ also i.i.d. with $\mathcal{N}$ as common law. Let $\mathrm{e}_1 \in \mathbb{R}^d$ be equal to $(1, 0, \ldots, 0)$. We define

1. the Markov chain $(Z_n)_n$ as follows: $Z_0 = \mathcal{N}_*$ where $\mathcal{N}_*$ is a random variable distributed as $\mathcal{N}$, and, for all $n \geq 0$,

$$Z_{n+1} = \delta_n(Z_n)\mathcal{N}'_n + (1-\delta_n(Z_n))Z_n \tag{3.12}$$

   where $\delta_n(Z_n)$ equals 1 if $\|\mathrm{e}_1 + \sigma\mathrm{N}'_n\|^2(1+\mathcal{N}'_n) - 1 \leq Z_n$ and 0 otherwise.

2. the sequence $(F(Z_n))_{n\geq 0}$ as follows: for $n \geq 0$,

$$F(Z_n) := \ln\left(\|\mathrm{e}_1 + \sigma\mathrm{N}'_n\mathbb{1}_{\{\|\mathrm{e}_1+\sigma\mathrm{N}'_n\|^2(1+\mathcal{N}'_n)<1+Z_n\}}\|\right). \tag{3.13}$$

Using these definitions, we can state the key point of our study in the following Proposition.

Proposition 3.11 (Link between the stability of $(Z_n)_n$ and log-linear convergence)**.** Let $(Z_n)_n$ and $(F(Z_n))_n$ be the Markov chains introduced in Definition 3.10. Then the following equality

$$\frac{1}{n}\ln\left(\frac{\|X_n\|}{\|X_0\|}\right) = \frac{1}{n}\sum_{k=0}^{n-1}F(Z_k) \tag{3.14}$$

holds in distribution.Therefore, if $\frac{1}{n}\sum_{k=0}^{n-1}F(Z_k)$ converges almost surely to a finite value that we will denote γ, $\frac{1}{n}\ln\left(\frac{\|X_n\|}{\|X_0\|}\right)$ will converge (in probability) to the same value γ.

The condition $\frac{1}{n}\sum_{k=0}^{n-1}F(Z_k) \to \gamma$ given in Proposition 3.11 holds if the LLN holds for the Markov chain $(Z_n)_n$. If in addition $\gamma \neq 0$, then the log-linear behavior holds, at least in probability, for the sequence $(\|\mathrm{X}_n\|)_n$ given in Eq. 3.9. In the following section, we investigate the establishment of a LLN for the Markov chain $(Z_n)_n$.

Stability

In Proposition 3.11, we have seen that log-linear convergence can be implied from the stability of the chain $(Z_n)_n$ introduced in Definition 3.10. The goal is to prove that the chain $(Z_n)_n$ is sufficiently stable so that a LLN can be stated. Before investigating the stability of $(Z_n)_n$ we recall some definitions and results about φ-irreducible Markov Chains that will be used in the sequel. We refer to the Meyn and Tweedie book for a complete presentation of this theory [97]. In the following $\mathfrak{B}(\mathbb{R})$ will denote the Borel σ-algebra on $\mathbb{R}$ and for a subset $S \subset \mathbb{R}$, $\mathfrak{B}(S)$ will denote the Borel σ-algebra on S.

Basics about Markov chains and definitions For a Markov chain $(Z_n)_n \subset \mathbb{R}$, the *transition kernel* $P(.,.)$ is defined for all $z \in \mathbb{R}$, for all $A \in \mathfrak{B}(\mathbb{R})$ as

$$P(z, A) = P(Z_1 \in A | Z_0 = z).$$

A chain $(Z_n)_n$ is *irreducible with respect to a measure* φ if:

$$\forall (z, A) \in \mathbb{R} \times \mathfrak{B}(\mathbb{R}) \text{ such that } \varphi(A) > 0, \exists\, n_0 \geq 0 \text{ such that } P^{n_0}(z, A) > 0\,, \qquad (3.15)$$

where $P^{n_0}(z, A)$ equals $P(Z_{n_0} \in A | Z_0 = z)$. Another equivalent definition for the φ-irreducibility of the Markov chain $(Z_n)_n$ is: $\forall z \in \mathbb{R}, \forall A \in \mathfrak{B}(\mathbb{R})$ such that $\varphi(A) > 0$, $P(\tau_A < \infty | Z_0 = z) > 0$ where, τ_A is the hitting time of Z_n on A, i.e.,

$$\tau_A = \min\{n \geq 1 \text{ such that } Z_n \in A\}.$$

If the last term of Eq. 3.15 is equal to one, the chain is *recurrent*. A φ-irreducible chain $(Z_n)_n$ is *Harris recurrent* if:

$$\forall A \in \mathfrak{B}(\mathbb{R}) \text{ such that } \varphi(A) > 0; P_z(\eta_A = \infty) = 1, z \in \mathbb{R}\,,$$

where η_A is the occupation time of A, i.e., $\eta_A = \sum_{n=1}^{\infty} \mathbb{1}_{\{Z_n \in A\}}$.
A chain $(Z_n)_n$ which is Harris-recurrent admits an *invariant measure*, i.e., a measure π on $\mathfrak{B}(\mathbb{R})$ satisfying:

$$\pi(A) = \int_{\mathbb{R}} \pi(\mathrm{d}z) P(z, A), A \in \mathfrak{B}(\mathbb{R})\,.$$

If in addition this measure is a probability measure, the chain is called *positive*. Positive, Harris-recurrent chains satisfy the Strong Law of Large Numbers (LLN) as stated in [97, Theorem 17.0.1] and recalled here.

Theorem 3.12 (LLN for Harris positive chains)**.** Suppose that $(Z_n)_n$ is a positive Harris chain with invariant probability measure π, then the LLN holds for any function G satisfying $\pi(\mathrm{G}) = \int \mathrm{G} \mathrm{d}\pi < \infty$, i.e.,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \mathrm{G}(Z_k) = \pi(\mathrm{G})\,. \qquad (3.16)$$

To show the different stability notions such as recurrence, Harris-recurrence or positivity of $(Z_n)_n$ it is possible to make use of practical drift conditions. Stronger stability criteria are called *uniform ergodicity* and *geometric ergodicity* (see [97, Eq. 16.6, Eq. 15.7] for the definitions). These stability notions imply the positivity and Harris recurrence of the chain. Drift conditions can be used to prove the geometric ergodicity of φ-irreducible chain. Uniform ergodicity can be obtained without the need to verify the φ-irreducibility, using the following theorem which is derived from a specific case of [97, Theorem 16.2.1, Theorem 16.2.4].

Theorem 3.13 (Condition for uniform ergodicity)**.** Suppose that there exists a finite measure ν on $\mathfrak{B}(\mathbb{R})$ such that a Markov chain $(Z_n)_n$ satisfies $P(z, A) \geq \nu(A)$ for all $z \in \mathbb{R}$ and $A \in \mathfrak{B}(\mathbb{R})$. Then $(Z_n)_n$is uniformly ergodic.

Using the equivalent property of uniform ergodicity (assertion (vi) in [97, Theorem 16.0.2]) in the assertion (ii) of [97, Theorem 10.4.10] one can conclude that if a Markov chain $(Z_n)_n$ is uniformly ergodic then it is φ-irreducible, aperiodic (see definition in [97, p. 121]) positive Harris-recurrent. Combining this with Theorem 3.13, we have the following corollary.

Corollary 2. Suppose that there exists a finite measure ν on $\mathfrak{B}(\mathbb{R})$ such that a Markov chain $(Z_n)_n$ satisfies $P(z, A) \geq \nu(A)$ for all $z \in \mathbb{R}$ and $A \in \mathfrak{B}(\mathbb{R})$. Then (Z_n) is φ-irreducible, aperiodic, positive Harris-recurrent.

Stability of Z_n In the following, we will study the Markov chain $(Z_n)_n$ introduced in Definition 3.10. Its stability will follow from the use of Corollary 2 and consequently the (LLN) given in Theorem 3.12 holds for $(Z_n)_n$.

Lemma 3.14.

$$Z_n \in supp(p_{\mathcal{N}}) = [m_{\mathcal{N}}, M_{\mathcal{N}}[\,.$$

Proposition 3.15 (Transition Kernel)**.** The transition kernel $P(.,.)$ of Z_n is split into an absolutely continuous part P_1 and a singular part P_2:

$$\forall z \in [m_{\mathcal{N}}, M_{\mathcal{N}}[,\, \forall A \in \mathfrak{B}([m_{\mathcal{N}}, M_{\mathcal{N}}[),\, P(z, A) = P_1(z, A) + \delta_{\{z\}}(A)P_2(z) \tag{3.17}$$

where $P_1(z, A)$ equals $P\left(\{\mathcal{N} \in A\} \cap \{\|\mathrm{e}_1 + \sigma \mathrm{N}\|^2 (1 + \mathcal{N}) < 1 + z\}\right)$, $\delta_{\{z\}}$ is the Dirac measure concentrated in $\{z\}$ and $P_2(z) = P\left(\|\mathrm{e}_1 + \sigma \mathrm{N}\|^2 (1 + \mathcal{N}) \geq 1 + z\right)$. An other expression for P_1 is

$$P_1(z, A) = \int_{\mathbb{R}^d} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \mathbb{1}_A(u) \mathbb{1}_{\{\|\mathrm{e}_1 + \sigma t\|^2 (1+u) < 1+z\}}(u, t) p_{\mathrm{N}}(t) p_{\mathcal{N}}(u) \mathrm{d}u\mathrm{d}t\,. \tag{3.18}$$

Proposition 3.16 (Doeblin condition or minoration condition)**.** In the case $m_{\mathcal{N}} \neq -1$,

$$\forall z \in [m_{\mathcal{N}}, M_{\mathcal{N}}[, \forall A \in \mathfrak{B}([m_{\mathcal{N}}, M_{\mathcal{N}}[), P_1(z, A) \geq \nu(A)$$

where ν is the measure defined as

$$\nu(A) = \int_{\mathbb{R}^d} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \mathbb{1}_A(u) \mathbb{1}_{\{\|\mathrm{e}_1 + \sigma t\|^2 (1+u) < 1 + m_{\mathcal{N}}\}}(u, t) p_{\mathrm{N}}(t) p_{\mathcal{N}}(u) \mathrm{d}u\mathrm{d}t$$

The following corollary holds as a direct consequence of the application of Corollary 2 using the result of Proposition 3.16.

Corollary 3. If $m_{\mathcal{N}} \neq -1$, the chain $(Z_n)_n$ is positive Harris recurrent.

The following Proposition will be useful when establishing the LLN for the Markov chain $(Z_n)_n$.

Proposition 3.17. Suppose that the Markov chain $(Z_n)_n$ admits an invariant probability measure denoted μ. Let γ be the quantity defined by

$$\gamma := \int E\left[\ln(\|\mathrm{e}_1 + \delta(z)\sigma \mathrm{N}(0, I_d)\|)\right] d\mu(z)\,, \tag{3.19}$$

where $\delta(z)$ equals 1 if $\|\mathrm{e}_1 + \sigma \mathrm{N}(0, I_d)\|^2 (1 + \mathcal{N}) - 1 \leq z$ and 0 otherwise. Then γ is finite for all $\sigma > 0$. Moreover, the application $\sigma \mapsto \gamma(\sigma)$ is continuous on $]0, +\infty[$.

We are now ready to state the main result of this section

Theorem 3.18. The $(1+1)$-ES minimizing the noisy sphere (Eq. 3.8) defined in Eq. 3.9 (and Eq. 3.10) converges almost surely to zero if $m_{\mathcal{N}} > -1$ and diverges almost surely to infinity when $m_{\mathcal{N}} < -1$. The convergence (or divergence) rate verfies the following equation

$$\frac{1}{n} \ln \|\mathrm{X}_n\| \to \gamma := \int E\left[\ln(\|\mathrm{e}_1 + \delta(z)\sigma \mathrm{N}(0, I_d)\|)\right] d\mu(z) \tag{3.20}$$

which holds in probability and where $\delta(z)$ equals 1 if $\|\mathrm{e}_1 + \sigma \mathrm{N}(0, I_d)\|^2 (1 + \mathcal{N}) - 1 \leq z$ and 0 otherwise and μ is the invariant probability measure of the Markov chain $(Z_n)_n$. Moreover, if $1 + m_{\mathcal{N}} > 0$ then the convergence rate $\gamma \leq 0$ and if $1 + m_{\mathcal{N}} < 0$ then $\gamma \geq 0$.

Remark 3.2.1. Theorem 3.18 does not state that the log-linear behavior holds for the sequence $(\|\mathrm{X}_n\|)_n$ where $(\mathrm{X}_n)_n$ is defined in Eq. 3.9. It gives only the expression of the convergence (or divergence) rate of the sequence $\ln((\|\mathrm{X}_n\|))_n$. To show rigorously the log-linear behavior, one has to show that the convergence rate given in Eq. 3.20 is not equal to 0 when $m_{\mathcal{N}} \neq -1$. However, a benefit of our study is that the convergence rate derived in Eq. 3.20 is easy to compute numerically using Monte Carlo simulations. Note finally that Figure 3.2 suggests that the convergence (or divergence) rate is not equal to zero for the value of σ represented.

3.2.3 Conclusion

The theoretical study using LLN for Markov chains shows that the scale-invariant $(1+1)$-ES minimizing the noisy sphere function with lower bounded noise satisfies $\frac{1}{n} \ln \|\mathrm{X}_n\| \xrightarrow{\mathcal{P}} \gamma$ where γ is a finite convergence (or divergence) rate which corresponds to the expectation $\int E\left[\ln(\|\mathrm{e}_1 + \delta(z)\sigma \mathrm{N}(0, I_d)\|)\right] d\mu(z)$ where $\delta(z)$ equals 1 if $\|\mathrm{e}_1 + \sigma \mathrm{N}(0, I_d)\|^2 (1 + \mathcal{N}) - 1 \leq z$ and 0 otherwise and μ is the invariant probability measure of the Markov chain $(Z_n)_n$. However, we have not been able to exclude the case of the convergence rate γ equal to

zero to state that the behavior of the algorithm investigated is log-linear according to the definition given in Eq. 3.7. Figure 3.2 suggests that the algorithm converges or diverges log-linearly if the infimum of the noise, $m_{\mathcal{N}}$, is such that $m_{\mathcal{N}} \neq -1$. Numerical simulations of the convergence rate derived in Theorem 3.18 can be used to exclude numerically the case of null convergence rate which seems to be equivalent to the case $m_{\mathcal{N}} = -1$. Finally, another point which has to be investigated in a future work is to show that the convergence given in Eq. 3.7 i.e., $\frac{1}{n} \ln \|\mathrm{X}_n\| \to \gamma$ holds also almost surely.

Appendix

The following Lemma will be useful for proofs.

Lemma 3.19. The sequence $(\mathrm{X}_n)_n$ introduced in Eq. 3.9 satisfies: for every $n \geq 0$, $\|\mathrm{X}_n\| \neq 0$ almost surely.

Proof :
The result is demonstrated inductively. The first parent is chosen randomly with $P(\|\mathrm{X}_0\| = 0) = 0$. Suppose that $P(\|\mathrm{X}_n\| = 0) = 0$. As the offspring $\tilde{\mathrm{X}}_n$ is obtained by adding to X_n a random vector admitting an absolutely continuous distribution with respect to the Lebesgue measure then $P(\|\tilde{\mathrm{X}}_n\| = 0) = 0$. Consequently, if the offspring is accepted then $P(\|\mathrm{X}_{n+1}\| = 0) = P(\|\tilde{\mathrm{X}}_n\| = 0) = 0$, otherwise $P(\|\mathrm{X}_{n+1}\| = 0) = P(\|\mathrm{X}_n\| = 0) = 0$. □

Proof of Proposition 3.9

Taking the norm in Eq. 3.9, we have for $n \geq 0$

$$\|\mathrm{X}_{n+1}\| = \|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n \mathbb{1}_{\{\|\mathrm{X}_n+\sigma\|\mathrm{X}_n\|\mathrm{N}_n\|^2(1+\mathcal{N}_k)<(1+O_n)\|\mathrm{X}_n\|\}}\|$$

Lemma 3.19 states that $n \geq 0$, $\|\mathrm{X}_n\| \neq 0$ almost surely. Then the previous equation can be rewritten as

$$\|\mathrm{X}_{n+1}\| = \|\mathrm{X}_n\|\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n \mathbb{1}_{\left\{\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\right\|^2(1+\mathcal{N}_n)<(1+O_n)\right\}}\right\| \text{ a.s.}$$

Taking the logarithm of the previous equation, one has for $n \geq 0$

$$\ln(\|\mathrm{X}_{n+1}\|) = \ln(\|\mathrm{X}_n\|) + \ln\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n \mathbb{1}_{\left\{\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\right\|^2(1+\mathcal{N}_n)<(1+O_n)\right\}}\right\|\right) \text{ a.s.} \quad (3.21)$$

Summing the equations (3.21) from 0 to $n-1$ and dividing by n, one gets

$$\begin{aligned}\frac{1}{n}\ln\left(\frac{\|\mathrm{X}_n\|}{\|\mathrm{X}_0\|}\right) &= \frac{1}{n}\sum_{k=0}^{n-1}\ln\left(\frac{\|\mathrm{X}_{k+1}\|}{\|\mathrm{X}_k\|}\right) \\ &= \frac{1}{n}\sum_{k=0}^{n-1}\ln\left(\left\|\frac{\mathrm{X}_k}{\|\mathrm{X}_k\|} + \sigma\mathrm{N}_k \mathbb{1}_{\left\{\left\|\frac{\mathrm{X}_k}{\|\mathrm{X}_k\|}+\sigma\mathrm{N}_k\right\|^2(1+\mathcal{N}_k)<1+O_k\right\}}\right\|\right).\end{aligned}$$

□

Proof of Proposition 3.11

Step 1: We show that the random variables Z_n (introduced in Definition 3.10) and O_n (defined in Eq. 3.10) follow the same distribution. We are going to prove inductively this result. For $n = 0$, the random variables O_0 and $Z_0 = \mathcal{N}_*$ follow the same noise distribution

$\mathcal{N}$. For $n \geq 0$, suppose that O_n and Z_n follow the same distribution. We have to show that $E(e^{itO_{n+1}}) = E(e^{itZ_{n+1}})$. According to Eq. 3.10 and using Lemma 3.19, we have

$$E(e^{itO_{n+1}}|\mathrm{X}_n, O_n) = E\left\{e^{it\mathcal{N}_n}\mathbb{1}_{\left\{\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\right\|^2(1+\mathcal{N}_n)<1+O_n\right\}}|\mathrm{X}_n, O_n\right\} + E\left\{e^{itO_n}\mathbb{1}_{\left\{\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\right\|^2(1+\mathcal{N}_n)\geq 1+O_n\right\}}|\mathrm{X}_n, O_n\right\}.$$

Let $R_n : \mathbb{R}^d \mapsto \mathbb{R}^d$ be an orthogonal transformation (rotation) such that $R_n\left(\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}\right) = \mathrm{e}_1$. Then, $\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\| = \|R_n\left(\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right)\| = \|\mathrm{e}_1 + R_n(\mathrm{N}_n)\|$ which gives

$$E(e^{itO_{n+1}}|\mathrm{X}_n, O_n) = E\left\{e^{it\mathcal{N}_n}\mathbb{1}_{\left\{\left\|\mathrm{e}_1+\sigma R_n(\mathrm{N}_n)\right\|^2(1+\mathcal{N}_n)<1+O_n\right\}}|\mathrm{X}_n, O_n\right\} + E\left\{e^{itO_n}\mathbb{1}_{\left\{\left\|\mathrm{e}_1+\sigma R_n(\mathrm{N}_n)\right\|^2(1+\mathcal{N}_n)\geq 1+O_n\right\}}|\mathrm{X}_n, O_n\right\}.$$

This equation can be rewritten as:

$$E(e^{itO_{n+1}}|\mathrm{X}_n, O_n) = \int_{[m_{\mathcal{N}}, M_{\mathcal{N}}[}\int_{\mathbb{R}^d} e^{ity}\mathbb{1}_{\left\{\left\|\mathrm{e}_1+\sigma R_n(x)\right\|^2(1+y)<1+O_n\right\}}p_{\mathrm{N}}(x)dx\; p_{\mathcal{N}}(y)dy + \int_{[m_{\mathcal{N}}, M_{\mathcal{N}}[}\int_{\mathbb{R}^d} e^{itO_n}\mathbb{1}_{\left\{\left\|\mathrm{e}_1+\sigma R_n(x)\right\|^2(1+y)\geq 1+O_n\right\}}p_{\mathrm{N}}(x)dx\; p_{\mathcal{N}}(y)dy\,.$$

Let us apply the change of variables $z = R_n(x)$. As the isotropic multivariate normal distribution is invariant by orthogonal transformation, the new variable follows also the same multivariate normal distribution and one can write

$$E(e^{itO_{n+1}}|\mathrm{X}_n, O_n) = \int_{[m_{\mathcal{N}}, M_{\mathcal{N}}[}\int_{\mathbb{R}^d} e^{ity}\mathbb{1}_{\left\{\left\|\mathrm{e}_1+\sigma z\right\|^2(1+y)<1+O_n\right\}}p_{\mathrm{N}}(z)dz\; p_{\mathcal{N}}(y)dy + \int_{[m_{\mathcal{N}}, M_{\mathcal{N}}[}\int_{\mathbb{R}^d} e^{itO_n}\mathbb{1}_{\left\{\left\|\mathrm{e}_1+\sigma z\right\|^2(1+y)\geq 1+O_n\right\}}p_{\mathrm{N}}(z)dz\; p_{\mathcal{N}}(y)dy\,.$$

Therefore, one gets:

$$E(e^{itO_{n+1}}|\mathrm{X}_n, O_n) = E\left\{e^{it\mathcal{N}_n}\mathbb{1}_{\left\{\left\|\mathrm{e}_1+\sigma\mathrm{N}_n\right\|^2(1+\mathcal{N}_n)<1+O_n\right\}} + e^{itO_n}\mathbb{1}_{\left\{\left\|\mathrm{e}_1+\sigma\mathrm{N}_n\right\|^2(1+\mathcal{N}_n)\geq 1+O_n\right\}}|O_n\right\}.$$

The right hand side of the previous equation can be written as $g_n(O_n)$ with g_n continuous[3] and bounded ($|g_n(O_n)| \leq 1$). As O_n and Z_n follow the same distribution (recurrence

[3]The continuity follows from the Lebesgue dominated convergence Theorem for continuity.

hypothesis), then $E(g_n(O_n)) = E(g_n(Z_n))$ which means that by taking the expectation of the previous equation, one gets

$$E(e^{itO_{n+1}}) = E\left\{e^{it\mathcal{N}_n}\mathbb{1}_{\left\{\left\|\mathrm{e}_1+\sigma\mathrm{N}_n\right\|^2(1+\mathcal{N}_n)<1+Z_n\right\}}\right\} + E\left\{e^{itZ_n}\mathbb{1}_{\left\{\left\|\mathrm{e}_1+\sigma\mathrm{N}_n\right\|^2(1+\mathcal{N}_n)\geq 1+Z_n\right\}}\right\} = E(e^{itZ_{n+1}})\,.$$

Step 2: We have shown that the random variables Z_n and O_n follow the same distribution. In the same manner, we want to show that, for $n \geq 0$, the random variables

$\mathcal{U}_n := \ln\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\mathbb{1}_{\left\{\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\right\|^2(1+\mathcal{N}_n)<1+O_n\right\}}\right\|\right)$ and

$F(Z_n) := \ln\left(\|\mathrm{e}_1 + \sigma\mathrm{N}'_n\mathbb{1}_{\{\|\mathrm{e}_1+\sigma\mathrm{N}'_n\|^2(1+\mathcal{N}'_n)<1+Z_n\}}\|\right)$ are equal in distribution.

$$E(e^{it\mathcal{U}_n}|\mathrm{X}_n, O_n) = E\left\{e^{\left\{it\ln\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\right\|\right)\mathbb{1}_{\left\{\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\right\|^2(1+\mathcal{N}_n)<1+O_n\right\}}\right\}}|\mathrm{X}_n, O_n\right\}\,.$$

Again, the invariance of the multivariate normal distribution by any orthogonal transformation R and the fact that $\|R(x)\| = \|x\|$ for any $x \in \mathbb{R}^d$ gives

$$E(e^{it\mathcal{U}_n}|\mathrm{X}_n, O_n) = E\left\{e^{\left\{it\ln(\|\mathrm{e}_1+\sigma\mathrm{N}_n\|)\mathbb{1}_{\{\|\mathrm{e}_1+\sigma\mathrm{N}_n\|^2(1+\mathcal{N}_n)<1+O_n\}}\right\}}|O_n\right\}\,.$$

The conditional expectation $E(e^{it\mathcal{U}_n}|\mathrm{X}_n, O_n)$ reduces then to a function of O_n and can be written as $h_n(O_n)$ where h_n is real valued bounded function and for which the continuity follows from the Lebesgue dominated convergence Theorem. As O_n and Z_n follow the same distribution, one has $E(h_n(O_n)) = E(h_n(Z_n))$ which gives $E\left[e^{it\mathcal{U}_n}\right] = E\left[e^{itF(Z_n)}\right]$. Therefore, for $n \geq 0$, $\mathcal{U}_n$ and $F(Z_n)$ follow the same distribution.

Step 3: Now, we have to show that, for $n \geq 1$, $\sum_{k=0}^{n-1}\mathcal{U}_k$ and $\sum_{k=0}^{n-1}F(Z_k)$ are equal in distribution. We are going to prove the result inductively. For $n = 1$, $\sum_{k=0}^{0}\mathcal{U}_k = \mathcal{U}_0$ and $\sum_{k=0}^{0}F(Z_k) = F(Z_0)$ are equal in distribution according to step 2. Suppose that, for $n \geq 1$, $S_n := \sum_{k=0}^{n-1}\mathcal{U}_k$ and $T_n := \sum_{k=0}^{n-1}F(Z_k)$ are equal in distribution. Let us prove that S_{n+1} and T_{n+1} are equal in distribution. We have to show that $E\left(e^{itS_{n+1}}\right) = E\left(e^{itT_{n+1}}\right)$. We define the filtration $\mathcal{T}_n$ as

$\mathcal{T}_n := \sigma\left(\mathrm{X}_0, \ldots, \mathrm{X}_n, O_0, \ldots, O_n, \mathrm{N}_0, \ldots, \mathrm{N}_{n-1}, \mathcal{N}_0, \ldots, \mathcal{N}_n\right)$. We have

$$\begin{aligned}
E\left(e^{itS_{n+1}}|\mathcal{T}_n\right) &= e^{itS_n}E\left(e^{it\mathcal{U}_n}|\mathcal{T}_n\right)\\
&= e^{itS_n}E\left(e^{it\ln\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\mathbb{1}_{\left\{\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\right\|^2(1+\mathcal{N}_n)<1+O_n\right\}}\right\|\right)}|\mathcal{T}_n\right)\\
&= e^{itS_n}E\left(e^{it\ln\left(\|\mathrm{e}_1+\sigma\mathrm{N}_n\mathbb{1}_{\{\|\mathrm{e}_1+\sigma\mathrm{N}_n\|^2(1+\mathcal{N}_n)<1+O_n\}}\|\right)}|\mathcal{T}_n\right)\\
&= e^{itS_n}E\left(e^{itF(Z_n)}|\mathcal{T}_n\right)
\end{aligned}$$

This gives $E\left(e^{itS_{n+1}}|\mathcal{T}_n\right) = E\left(e^{itS_n+F(Z_n)}|\mathcal{T}_n\right)$. Taking the expectation of this equation, one gets $E\left(e^{itS_{n+1}}\right) = E\left(e^{itS_n+F(Z_n)}\right)$. This can be rewritten as

$$\begin{aligned} E\left(e^{itS_{n+1}}\right) &= E\left[E\left(e^{itS_n+F(Z_n)}|\mathrm{N}'_n, \mathcal{N}'_n, Z_n\right)\right] \\ &= E\left[E\left(e^{itS_n}|\mathrm{N}'_n, \mathcal{N}'_n, Z_n\right) e^{itF(Z_n)}\right] \\ &= E\left[E\left(e^{itS_n}\right) e^{itF(Z_n)}\right] \\ &= E\left[E\left(e^{itT_n}\right) e^{itF(Z_n)}\right] \\ &= E\left(e^{itT_{n+1}}\right) \end{aligned}$$

Consequently, for $n \geq 1$, $\frac{1}{n}\sum_{k=0}^{n-1} \mathcal{U}_k$ and $\frac{1}{n}\sum_{k=0}^{n-1} F(Z_k)$ are equal in distribution.
By Proposition 3.9, one has

$$\frac{1}{n}\ln\left(\frac{\|\mathrm{X}_n\|}{\|\mathrm{X}_0\|}\right) = \frac{1}{n}\sum_{k=0}^{n-1}\ln\left(\left\|\frac{\mathrm{X}_k}{\|\mathrm{X}_k\|} + \sigma\mathrm{N}_k \mathbb{1}_{\left\{\left\|\frac{\mathrm{X}_k}{\|\mathrm{X}_k\|}+\sigma\mathrm{N}_k\right\|^2(1+\mathcal{N}_k)<1+O_k\right\}}\right\|\right) a.s.$$

Then $\frac{1}{n}\ln\left(\frac{\|\mathrm{X}_n\|}{\|\mathrm{X}_0\|}\right)$ equals in distribution $\frac{1}{n}\sum_{k=0}^{n-1} F(Z_k)$. Consequently if the Markov chain $(Z_n)_n$ is stable such that it verifies the (LLN) for Markov chains, the result holds in the sense that $\frac{1}{n}\sum_{k=1}^{n} F(Z_k)$ converges to some γ a.s. It follows that $\frac{1}{n}\ln\|\mathrm{X}_n\|$ converges to γ in probability. □

Proof of Lemma 3.14

The result is proven inductively. For $n = 0$, by Definition 3.10, $Z_0 = \mathcal{N}_* \in [m_\mathcal{N}, M_\mathcal{N}[$. For $n \geq 0$, suppose that $Z_n \in supp(p_\mathcal{N}) = [m_\mathcal{N}, M_\mathcal{N}[$. By Eq. 3.12, Z_{n+1} equals $\mathcal{N}'_n \in supp(p_\mathcal{N}) = [m_\mathcal{N}, M_\mathcal{N}[$ or Z_{n+1} equals Z_n which is in $supp(p_\mathcal{N}) = [m_\mathcal{N}, M_\mathcal{N}[$ by the recurrence hypothesis. Then $Z_{n+1} \in supp(p_\mathcal{N}) = [m_\mathcal{N}, M_\mathcal{N}[$. □

Proof of Proposition 3.15

The transition kernel $P(z, A)$ is the probability that Z_1 belongs to A conditionally to $Z_0 = z$. By Eq. 3.12, Z_1 equals $\mathcal{N}'_0$ if $\|\mathrm{e}_1 + \sigma\mathrm{N}\|^2(1+\mathcal{N}) < 1+z$, otherwise Z_1 equals z. Let $P_1(z, A)$ represent the probability to have $Z_1 = \mathcal{N}'_0$ and $Z_1 \in A$ and $P_2(z)$ represent the probability to have $\|\mathrm{e}_1 + \sigma\mathrm{N}\|^2(1+\mathcal{N}) \geq 1+z$. The expression of $P(z, A)$ given in Eq. 3.18 follows. □

Proof of Proposition 3.16

Let us show that $\nu : \mathfrak{B}([m_\mathcal{N}, M_\mathcal{N}[) \mapsto \mathbb{R}^+ \cup \{+\infty\}$ defined as

$$\nu(A) = \int_{\mathbb{R}^d}\int_{m_\mathcal{N}}^{M_\mathcal{N}} \mathbb{1}_A(u)\mathbb{1}_{\{\|\mathrm{e}_1+\sigma t\|^2(1+u)<1+m_\mathcal{N}\}}(u,t)p_\mathrm{N}(t)p_\mathcal{N}(u)\mathrm{d}u\mathrm{d}t$$

is a finite measure. First, we have $\nu(\emptyset) = 0$. Second, if E_1 and E_2 are two disjoint sets then $\nu(E_1 \cup E_2) = \nu(E_1) + \nu(E_2)$ as the function $\mathbb{1}_{E_1\cup E_2}$ is identically equal to $\mathbb{1}_{E_1} + \mathbb{1}_{E_2}$

when $E_1 \cap E_2 = \emptyset$. Third,

$$\nu([m_{\mathcal{N}}, M_{\mathcal{N}}[) = \int_{\mathbb{R}^d} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \mathbb{1}_{\{\|\mathrm{e}_1+\sigma t\|^2(1+u)<1+m_{\mathcal{N}}\}}(u,t) p_{\mathrm{N}}(t) p_{\mathcal{N}}(u) \mathrm{d}u\mathrm{d}t \leq 1\,.$$

Now, if $m_{\mathcal{N}} = -1$ then the indicator function $\mathbb{1}_{\{\|\mathrm{e}_1+\sigma t\|^2(1+u)<1+m_{\mathcal{N}}\}}(u,t)$ equals zero for any $t \in \mathbb{R}^d$ and $u \in [-1, M_{\mathcal{N}}[$ almost surely. Therefore, ν is identically equal to zero. However, if $m_{\mathcal{N}} \neq -1$, then, for $A \in \mathfrak{B}([m_{\mathcal{N}}, M_{\mathcal{N}}[)$ with a strictly positive Lebesgue measure, the set

$$\mathcal{A} := \{(u,t) \in ([m_{\mathcal{N}}, M_{\mathcal{N}}[\cap A) \times \mathbb{R}^d \text{ such that } \|\mathrm{e}_1 + \sigma t\|^2 (1+u) < 1 + m_{\mathcal{N}}\}$$

has a stricly positive measure with respect to a Lebesgue measure defined on $\mathfrak{B}(\mathbb{R}^d \times [m_{\mathcal{N}}, M_{\mathcal{N}}[)$. This implies that ν is non identically equal to zero if and only if $m_{\mathcal{N}} \neq -1$. Moreover, for $t \in \mathbb{R}^d$, $(u,z) \in [m_{\mathcal{N}}, M_{\mathcal{N}}[^2$

$$\|\mathrm{e}_1 + \sigma t\|^2 (1+u) < 1 + m_{\mathcal{N}} \Rightarrow \|\mathrm{e}_1 + \sigma t\|^2 (1+u) < 1 + z$$

which gives that $\forall z \in [m_{\mathcal{N}}, M_{\mathcal{N}}[, \forall A \in \mathfrak{B}([m_{\mathcal{N}}, M_{\mathcal{N}}[), P_1(z,A) \geq \nu(A)$. □

Proof of Proposition 3.17

Let $g : \mathbb{R}^d \times \mathbb{R}_+^* \times \mathbb{R} \times \mathbb{R}$ be defined for (x, σ, y, z) in $\mathbb{R}^d \times \mathbb{R}_+^* \times \mathbb{R} \times \mathbb{R}$ by

$$g(x, \sigma, y, z) = \|\mathrm{e}_1 + \mathbb{1}_{\{\|\mathrm{e}_1+\sigma x\|^2(1+y)-1<z\}}(x,y,z)\sigma x\|\,.$$

The quantity γ defined in Eq. 3.19 results from the integration of the function $\ln(g)$ with respect to the variables x, y and z. We notice that $g((x_1, x_2, \ldots, x_d), \sigma, y, z) = g((x_1, \epsilon_2 x_2, \ldots, \epsilon_d x_d), \sigma, y, z)$ for all $(\epsilon_2, \ldots, \epsilon_d)$ in $\{-1, +1\}^{d-1}$ and $(x_1, x_2, \ldots, x_d)$ in $\mathbb{R}^d$. Therefore, we can restrict the integration with respect to the variable x to the domain $\mathcal{D} := \mathbb{R}^* \times]0, +\infty[^{d-1}$, more precisely the quantity γ can be rewritten as

$$\gamma = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{D}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \ln\left(g(x, \sigma, y, z)\right) e^{-\frac{\|x\|^2}{2}} p_{\mathcal{N}}(y) dx dy d\mu(z)\,.$$

where μ is the invariant probability measure of the Markov chain $(Z_n)_n$ introduced in Definition 3.10 which we supposed that it exists in the hypothesis of Proposition 3.17. We introduce γ^+ as:

$$\gamma^+ = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{D}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \ln^+\left[g(x, \sigma, y, z)\right] e^{-\frac{\|x\|^2}{2}} p_{\mathcal{N}}(y) dx dy d\mu(z)$$

and γ^- as:

$$\gamma^- = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{D}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \ln^-\left[g(x, \sigma, y, z)\right] e^{-\frac{\|x\|^2}{2}} p_{\mathcal{N}}(y) dx dy d\mu(z)$$

such that $\gamma = \gamma^+ - \gamma^-$. The quantities γ^+ and γ^- are well defined but could be infinite. Using spherical coordinates (with $d \geq 2$) we obtain after partial integration

$$\gamma^- = \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{1}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty} \int_0^{\frac{\pi}{2}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \ln^-\left[h(r,\theta,\sigma,y,z)\right] r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta) p_{\mathcal{N}}(y)\, dr\, d\theta dy d\mu(z)\,,$$

and

$$\gamma^+ = \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{1}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty} \int_0^{\pi} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \ln^+\left[h(r,\theta,\sigma,y,z)\right] r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta) p_{\mathcal{N}}(y)\, dr\, d\theta dy d\mu(z)\,,$$

where h is the positive function defined on $\mathbb{R}^+ \times [0,\pi] \times \mathbb{R}_+^* \times \mathbb{R} \times \mathbb{R}$ by

$$h(r,\theta,\sigma,y,z) = \|\mathbb{1}_{\{\|\sigma r - e^{i\theta}\|^2(1+y)-1<z\}}(r,\theta,y,z)\sigma r - e^{i\theta}\|\,.$$

For (r,θ,σ,y,z) in $\mathbb{R}^+ \times [0,\pi] \times \mathbb{R}_+^* \times \mathbb{R} \times \mathbb{R}$, we have

$$\ln^+(h(r,\theta,\sigma,y,z)) \leq \ln^+(1+\sigma r) \leq \sigma r \tag{3.22}$$

and

$$\ln^-(h(r,\theta,\sigma,y,z)) \leq \ln^-(\sin(\theta))\,. \tag{3.23}$$

This gives

$$\gamma^+ \leq \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{\sigma\pi}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty} r^d e^{-\frac{r^2}{2}}\, dr < +\infty\,,$$

and

$$\begin{aligned}
\gamma^- &\leq \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{1}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty} \int_0^{\frac{\pi}{2}} \ln^-\left(sin(\theta)\right) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta)\, dr\, d\theta \\
&\leq \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{2}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty} r^{d-1} e^{-\frac{r^2}{2}}\, dr \int_0^{\frac{\pi}{2}} \sin^{d-\frac{5}{2}}(\theta)\, d\theta < +\infty\,.
\end{aligned}$$

For the remaining case $d = 1$, we have

$$\gamma^+ \leq \frac{\sigma}{\sqrt{2\pi}} \int_{\mathbb{R}} |x| e^{-\frac{x^2}{2}} dx = \frac{2\sigma}{\sqrt{2\pi}} \int_{\mathbb{R}^+} x e^{-\frac{x^2}{2}} < +\infty\,,$$

For γ^-, after a change of variables $(v = \sigma x)$, we get

$$\begin{aligned}
\gamma^- &\leq \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \int_{-2}^{0} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \frac{\ln\left(|1 + \mathbb{1}_{\{|1+v|^2(1+y)-1<z\}}(v,y,z)v|\right)}{v} p_{\mathcal{N}}(y) dv dy d\mu(z) \\
&= \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \int_{-2}^{0} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \frac{\ln\left(|1+v|\right)}{v} \mathbb{1}_{\{|1+v|^2(1+y)-1<z\}}(v,y,z) p_{\mathcal{N}}(y) dv dy d\mu(z) \\
&\leq \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \int_{-2}^{0} \int_{m_{\mathcal{N}}}^{M_{\mathcal{N}}} \frac{\ln\left(|1+v|\right)}{v} p_{\mathcal{N}}(y) dv dy d\mu(z) \\
&= \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \int_{-2}^{0} \frac{\ln\left(|1+v|\right)}{v} dv < +\infty\,.
\end{aligned}$$

The continuity with respect to σ is shown , using the Lebesgue dominated convergence theorem (for continuity), on every range $]0, M[$ and then for the whole $]0, +\infty[$ thanks to the inequalities given in Eq. 3.22 and Eq. 3.23. This gives the result for $d > 1$.
For the case $d = 1$, the integrand in γ^+ is continuous with respect to σ for almost all (x, y, z) in $\mathbb{R} \times [m_{\mathcal{N}}, M_{\mathcal{N}}[\times[m_{\mathcal{N}}, M_{\mathcal{N}}[$ and is dominated by $\frac{2}{\sqrt{2\pi}} S x e^{-\frac{x^2}{2}}$ for $(x, \sigma, y, z) \in \mathbb{R}^+ \times]0, S] \times [0, +\infty[\times[m_{\mathcal{N}}, M_{\mathcal{N}}[\times[m_{\mathcal{N}}, M_{\mathcal{N}}[$ which gives the continuity of γ^+ with respect to σ by the Lebesgue dominated convergence Theorem. For γ^-, and after the change of variables $v = \sigma x$, the integrand will be dominated by $\frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \frac{\ln(|1+v|)}{v}$ for $(v, \sigma, y, z) \in]-2, 0] \times]0, +\infty[\times[m_{\mathcal{N}}, M_{\mathcal{N}}[\times[m_{\mathcal{N}}, M_{\mathcal{N}}[$ and the continuity of γ^- with respect to σ follows from the dominated convergence Theorem. □

Proof of Theorem 3.18

The almost sure convergence or divergence was already given in Theorem 3.8. Now, we give interest to the convergence (or divergence) rate.Corollary 3 states that, for $m_{\mathcal{N}} \neq -1$, the Markov chain $(Z_n)_n$ is positive and Harris recurrent. Therefore it satisfies the (LLN) given in Theorem 3.12. Let μ the invariant probability measure of the chain$(Z_n)_n$. Then, we can define the quantity $\gamma := \int E\left[\ln(\|\mathrm{e}_1 + \delta(z)\sigma \mathrm{N}(0, I_d)\|)\right] d\mu(z)$ where $\delta(z)$ equals 1 if $\|\mathrm{e}_1 + \sigma \mathrm{N}(0, I_d)\|^2 (1 + \mathcal{N}) - 1 \leq z$ and 0 otherwise. By Proposition 3.17, γ is finite. As $(Z_n)_n$ satisfies the conditions of the LLN and $\gamma < +\infty$, the right hand side of Eq. 3.14 converges almost surely to γ. Then the sequence $\frac{1}{n}\ln(\|\mathrm{X}_n\|)_n$ converges in distribution to γ. As γ is a constant, the convergence of the sequence $\frac{1}{n}\ln(\|\mathrm{X}_n\|)_n$ to γ holds also in probability.
The convergence in probability of the sequence $\frac{1}{n}\ln(\|\mathrm{X}_n\|)$ (when $m_{\mathcal{N}} \neq -1$) implies that there is a subsequence which writes as $\left(\frac{1}{\phi(n)}\ln(\|\mathrm{X}_{\phi(n)}\|)\right)_n$ and which converges almost surely to the same limit γ.

Moreover, by Theorem 3.8, the sequence $(\ln(\|\mathrm{X}_n\|))_n$ converges almost surely to $+\infty$ if $m_{\mathcal{N}} < -1$ and to $-\infty$ if $m_{\mathcal{N}} > -1$. Combining this with the fact that the sequence $\left(\frac{1}{\phi(n)}\ln(\|\mathrm{X}_{\phi(n)}\|)\right)_n$ converges almost surely to γ, we deduce that $\gamma \geq 0$ if $m_{\mathcal{N}} > -1$ and $\gamma \leq 0$ if $m_{\mathcal{N}} > -1$. □

3.3 Additional convergence/divergence results

In this section, we generalize convergence/divergence results that have been derived in [76] (Section 3.1) for the objective function defined by Eq. 3.4 to the following (noisy) objective function:

$$\mathcal{F}_\alpha(x) = (\|x\|^2 + \alpha)(1 + \mathcal{N}) \tag{3.24}$$

where α is a positive constant. The noise random variable $\mathcal{N}$ has a finite expectation such that $E(\mathcal{N}) > -1$ and has a density function $p_{\mathcal{N}}$ which lies in the range $[m_{\mathcal{N}}, M_{\mathcal{N}}[$ where $-\infty \leq m_{\mathcal{N}} < M_{\mathcal{N}} \leq +\infty$[4] , $M_{\mathcal{N}} > -1$ and $m_{\mathcal{N}} \neq -1$. Some of the proofs of the following results are based on the second Borel-Cantelli Lemma (see Lemma 3.2). It is worth noticing that the log-linear behavior observed in Figures 3.2 and 3.3 and theoretically shown in Section 3.2 when $\alpha = 0$ does not hold anymore for $\alpha > 0$ as the variance of the noise random variable does not reduce to zero close to the optimum. We recall here that:

- The random vector $\mathrm{N}(0, I_d)$ is the multivariate isotropic normal distribution on $\mathbb{R}^d$ with mean $(0, \ldots, 0) \in \mathbb{R}^d$ and covariance matrix the identity I_d.
- The random vectors N_n $(n \geq 0)$ are independent realizations of $\mathrm{N}(0, I_d)$.
- The random variables $\mathcal{N}_n$ $(n \geq 0)$ are independent realizations of $\mathcal{N}$.
- The vector e_1 is a unit vector in $\mathbb{R}^d$ which equals $(1, 0, \ldots, 0)$.

In the case of the minimization of the objective function (Eq. 3.24) using a scale-invariant $(1+1)$-ES, the solution at an iteration n, X_n, is updated as follows:

$$\begin{aligned} \mathrm{X}_{n+1} &= \mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n \text{ if } \left[\left\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n\right\|^2 + \alpha\right](1 + \mathcal{N}_n) < (\|\mathrm{X}_n\|^2 + \alpha)(1 + O_n) \,, \\ &= \mathrm{X}_n \text{ otherwise} \,, \end{aligned} \tag{3.25}$$

and the new normalized overvaluation O_{n+1} is then:

$$\begin{aligned} O_{n+1} &= \mathcal{N}_n \text{ if } \left[\left\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n\right\|^2 + \alpha\right](1 + \mathcal{N}_n) < (\|\mathrm{X}_n\|^2 + \alpha)(1 + O_n) \,, \\ &= O_n \text{ otherwise} \,. \end{aligned} \tag{3.26}$$

The results depend, similarly to the case of the noisy objective function given by Eq. 3.4, on the infimum of the noise $m_{\mathcal{N}}$. The results are summarized in the two following sections.

3.3.1 Convergence in the case $m_{\mathcal{N}} > -1$

The result is stated in the following proposition.

[4]Note that, comparing to Section 3.1, the hypothesis on the variable $\mathcal{N}$ are more general in this section: The infimum of the noise can be infinite, i.e., $m_{\mathcal{N}} = -\infty$.

Proposition 3.20 (Convergence for $m_{\mathcal{N}} > -1$)**.** Consider the sequences $(O_n)_n$ and $(\mathrm{X}_n)_n$ defined by the recurrence relations Eq. 3.25 and Eq. 3.26 for the minimization of the objective function defined in Eq. 3.24. If $m_{\mathcal{N}} + 1 > 0$ then the sequences $(\mathcal{F}_\alpha(\mathrm{X}_n))_n$ and $(\|\mathrm{X}_n\|)_n$ converge respectively to $\alpha(1 + m_{\mathcal{N}})$ and zero almost surely.

Proof :
The convergence in the case $\alpha = 0$ has been already stated in Proposition 3.3. Let us now demonstrate the result for $\alpha > 0$.
Step 1: Note in the beginning that the sequence $(\mathcal{F}_\alpha(\mathrm{X}_n))_n$ is decreasing due to the acceptance condition used in the $(1+1)$-ES. Let us show that the sequence $(\mathcal{F}_\alpha(\mathrm{X}_n))_n$ is positive, lower bounded and that the sequence $(\|\mathrm{X}_n\|)_n$ is upper bounded. The decrease of the sequence $(\mathcal{F}_\alpha(\mathrm{X}_n))_n$ and the fact the random variable $\mathcal{N}$ is lower bounded by $m_{\mathcal{N}} > -1$ imply, for $n \geq 0$, that:

$$\mathcal{F}_\alpha(\mathrm{X}_0) \geq \mathcal{F}_\alpha(\mathrm{X}_n) = \left(\|\mathrm{X}_n\|^2 + \alpha\right)(1 + O_n) \geq \left(\|\mathrm{X}_n\|^2 + \alpha\right)(1 + m_{\mathcal{N}}) \geq \alpha(1 + m_{\mathcal{N}}) \geq 0\,. \tag{3.27}$$

The decreasing sequence $(\mathcal{F}_\alpha(\mathrm{X}_n))_n$ is then positive and lower bounded. Therefore it converges almost surely. Moreover, by the previous equation, one gets

$$\|\mathrm{X}_n\|^2 \leq M \tag{3.28}$$

where M is defined as $M := \frac{\mathcal{F}_\alpha(\mathrm{X}_0) - \alpha(1+m_{\mathcal{N}})}{1+m_{\mathcal{N}}}$. This means that the sequence $(\|\mathrm{X}_n\|)_n$ is upper bounded by M.
Step 2: Let us show that the sequence $(\mathcal{F}_\alpha(\mathrm{X}_n))_n$ converges almost surely to $\alpha(1+m_{\mathcal{N}})$. Let $\epsilon > 0$, we are going to show that $\exists\ n_0 \geq 0$ such that $\mathcal{F}_\alpha(\mathrm{X}_n) < \alpha\,(1 + m_{\mathcal{N}}) + \epsilon$, $\forall n \geq n_0$. Let $n \geq 0$. For $\epsilon > 0$, $\exists\ K(\epsilon) > 1$ such that the probability to have $\mathcal{N}_n \in \ \mathrm{supp}(p_{\mathcal{N}})$ and $\alpha(1 + \mathcal{N}_n) < \alpha(1 + m_{\mathcal{N}}) + \frac{\epsilon}{K(\epsilon)}$ is strictly positive. Let $a := \frac{K(\epsilon)-1}{K(\epsilon)} \frac{1+m_{\mathcal{N}}}{1+m_{\mathcal{N}}+\frac{\epsilon}{K(\epsilon)\alpha}} \frac{\epsilon}{\mathcal{F}_\alpha(\mathrm{X}_0)-\alpha(1+m_{\mathcal{N}})} > 0$ and $b := 1+m_{\mathcal{N}}+\frac{\epsilon}{K(\epsilon)\alpha} > 0$. Suppose that the events $\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}_n\right\|^2 \leq a\right)$ and $(1 + \mathcal{N}_n < b)$ hold. Therefore, using in addition Eq. 3.28, the fitness of the offspring $\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n$ at an iteration n verifies

$$\begin{aligned}
\mathcal{F}_\alpha(\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n) &= \|\mathrm{X}_n\|^2 \left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}_n\right\|^2 (1 + \mathcal{N}_n) + \alpha\,(1 + \mathcal{N}_n) \\
&\leq Mab + \alpha b \\
&= \epsilon \frac{K(\epsilon) - 1}{K(\epsilon)} + \alpha\,(1 + m_{\mathcal{N}}) + \frac{\epsilon}{K(\epsilon)} = \alpha\,(1 + m_{\mathcal{N}}) + \epsilon\,.
\end{aligned}$$

Then, for $n \geq 0$, the event $A_n = \left(\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}_n\right\|^2 \leq a\right) \cap (1 + \mathcal{N}_n \leq b)\right)$ implies the event $\mathcal{F}_\alpha(\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n) \leq \alpha\,(1 + m_{\mathcal{N}}) + \epsilon$ and therefore $\mathcal{F}_\alpha(\mathrm{X}_{n+1}) \leq \mathcal{F}_\alpha(\mathrm{X}_n +$

$\sigma\|\mathrm{X}_n\|\mathrm{N}_n) \leq \alpha\,(1+m_{\mathcal{N}}) + \epsilon$. Moreover, the event A_n has a probability which verifies:

$$\begin{aligned} P(A_n) &= P\left(\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right\|^2 \leq a\right) \cap (1+\mathcal{N}_n \leq b)\right) \\ &= P\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right\|^2 \leq a\right) P\,(1+\mathcal{N}_n \leq b) \\ &= P\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right\|^2 \leq a\right) P\,(1+\mathcal{N} \leq b)\,. \end{aligned} \tag{3.29}$$

By Lemma 3.21, the quantity $P\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right\|^2 \leq a\right)$ equals $P\left(\|\mathrm{e}_1 + \sigma\mathrm{N}\|^2 \leq a\right)$. Therefore, $P(A_n)$ equals the constant value $P\left(\|\mathrm{e}_1 + \sigma\mathrm{N}\|^2 \leq a\right) P\,(1+\mathcal{N} \leq b)$ which implies that $\sum_{n\geq 0} P(A_n) = +\infty$. Moreover, by the same Lemma, we have the independence of the events $\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right\|^2 \leq a\right)$ and therefore that of the events A_n. Thus, the Borel-Cantelli Lemma (Lemma 3.2) can be applied and shows that the event A_n happens almost surely and then the event $\mathcal{F}_\alpha(\mathrm{X}_{n+1}) \leq \alpha\,(1+m_{\mathcal{N}}) + \epsilon$ happens almost surely. Therefore, the sequence $(\mathcal{F}_\alpha(\mathrm{X}_n))_n$ converges almost surely to $\alpha\,(1+m_{\mathcal{N}})$.

Step 3: Now we have to show that the sequence $(\|\mathrm{X}_n\|)_n$ converges to 0 almost surely. From Eq. 3.27, we have for $n \geq 0$,

$$\mathcal{F}_\alpha(\mathrm{X}_n) \geq \left(\|\mathrm{X}_n\|^2 + \alpha\right)(1+m_{\mathcal{N}})\,.$$

Using the fact that $m_{\mathcal{N}} + 1 > 0$, the previous equation implies that, for $n \geq 0$, $0 \leq \|\mathrm{X}_n\|^2 \leq \frac{\mathcal{F}_\alpha(\mathrm{X}_n)}{1+m_{\mathcal{N}}} - \alpha$. As both the right and left hand sides of this equation converge to zero, the sequence $(\|\mathrm{X}_n\|)_n$ converges also to zero. □

Lemma 3.21. Let $(\mathrm{X}_n)_n$ be the sequence of random vectors in $\mathbb{R}^d$ defined in Eq. 3.25 and $(\mathrm{N}_n)_n$ the relative sequence of independent random vectors following the same distribution $\mathrm{N}(0, I_d)$ used to define the sequence $(\mathrm{X}_n)_n$ as shown in Eq. 3.25. Then the variables $Y_n := \left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right\|$ are independent and follow the same distribution as $\|\mathrm{e}_1 + \sigma\mathrm{N}(0, I_d)\|$.

Proof :

In the beginning, let us show that, for $n \geq 0$, the random variable Y_n follows the same distribution as $\|\mathrm{e}_1 + \sigma\mathrm{N}(0, I_d)\|$. Let $t \in \mathbb{R}$, the expectation $E(e^{itY_n})$ writes as follows:

$$E(e^{itY_n}) = E\left[E\left(e^{it\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right\|} | \mathrm{X}_n\right)\right] \tag{3.30}$$

Let R_n an orthogonal transformation (rotation) such that $R_n\left(\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}\right) = \mathrm{e}_1$. The previous equation becomes:

$$E(e^{itY_n}) = E\left[E\left(e^{it\|\mathrm{e}_1 + \sigma R_n(\mathrm{N}_n)\|} | \mathrm{X}_n\right)\right]\,. \tag{3.31}$$

Applying a change of variables $U_n = R_n(\mathrm{N}_n)$, the variables U_n and N_n follow the same distribution due to the fact that the distribution of $\mathrm{N}(0, I_d)$ is spherical. Therefore, one gets:

$$E(e^{itY_n}) = E\left[E\left(e^{it\|\mathrm{e}_1 + \sigma\mathrm{N}_n\|}\right)\right] = E\left(e^{it\|\mathrm{e}_1 + \sigma\mathrm{N}(0, I_d)\|}\right)\,. \tag{3.32}$$

Now, we have to show that the variables Y_n ($n \geq 0$) are independent. Let n, m in $\mathbb{N}$ such that $m \neq n$. We suppose, without loss of generality, that $n < m$. Let $t_1, t_2 \in \mathbb{R}$. We are going to show that $E\left(e^{it_1Y_n+it_2Y_m}\right) = E\left(e^{it_1Y_n}\right) E\left(e^{it_2Y_m}\right)$. We have

$$E\left(e^{it_1Y_n+it_2Y_m}\right) = E\left[E\left(e^{it_1Y_n+it_2Y_m}|\mathrm{X}_n, \mathrm{X}_m, \mathrm{N}_n\right)\right] . \tag{3.33}$$

The random variable Y_n is $\sigma(\mathrm{X}_n, \mathrm{N}_n)$-measurable, so that

$$E\left(e^{it_1Y_n+it_2Y_m}\right) = E\left[e^{it_1Y_n} E\left(e^{it_2\left\|\frac{\mathrm{X}_m}{\|\mathrm{X}_m\|}+\sigma\mathrm{N}_m\right\|}|\mathrm{X}_n, \mathrm{X}_m, \mathrm{N}_n\right)\right] \tag{3.34}$$

Using the independence of N_m with the random vectors X_n, NN_n and X_m, we get

$$\begin{aligned} E\left(e^{it_2\left\|\frac{\mathrm{X}_m}{\|\mathrm{X}_m\|}+\sigma\mathrm{N}_m\right\|}|\mathrm{X}_n, \mathrm{X}_m, \mathrm{N}_n\right) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{it_2\left\|\frac{\mathrm{X}_m}{\|\mathrm{X}_m\|}+\sigma x\right\|} e^{-\frac{\|x\|^2}{2}} dx \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{it_2\|\mathrm{e}_1+\sigma x\|} e^{-\frac{\|x\|^2}{2}} dx \\ &= E\left[e^{it_2\|\mathrm{e}_1+\sigma\mathrm{N}_m\|}\right] . \end{aligned} \tag{3.35}$$

Therefore, we get

$$\begin{aligned} E\left(e^{it_1Y_n+it_2Y_m}\right) &= E\left(e^{it_1Y_n} E\left(e^{it_2\|\mathrm{e}_1+\sigma\mathrm{N}_m\|}\right)\right) \\ &= E\left(e^{it_2\|\mathrm{e}_1+\sigma\mathrm{N}_m\|}\right) E\left(e^{it_1Y_n}\right) \\ &= E\left(e^{it_1Y_n}\right) E\left(e^{it_2Y_m}\right) . \end{aligned} \tag{3.36}$$

□

3.3.2 Divergence in the case $-\infty \leq m_{\mathcal{N}} < -1$

In the case where $-\infty \leq m_{\mathcal{N}} < -1$, the key idea is that objective functions are negative after a finite number of iterations. This is stated in the following lemma.

Lemma 3.22. Consider the sequences $(O_n)_n$ and $(\mathrm{X}_n)_n$ defined by the recurrence relations Eq. 3.25 and Eq. 3.26 for the minimization of the objective function defined in Eq. 3.24. If $-\infty \leq m_{\mathcal{N}} < -1$, then objective functions are negative after a finite number of iterations i.e., $\exists\ n_1 \geq 0$ such that $\mathcal{F}_\alpha(\mathrm{X}_n) < 0$ for $n \geq n_1$ almost surely.

Proof :
The proof is similar to the proof of the assertion 1 of Lemma 3.7 in the non shifted case (i.e., relative to the objective function given by Eq. 3.4). Let us show that the event $\mathcal{A} := \{\exists\ n_1 \geq 0$ such that $\forall\ n \geq n_1\ \mathcal{F}_\alpha(\mathrm{X}_n) < 0\}$ is equal to the event $\mathcal{B} := \{\exists\ p_0 \geq 0$ such that $\mathcal{N}_{p_0} < -1\}$.
First, we show that $\mathcal{A} \subset \mathcal{B}$. This is equivalent to show that $\mathcal{B}^c \subset \mathcal{A}^c$. If $\forall\ p \geq 0$, $\mathcal{N}_p \geq -1$ then $\forall p \geq 0$, $\mathcal{F}_\alpha(X_p) \geq 0$ (because
$\mathcal{F}_\alpha(\|X_p\|) = (\|X_p\|^2 + \alpha)(1 + O_p) = (\|X_p\|^2 + \alpha)\left(1 + \mathcal{N}_{\psi(p)}\right)$ where $\psi(p) \leq p$ is the index of last acceptance).
Now we have to show that $\mathcal{B} \subset \mathcal{A}$: Suppose that $\exists\ p_0 \geq 0$ such that $\mathcal{N}_{p_0} < -1$. We denote by $p_1 \geq 0$ the integer defined by $p_1 = \min\{p \in \mathbb{N}$ such that $\mathcal{N}_p < -1\}$. Then

$\mathcal{F}_\alpha(X_{p_1}) < 0$ and $\mathcal{F}_\alpha(X_p) \geq 0$ for all $0 \leq p \leq p_1 - 1$. Then, as $(\mathcal{F}_\alpha(\mathrm{X}_n))_n$ is a decreasing sequence, $\forall n \geq p_1 \;\; \mathcal{F}_\alpha(\mathrm{X}_n) < 0$.
This implies that $P(\mathcal{A}) = P(\mathcal{B})$. Now, we have for all $n \geq 0$,

$$P(\mathcal{B}^c) = P(\cap_{p=0}^{+\infty}(\mathcal{N}_p \geq -1)) \leq \Pi_{i=0}^{n} P(\mathcal{N} \geq -1)) = (P(\mathcal{N} \geq -1))^n.$$

Let $a := P(\mathcal{N} \geq -1)$. As $-\infty \leq m_{\mathcal{N}} < -1$, then $a < 1$ and consequently $P(\mathcal{B}^c) = 0$ and $P(\mathcal{A}) = 1$. Then $\exists\ n_1 \geq 0$ such that $\mathcal{F}_\alpha(\mathrm{X}_n) < 0$ for $n \geq n_1$ almost surely. The sequence $(\mathcal{F}_\alpha(\mathrm{X}_n))_n$ is decreasing (because of the elitist selection). Then for $n \geq n_1$, $\mathcal{F}_\alpha(\mathrm{X}_n) \leq \mathcal{F}_\alpha(X_{n_1}) < 0$.

□

We are now ready to state the main result.

Proposition 3.23 (Divergence for $-\infty \leq m_{\mathcal{N}} < -1$)**.** Consider the sequences $(O_n)_n$ and $(\mathrm{X}_n)_n$ defined by the recurrence relations Eq. 3.25 and Eq. 3.26 for the minimization of the objective function defined in Eq. 3.24. If $m_{\mathcal{N}} + 1 < 0$ then:

1. Objective functions are negative after a finite number of iterations i.e., $\exists\ n_1 \geq 0$ such that $\mathcal{F}_\alpha(\mathrm{X}_n) < 0$ for $n \geq n_1$ almost surely.

2. For $n \geq n_1$, the sequence of the expectations of the distances squared to the optimum of the non noisy objective function is increasing in the sense that $E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \mid \mathrm{X}_n, O_n, \mathcal{N}_n\right) \geq 1$.

Therefore, for $n \geq n_1$, $E(\|\mathrm{X}_n\|^2) \geq E(\|X_{n_1}\|^2) > 0$, and the sequence $(E(\|\mathrm{X}_n\|^2))_n$ cannot converge to zero.

This result include the particular case of Gaussian noise ($m_{\mathcal{N}} = -\infty$). Therefore, in the case of a Gaussian noise, the algorithm cannot converge in the sense that the L^2-norm of the sequence $(\|\mathrm{X}_n\|)_n$ can not converge to zero. This result seems in contradiction with the result of Arnold and Beyer [8] in which they show that convergence (in expectation) occurs due to a positive expected progress rate. The reason for this apparent contradiction is due to the model investigated by Arnold and Beyer. Arnold and Beyer's model writes as:

$$\mathrm{f}(x) = \|x\|^2 \left(1 + \frac{2\sigma_\epsilon^*}{d}\mathrm{N}(0,1)\right) \tag{3.37}$$

where d is the search space dimension, σ_ϵ^* is a strictly positive constant called the normalized noise strength and $\mathrm{N}(0,1)$ is the Gaussian random variable with mean 0 and variance 1. Our study shows that whenever, a negative fintess value is sampled, the algorithm start to diverge. In [8, Fig 8], and for the values $\sigma_\epsilon^* = 2$ and $d = 80$, the probability that a negative fitness value is sampled is upper bounded by 10^{-88} as already stated in Section 3.1.4. Therefore, the average value of the moment n_1 defined in Lemma 3.22 is 10^{88}. As in practice, the algorithm does not run such a number of iterations, fitness functions values sampled are positive and a convergence is observed.
Proof :
Note that the case $\alpha = 0$ and $-\infty < m_{\mathcal{N}} < -1$ leads to a divergence of the algorithm as

already stated in Proposition 3.4. Now we investigate the more general result where $\alpha \geq 0$ and $-\infty \leq m_{\mathcal{N}} < -1$. The first point of the proof is demonstrated in Lemma 3.22. The fact that $\exists\, n_1 \geq 0$ such that $\mathcal{F}_\alpha(\mathrm{X}_n) < 0$ for $n \geq n_1$ almost surely implies that $1 + O_n < 0$ for all $n \geq n_1$. For $n \geq 0$, as $P(\|\mathrm{X}_n\| = 0) = 0$, one can divide the the acceptance event inequality (see Eq. 3.25 and Eq. 3.26) by $\|\mathrm{X}_n\|^2$. The resulting inequality writes as:

$$\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}_n\right\|^2 + \frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1 + \mathcal{N}_n) < \left(1 + \frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1 + O_n)\ .$$

In the sequel, we suppose $n \geq n_1$. We have:

$$\begin{aligned} E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right) = \\ E\left(\mathbb{1}_{\left\{\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\right\|^2+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+\mathcal{N}_n) > \left(1+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+O_n)\right\}} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right) \\ + E\left(\frac{\|\mathrm{X}_n\|^2 \left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right\|^2}{\|\mathrm{X}_n\|^2} \mathbb{1}_{\left\{\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_n\right\|^2+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+\mathcal{N}_n) < \left(1+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+O_n)\right\}} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right)\ . \end{aligned}$$

As the multivariate normal distribution is isotropic, we get

$$\begin{aligned} E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right) = \\ E\left(\mathbb{1}_{\left\{\left(\|\mathrm{e}_1+\sigma\mathrm{N}_n\|^2+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+\mathcal{N}_n) > \left(1+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+O_n)\right\}} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right) \\ + E\left(\|\mathrm{e}_1 + \sigma\mathrm{N}_n\|^2 \mathbb{1}_{\left\{\left(\|\mathrm{e}_1+\sigma\mathrm{N}_n\|^2+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+\mathcal{N}_n) < \left(1+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+O_n)\right\}} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right)\ . \end{aligned}$$

Let $\mathrm{N}_{n,1}$ denote the first coordinate of the variable N_n. The quantity $\|\mathrm{e}_1 + \sigma\mathrm{N}_n\|^2$ equals $1 + 2\sigma\mathrm{N}_{n,1} + \sigma^2\|\mathrm{N}_n\|^2$ and we have

$$\begin{aligned} E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right) = 1 \\ + \sigma^2 E\left(\|\mathrm{N}_n\|^2 \mathbb{1}_{\left\{\left(\|\mathrm{e}_1+\sigma\mathrm{N}_n\|^2+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+\mathcal{N}_n) < \left(1+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+O_n)\right\}} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right) \\ + 2\sigma E\left(\mathrm{N}_{n,1} \mathbb{1}_{\left\{\left(\|\mathrm{e}_1+\sigma\mathrm{N}_n\|^2+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+\mathcal{N}_n) < \left(1+\frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1+O_n)\right\}} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right)\ . \end{aligned}$$

For $n \geq n_1$, we have $1 + O_n < 0$. Therefore, the event

$$\left(\left(\|\mathrm{e}_1 + \sigma\mathrm{N}_n\|^2 + \frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1 + \mathcal{N}_n) < \left(1 + \frac{\alpha}{\|\mathrm{X}_n\|^2}\right)(1 + O_n)\right)$$

is equivalent to the event

$$\left\{(1 + \mathcal{N}_n < 0) \cap \left(\|\mathrm{e}_1 + \sigma\mathrm{N}_n\|^2 > A\left(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n\right)\right)\right\}$$

where $A\left(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n\right)$ is defined as $A\left(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n\right) := \left(1 + \frac{\alpha}{\|\mathrm{X}_n\|^2}\right)\frac{1+O_n}{1+\mathcal{N}_n} - \frac{\alpha}{\|\mathrm{X}_n\|^2}$. Therefore, we get:

$$
\begin{aligned}
E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right) = 1+ \\
2\sigma \mathbb{1}_{\{\mathcal{N}_n<0\}} E\left(\mathrm{N}_{n,1}\mathbb{1}_{\{1+2\sigma\mathrm{N}_{n,1}+\sigma^2\|\mathrm{N}_n\|^2>A(O_n,\|\mathrm{X}_n\|,\mathcal{N}_n)\}} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right) \\
+ \sigma^2 \mathbb{1}_{\{\mathcal{N}_n<0\}} E\left(\|\mathrm{N}_n\|^2 \mathbb{1}_{\{1+2\sigma\mathrm{N}_{n,1}+\sigma^2\|\mathrm{N}_n\|^2>A(O_n,\|\mathrm{X}_n\|,\mathcal{N}_n)\}} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right)
\end{aligned}
$$

Now, we will show that $M(\mathrm{X}_n, O_n, \mathcal{N}_n) := E\left(\mathrm{N}_{n,1}\mathbb{1}_{\{\|\mathrm{e}_1+\sigma\mathrm{N}_n\|^2>A(O_n,\|\mathrm{X}_n\|,\mathcal{N}_n)\}} \,|\, \mathrm{X}_n, O_n, \mathcal{N}_n\right) \geq 0$. The quantity $M(\mathrm{X}_n, O_n, \mathcal{N}_n)$ can be rewritten as

$$
M(\mathrm{X}_n, O_n, \mathcal{N}_n) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} x_1 \mathbb{1}_{\{\|\mathrm{e}_1+\sigma x\|^2>A(O_n,\|\mathrm{X}_n\|,\mathcal{N}_n)\}}(x) dx \,. \tag{3.38}
$$

Let O_n, $\|\mathrm{X}_n\|$ and $\mathcal{N}_n$ be fixed and let $(x_1, \ldots, x_d) \in \mathbb{R}^d$. If x_1 is such that

$$
x_1 < 0 \text{ and } 1 + 2\sigma x_1 + \sigma^2\|x\|^2 > A\left(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n\right)
$$

then

$$
1 + 2\sigma(-x_1) + \sigma^2\left((x_1)^2 + \sum_{i=2}^{d}(x_i)^2\right) \geq 1 + 2\sigma x_1 + \sigma^2\|x\|^2 > A\left(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n\right)
$$

Let $B\left(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n, x\right)$ denote the quantity $\frac{A(O_n,\|\mathrm{X}_n\|,\mathcal{N}_n)-1-\sigma^2\|x\|^2}{2\sigma}$. Then

$$
B\left(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n, (x_1, x_2 \ldots, x_d)\right) = B\left(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n, (-x_1, x_2 \ldots, x_d)\right) \,, \tag{3.39}
$$

and we have

$$
\text{if } x_1 < 0 \text{ then } \mathbb{1}_{\{x_1>B(O_n,\|\mathrm{X}_n\|,\mathcal{N}_n,(x_1,x_2,\ldots,x_d))\}} \leq \mathbb{1}_{\{-x_1>B(O_n,\|\mathrm{X}_n\|,\mathcal{N}_n,(-x_1,x_2,\ldots,x_d))\}} \,. \tag{3.40}
$$

The quantity $M(\mathrm{X}_n, O_n, \mathcal{N}_n)$ can be rewritten as

$$
\begin{aligned}
M(\mathrm{X}_n, O_n, \mathcal{N}_n) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} x_1 \mathbb{1}_{\{x_1\leq 0\}} \mathbb{1}_{\{\|\mathrm{e}_1+\sigma x\|^2>A(O_n,\|\mathrm{X}_n\|,\mathcal{N}_n)\}}(x) dx_1\right] dx_2 \ldots dx_d \\
+ \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} x_1 \mathbb{1}_{\{x_1\geq 0\}} \mathbb{1}_{\{\|\mathrm{e}_1+\sigma x\|^2>A(O_n,\|\mathrm{X}_n\|,\mathcal{N}_n)\}}(x) dx_1\right] dx_2 \ldots dx_d \,.
\end{aligned}
$$

Applying a change of variables in the second term ($u_1 = -x_1, u_2 = x_2, \ldots, u_d = x_d$), and using Eq. 3.39, one gets

$$
\begin{aligned}
M(\mathrm{X}_n, O_n, \mathcal{N}_n) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} x_1 \mathbb{1}_{\{x_1\leq 0\}} \mathbb{1}_{\{x_1>B(O_n,\|\mathrm{X}_n\|,\mathcal{N}_n,x)\}}(x) dx_1\right] dx_2 \ldots dx_d \\
+ \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} -u_1 \mathbb{1}_{\{u_1\leq 0\}} \mathbb{1}_{\{-u_1>B(O_n,\|\mathrm{X}_n\|,\mathcal{N}_n,u)\}}(u) du_1\right] du_2 \ldots du_d \,.
\end{aligned}
$$

This gives

$$M(\mathrm{X}_n, O_n, \mathcal{N}_n) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} x_1 \mathbb{1}_{\{x_1 \leq 0\}} \left(\mathbb{1}_{\{x_1 > B(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n, x)\}}(x) - \mathbb{1}_{\{-x_1 > B(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n, x)\}}(x) \right) dx_1 \right] dx_2 \dots dx_d \,.$$

By Eq. 3.40, one has $x_1 \mathbb{1}_{\{x_1 \leq 0\}} \left(\mathbb{1}_{\{x_1 > B(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n, x)\}}(x) - \mathbb{1}_{\{-x_1 > B(O_n, \|\mathrm{X}_n\|, \mathcal{N}_n, x)\}}(x) \right) \geq 0$ for all $x \in \mathbb{R}^d$. Consequently $M(\mathrm{X}_n, O_n, \mathcal{N}_n) \geq 0$ which implies that $E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \mid \mathrm{X}_n, O_n, \mathcal{N}_n \right) \geq 1$ for $n \geq n_1$. □

Chapter 4

Log-linear Behavior of the Scale-invariant $(1, \lambda)$-ES in Noisy Spherical Environments

The material in this chapter is the basis for a journal paper that we intend to submit soon. In Chapter 3, we investigated the effect of the elitist selection procedure of the scale-invariant (1 + 1)-ES when minimizing noisy objective functions. For a class of noisy objective functions with positive non-noisy part, we have shown that almost sure convergence cannot occur if negative noisy objective functions values can be sampled with a strictly positive probability. In this chapter, we investigate the behavior of the non elitist $(1, \lambda)$ isotropic ES when minimizing noisy objective functions. The adaptation rule is the scale-invariant rule (i.e., $\sigma_n = \sigma \|\mathrm{X}_n\|$) that had been previously shown to be optimal for comma strategies [17]. The general model of the noisy objective function is given by the following equation

$$\mathrm{f}(x) = \|x\|(1 + \sigma_\epsilon \mathcal{N}) \tag{4.1}$$

where $x \in \mathbb{R}^d$, $\mathcal{N}$ is an independent random variable that models the noise and σ_ϵ is a strictly positive constant which represents a scaling parameter for the noise level. We will refer to σ_ϵ as the noise strength. The noise random variable $\mathcal{N}$ is supposed to be absolutely continuous with respect to the Lebesgue measure.
Moreover, we investigate two models relative to the computation of the fitness of the offspring that we denote model **pf** and model **apf** respectively. Let $x \in \mathbb{R}^d$ denote a parent and $y \in \mathbb{R}^d$ its offspring. In the model **pf**, the fitness of the offspring is $\mathrm{f}(y) = \|y\| + \sigma_\epsilon \|y\| \mathcal{N}$. In the model **apf**, the fitness of the offspring is $\mathrm{f}(y) = \|y\| + \sigma_\epsilon \|x\| \mathcal{N}$. The model **apf** was used in [8] as a reliable approximation in the limit of infinite dimension of the search space.

The work can be divided into three parts, that we summarize below.

Part 1: Log-linear behavior for fixed finite dimension In this part, we investigate the log-linear behavior of the algorithm for a fixed search space dimension. The log-linear

behavior of the algorithm is proven in Theorem 4.8 for the models **pf** and **apf**. The result is established using the Law of Large Numbers for orthogonal random variables. The result is that $\lim_n \frac{1}{n}\ln(\|X_n\|) = F(\sigma,\sigma_\epsilon)$ or $\tilde{F}(\sigma,\sigma_\epsilon)$ where $F(\sigma,\sigma_\epsilon)$ (respectively $\tilde{F}(\sigma,\sigma_\epsilon)$) represents the convergence rate for the model **pf** (respectively **apf**). This theorem not only states that the behavior of the algorithm is log-linear (whenever the quantites $F(\sigma,\sigma_\epsilon)$ and $\tilde{F}(\sigma,\sigma_\epsilon)$ are nonzero), but also gives a quantitative information relative to the convergence (or divergence) speed that can be numerically computed (see Part 3).
.

Part 2: Infinite dimension study The hypothesis used in [8] suggests that the model **pf** is well approximated by the model **apf** for infinite dimension of the search space. In this part, we show rigorously that such an approximation is reliable when the search space dimension goes to infinity. Moreover, we investigate how the convergence rate $F(\sigma,\sigma_\epsilon)$ (or $\tilde{F}(\sigma,\sigma_\epsilon)$) varies as a function of the search space dimension d. Therefore, we investigate the limit of the so-called normalized convergence rates $dF(\sigma,\sigma_\epsilon)$ and $d\tilde{F}(\sigma,\sigma_\epsilon)$ with σ equal to $\frac{\sigma^*}{d}$ and σ_ϵ equal to $\frac{\sigma^*_\epsilon}{d}$. The strictly positive constants σ^* and σ^*_ϵ are respectively called normalized step-size mutation and normalized noise strength. The result of this computation relies on proving the uniform integrability of the underlying random variables and is given in Theorem 4.9. It is proven that the quantities $dF(\frac{\sigma^*}{d},\frac{\sigma^*_\epsilon}{d})$ and $d\tilde{F}(\frac{\sigma^*}{d},\frac{\sigma^*_\epsilon}{d})$ have the same limit, depending on λ, σ^* and σ^*_ϵ, that we will denote $l(\lambda,\sigma,\sigma^*_\epsilon)$. This result allows us to conclude that:

1. The convergence rate varies asymptotically linearly with the inverse of the search space dimension in the sense that $F(\frac{\sigma^*}{d},\frac{\sigma^*_\epsilon}{d}) \sim \frac{l(\lambda,\sigma,\sigma^*_\epsilon)}{d}$ and $\tilde{F}(\frac{\sigma^*}{d},\frac{\sigma^*_\epsilon}{d}) \sim \frac{l(\lambda,\sigma,\sigma^*_\epsilon)}{d}$.

2. The approximation used in [8] is reliable when the search space dimension goes to infinity.

Part 3: Specific case of Gaussian noise In this part, we focus on the particular case of Gaussian noise. First, we give in Theorem 4.10 a simplified expression of the limit $l(\lambda,\sigma,\sigma^*_\epsilon)$ of the normalized convergence rates $dF(\frac{\sigma^*}{d},\frac{\sigma^*_\epsilon}{d})$ and $d\tilde{F}(\frac{\sigma^*}{d},\frac{\sigma^*_\epsilon}{d})$. The proof in Theorem 4.10 uses the same techniques that were used in [25], and mainly relies on the fact that mutations follow the multivariate normal distribution.The limit of the normalized convergence rate (given in Eq. 4.20) is found to be equal to the opposite of the limit of the progress rate derived in [25]. This result generalizes the result derived in [17] for the non-noisy sphere function. Moreover, the expression derived shows that:

1. For sufficiently large search space dimensions, if $\sigma^*_\epsilon < 2c(1,\lambda)$, the algorithm converges provided that $\sigma^{*2} + \sigma^{*2}_\epsilon < 4c^2(1,\lambda)$ (strictly negative normalized convergence rate) and if $\sigma^*_\epsilon > 2c(1,\lambda)$ the algorithm diverges (strictly positive normalized convergence rate).

2. For fixed σ^* and λ, the limit of the normalized convergence rate when the search space goes to infinity is increasing as a function of σ^*_ϵ, i.e., the noise slows down a possible convergence of the algorithm or speeds up a possible divergence of the algorithm.

3. The limit of the normalized convergence rate when the search space dimension goes to infinity is a decreasing function of λ, i.e., increasing λ speed up a possible convergence of the algorithm.

Second, in the divergence case given by $\sigma_\epsilon^* > 2c(1,\lambda)$, we compare the strategies of

1. increasing λ,

2. re-sampling the offspring fitness N times and averaging its fitness through the N samplings.

By increasing λ or averaging (which decreases the normalized noise strength from σ_ϵ^* to $\frac{\sigma_\epsilon^*}{\sqrt{N}}$) one can be in the convergence situation given by $\sigma_\epsilon^* < 2c(1,\lambda)$. It is established, for sufficiently large values of σ_ϵ^*, that it is better for the $(1,\lambda)$-ES (in term of evaluation cost per generation), to reevaluate the offspring fitness than to increase the number of offspring λ. Note that a similar study had been previously done in [25].
Third, a contribution of this study is Theorem 4.8, which has been derived using a LLN for orthogonal random variables, and gives the explicit expression of the convergence (or divergence) rate. This expression is given in terms of an expectation of an underlying random variable and therefore, according to the LLN, can be numerically computed using Monte Carlo simulations. Monte Carlo simulations of the normalized convergence rates are plotted as a function of the normalized step-size mutation for different normalized noise strengths, different dimensions and both models **pf** and **apf**. Strictly positive (respectively negative) values of the normalized convergence rate mean that the algorithm converges (respectively diverges). In particular, it can be seen that for almost all parameter settings (normalized step-size muation, normalized noise strength, number of offspring), the convergence rate is nonzero, which gives the log-linear behavior of the algorithm.
Fourth, curves representing the normalized convergence rates for finite dimensions and the infinite dimension ($d \to +\infty$) are plotted (Figures 4.5 and 4.6) as a function of the normalized step-size mutation for the models **pf** and **apf** and two values of the normalized noise strength. These plots reveal that, for same parameter values of the algorithm and of the normalized strength, finite convergence rates can have strictly negative sign, suggesting a convergence of the algorithm, while the limit expression of the convergence rate is strictly positive, suggesting the divergence of the algorithm in the limit of infinite dimensions. Therefore, infinite dimension results have to be taken with care in some cases. Moreover, the comparison of the curves relative to the model **pf** to those relative to the model **apf** reveals that, for the same parameter values and finite dimensions, convergence can be predicted for one of the two models, while divergence occurs for the other model. These two observations prove the limits of adopting, for finite dimensions, infinite dimension results and for approximating the model **pf** by the model **apf**.

Finally, optimal convergence rates, optimal normalized step-size mutations, and upper bounds for the step-size mutation allowing to have a convergence of the algorithm are plotted, for finite and infinite dimensions, as a function of the normalized noise strength σ_ϵ^*.

Log-linear Behavior of the Scale-invariant $(1, \lambda)$-ES in Noisy Spherical Environments

4.1 Introduction

Optimization is a recurrent task in engineering problems and a research field investigated by applied mathematicians and by computer scientists as well. Mathematically speaking, the goal is to minimize (or maximize [1]) a real valued function f, called objective function, and defined on a search space Ω. The general context of this chapter is non linear unconstrained continuous optimization. This means that f is non linear, the search space Ω is non restricted and is (or contains) one or many open subsets of $\mathbb{R}^d$.
The difficulty of an unconstrained optimization problem is related to the dimension of the search space Ω and to the characteristics of the underlying objective function f. In real-world optimization problems, objective functions can be non-convex, non-smooth, discontinuous, noisy, multi-modals, ill-conditioned, non separable .... The algorithms developed to solve these problems explore the search space by generating, at each iteration, new trial point(s) either deterministically or randomly using some search distribution. Randomized search methods are well known global methods which prove to be more robust than deterministic search methods when optimization problems are 'difficult' [9, 106, 78]. Randomized search methods designed for continuous optimization include Pure Random Search (PRS) [31], Pure Adaptive Search (PAS) [148], Evolution Strategies (ES) [25], Differential Evolution (DE) [131, 132, 133], Particle Swarm Optimization (PSO) [34, 81, 126, 127], (continuous) Estimation of Distribution Algorithms (continuous (EDA)) [91] and Simulated Annealing (SA) [3][2]. According to the comparison of some widely used continuous randomized search methods which has been done during the Congress of Evolutionary Computation (CEC 2005) [2], the state of the art of ES called Covariance Matrix Adaptation-Evolution Strategy (CMA-ES) was highly competitive by solving all problems including multi-modal problems and robust as its performance was not affected by non-separability or non-convexity. Moreover, the performance of CMA-ES degrades slower than performance of the other methods when the test function is being less and less conditioned.

Pure Random Search is the simplest randomized search method. At each iteration, trial points are independent identically distributed (i.i.d.) and the best solution is retained. In particular, points are always sampled around the same point and the search

[1] Minimizing a real valued function f is equivalent to maximize $-f$.

[2] Simulated Annealing can be seen as a particular ES with a randomized rule for the acceptance of a new trial point.

distribution parameters such as the radius and favorite directions of the search are unchanged during the optimization process. It has been proven [149] that this intuitive procedure ensures a global convergence in the sense that the algorithm converges to the global minimum with probability 1 for every objective function for which the neighborhood of the global optimum can be reached with strictly positive probability. However, the research parameters of Pure Random Search are not adapted relatively to the history of the search and/or the shape of the objective function. Thus, its convergence time is very large increasing exponentially with the search space dimension. This makes Pure Random Search not useful in practice on some problems presenting a structure that could be exploited. This exponential dependency of the convergence time of the PRS with respect to the dimension is decreased to a linear dependency for the so-called Pure Adaptive Search (PAS) [148] but the PAS is an algorithm, not only move iff a better point is sampled but also does not adapt its research parameters. The run time will be then very large in practice.

On the other hand, ES, which are evolutionary algorithms (EA) designed for continuous optimization, were successful due to the adaptation mechanisms of research parameters they implement. ES, as other EA, use bio-inspired techniques at each iteration (called also generation) to evolve a set (or population) of solutions. Solutions in the beginning of an iteration n are called parents. Then the search step is based on the so-called mutations. A mutation is a perturbation of a parent which corresponds to adding a random sampling of a multivariate normal distribution. The resulting point is called offspring. At an iteration n, let X_n be the parent, the offspring Y_n equals

$$Y_n = \mathrm{X}_n + \sigma_n \mathrm{N}(0, C_n), \tag{4.2}$$

where σ_n is a strictly positive constant and $\mathrm{N}(0, M)$ denotes a sampling of the multivariate normal distribution with mean $(0, \ldots, 0) \in \mathbb{R}^d$ and a covariance matrix M. The parameter σ_n and the matrix C_n are the search distribution parameters. The parameter σ_n corresponds to the 'radius' of the search and is called the step-size mutation. The matrix C_n gives the favorite directions of the search at the iteration n and is abusively called the covariance matrix of the mutation. An efficient ES has to adapt its research parameters (σ_n and C_n) based on the history of the search. The simplest ES, is the so-called $(1+1)$-ES, which evolves a single solution and accept, at each iteration, the new trial point iff it is better than the previous sampled points. If the step-sizes σ_n $(n \geq 0)$ are set equals to a constant σ_0 and the covariance matrices C_n are set equal to the identity matrix of $\mathbb{R}^d$ which we denote I_d[3], it has been shown [117, 33] that almost sure convergence toward the global optimum holds when the objective function is continuous. If the step-sizes σ_n $(n \geq 0)$ are deterministically updated, it has been shown that global convergence[4] holds for isotropic ES whenever a sufficient condition on the sequence of step-sizes is satisfied [150]. Several adaptation schemes have been introduced. The one-fifth success rule [114, 82] is the oldest known technique which adapts only the step-size. Self-adaptive Strategies [114] and Meta-ES [63] employ the evolution itself to adjust the search parameter values. The state of the art of adaptive ES is the CMA-ES [61, 59, 57, 16] in which

[3]ES with $C_n = I_d$ are called isotropic ES.

[4]Global convergence studies refer to theoretical studies where objective function is not subject to many hypothesis and in particular these studies concern multi-modal functions.

the step-size and all the directions of the search are updated at each iteration.
The adaptation in ES makes them practically more effective and more rapid than PRS as it is the case of CMA-ES for which it is stated in [16] that: "On Convex-quadratic functions, the adaptation mechanisms for σ and C allow to achieve log-linear convergence after an adaptation time which scales between 0 and the search space dimension squared". The log-linear convergence, numerically observed in many numerical studies of optimization using ES, means that the logarithm of the distance to the optimum decreases linearly with the number of iterations after an adaptation time. Mathematically speaking, if we denote d_n the distance of the solution at an iteration n to the optimum, the (log)-linear (asymptotic) convergence means that

$$\lim_{n\to\infty} \frac{1}{n} \ln(d_n) = c \tag{4.3}$$

for some $c \neq 0$. The limit c is called convergence rate (of the sequence $(\ln(d_n))_n$). The sequence $(d_n)_n$ converges whenever $c < 0$. If $c > 0$, the algorithm diverges. It has been proven [17, 77] for isotropic ES that the convergence of ES on uni-modal objective functions is at most log-linear for any adaptation scheme of the sequence (σ_n) and that the optimal convergence rate is reached for a specific objective function and a specific adaptation rule of the sequence (σ_n). The specific objective function is the so-called sphere function which is the function mapping $\mathbb{R}^d$ into $\mathbb{R}$ and defined as $f(x) = \|x\|^2$ for $x \in \mathbb{R}^d$ where $\|.\|$ denotes the euclidean norm on $\mathbb{R}^d$. The minimum of this function is reached on $(0, \ldots, 0)$. The specific adaptation rule is the so-called scale-invariant technique for which the step-size is set proportionally to the distance to the optimum at each iteration[5] in the sense that, in the case of an optimum in $(0, \ldots, 0)$, this rule writes as

$$\sigma_n = \sigma \|\mathrm{X}_n\| \,, \tag{4.4}$$

where σ is a strictly positive constant called normalized step-size mutation. This adaptation rule has been widely investigated in the context of progress rate theory [114, 25] in which the exact expression of the scale-invariant mutation is to set σ in Eq. 4.4 equal to $\frac{\sigma^*}{d}$ where σ^* is a strictly positive constant called the normalized step-size mutation[6]. In progress rate theory, the goal is to maximize the expected progress to the optimum at each iteration (called progress rate) and the results derived hold asymptotically in the dimension of the search space. In the case of a realistic adaptation rule, an idea proposed in [27] of investigating the stability of Markov chains relative to the ES dynamics to study their behavior was exploited in [13] to rigorously prove that isotropic ES do converge (or diverge) log-linearly when minimizing the sphere function.

Noisy objective functions are frequently encountered in real-world optimization problems. Noise can have various origins as physical measurement limitations or the use of stochastic simulation procedures such as Monte-Carlo simulations. Note that these examples share the property that a reevaluation of a same solution lead to different objective

[5] This rule is artificial as in practice one does not know the optimum location.

[6] Note that we used the same terminology 'normalized step-size mutation' to denote σ in Eq. 4.4 and σ^* when σ is replaced by $\frac{\sigma^*}{d}$.

function values. Therefore, the noise investigated here is random.
The problem when dealing with noisy objective functions is that the noisy part of the function can deceive the decision making. The comparison of two solutions is no more reliable: the noisy objective function of a solution with low noise value can be better than the objective function value of a solution with a better (ideal) function value but large noise value. If this event happens frequently, the algorithm may diverges. Beyer [24] noticed that the behavior of evolutionary algorithms in noisy environments is similar, independently of the nature of the search space (continuous or discrete): noisy objective functions lead to the decrease of the convergence speed and to a deterioration of the final optimum location quality. ES are robust when solving noisy optimization problems [9, 106] compared to other deterministic or randomized search methods. In [9], it is shown that ES perform better than some deterministic method which can stagnate. In particular, it is shown that, for large values of noise, ES can perform even better than the implicit filtering method [47, 80] which belongs to the field of stochastic approximation algorithms [115, 83, 86, 87] which are optimization methods specifically designed for the optimization of stochastic and in particular noisy objective functions. In [106], it is stated that ES perform the best among population-based methods on noisy environments. However, there are few rigorous mathematical studies of the convergence of ES with respect to the noise properties. Theoretical studies of ES in presence of noise have been carried out by Rechenberg [114], Arnold and Beyer [25, 7, 5, 10, 6, 8, 24], using asymptotic estimations when the dimension of the search space tends to infinity. In [8], the noisy objective function used by the authors is

$$\|x\|^2 \left(1 + \frac{2\sigma_\epsilon^*}{d}\mathcal{N}\right) \tag{4.5}$$

in which the term σ_ϵ^* is a strictly positive constant called normalized noise strength and $\mathcal{N}$ is a Gaussian variable. Note that the noise model here is multiplicative, i.e., the noise is the ratio between the noisy and ideal objective function. It is worth noticing that a multiplicative noise model is a realistic model for modeling the noise, as the performance of the algorithm depends on how the noisy value of the objective function compares to the ideal value. Moreover, an hypothesis of an additional noise with a fixed variance will lead to a random behavior of the algorithm when the ideal objective functions values become, after some iterations, very small compared to the noise variance. In our study, we theoretically investigate the behavior of the so-called $(1, \lambda)$-ES[7] using the optimal scale-invariant adaptation rule on the minimization of an objective function perturbed by a multiplicative noise. The noisy objective function investigated here has a similar expression to that of Eq. 4.5 and simplifies to the function $f(x) = \|x\|$ in the absence of noise. We will denote the non-noisy function $f(x) = \|x\|$ the sphere function and the relative noisy function, that we will investigate here, noisy sphere function. Note that in general, the terminology 'sphere function' is in general used to denote the function $f(x) = \|x\|^2$, but in our case we used this terminology to refer to $f(x) = \|x\|$. The study is similar for the two functions. We investigate two noise models relative to two ways the offspring objective function

[7] In an iteration of the $(1, \lambda)$-ES, the new parent is the best offspring among the λ offspring newly generated.

computation is done. In the first model, the noise level of the offspring is proportional to its (ideal) objective function $f(x) = \|x\|$. The second model has been used by Arnold and Beyer in [8] as a reliable approximation of the first one for high search space dimensions: using the scale invariant algorithm with a Gaussian noise distribution for the noisy objective function, they claim that the noise level of an offspring (which corresponds to the standard deviation of the noise distribution) is well approximated by that of its parent when the search space dimension d goes to infinity. The first model will be referred to as model **pf** and the second one will be denoted model **apf**.
The behavior of ES on noisy objective functions is important to study. The randomized part of these functions covers many real objective function cases for which a little information is given and therefore any kind of irregularity is included on this kind of functions.

In this chapter, we want to see if, similarly to the non-noisy case, the behavior of the scale-invariant $(1,\lambda)$-ES is log-linear on the noisy spherical objective functions. For this purpose, we introduce in Section 4.2 the mathematical model for the objective function and the scale-invariant $(1,\lambda)$-ES minimizing this function with its two versions relative to the models **pf** and **apf**. In Sections 4.3 and 4.4 we investigate the log-linear behavior of the algorithm and derive the convergence (or divergence) rate in Theorem 4.8. Section 4.5 is dedicated to the study of the dependency of the convergence rate of the algorithm with respect of the search space dimension: we compute a common limit (Theorem 4.9) for the two models **pf** and **apf** of the so-called normalized convergence (or divergence) rate when the search space dimensions goes to infinity and derive its expression on the specific case of Gaussian noise (Theorem 4.10). In Section 4.6, the distinction between the cases where convergence or divergence happens is investigated theoretically and numerically for finite and infinite dimension cases. Note finally that for the sake of readability, most of the proofs of this chapter are sent into an appendix section.

4.2 Mathematical model for the scale-invariant $(1,\lambda)$-ES minimizing noisy sphere functions

4.2.1 Objective function model

The general noisy spherical model investigated here is the multiplicative noise model which writes as

$$\mathrm{f}(x) = \|x\|(1 + \sigma_\epsilon \mathcal{N}) \tag{4.6}$$

where $x \in \mathbb{R}^d$, $\mathcal{N}$ is an independent random variable that models the noise and σ_ϵ is a strictly positive constant which represents the scaling parameter for the noise level. We will refer to σ_ϵ as the noise strength. The noise random variable $\mathcal{N}$ is supposed to be absolutely continuous with respect to the Lebesgue measure. Its probability density function is denoted $p_{\mathcal{N}}$. The expression of the noise level $\sigma_\epsilon\|x\|$ conveys the idea of setting the variance of the noise proportional to the (ideal) objective function which is the sphere function $\|x\|$ here.

In our study, we investigate two noise models relative to two different expressions for

the computation of the offspring objective function. Let x denote the parent and y its offspring. The model **pf** is the original model given by Eq. 4.6 and then verifying that the noise level of the offspring is proportional to its (ideal) objective function, i.e., the fitness of the offspring y writes as $\|y\| + \sigma_\epsilon \|y\| \mathcal{N}$. The model **apf** is relative to the approximation used by Arnold and Beyer in [8]. In fact, Arnold and Beyer [8] state that for high dimension of the search space the parent and its offspring are so close that the noise level of the offspring (which is $\sigma_\epsilon \|y\| \mathcal{N}$ in the original model **pf**) will be well approximated by that of its parent, i.e., $\sigma_\epsilon \|x\| \mathcal{N}$. Thus, the fitness of the offspring y in model **apf** equals $\|y\| + \sigma_\epsilon \|x\| \mathcal{N}$. The model (**apf**) was also investigated in [136] as a model where the noise level is scaled proportionally to the step-size mutation.

4.2.2 The algorithm: the scale-invariant $(1, \lambda)$-ES minimizing the objective function defined in Eq. 4.6

In the context of minimization of a real valued function defined on a continuous subset of $\mathbb{R}^d$ $(d \geq 1)$, the $(1, \lambda)$-ES is a simple ES which evolves a single solution. The solution at an iteration n is the parent denoted X_n. An iteration n of a $(1, \lambda)$-ES is composed of three steps:

• **Search step:**
In this step, λ mutations are performed as in Eq. 4.2 resulting on λ new trial points (the offspring) $\mathrm{Y}_{i,n} := \mathrm{X}_n + \sigma_n \mathrm{N}_{i,n}(0, I_d)$, $i = 1, \ldots, \lambda$. The quantities $\mathrm{N}_{i,n}(0, I_d)$, $i = 1, \ldots, \lambda$ are independent realizations of the multivariate isotropic normal distribution on $\mathbb{R}^d$, $\mathrm{N}(0, I_d)$, which we will denote $\mathrm{N}^{(d)}$. For $d = 1$, $\mathrm{N}^{(1)}$ will be simply denoted NUsing a scale-invariant mutation described in Eq. 4.4, the expressions of the offspring can be rewritten as: $\mathrm{Y}_{i,n} = \mathrm{X}_n + \sigma \|\mathrm{X}_n\| \mathrm{N}_{i,n}(0, I_d)$, $i = 1, \ldots, \lambda$.

• **Evaluation step:**
In this step, objective functions of the offspring created are computed. The noisy objective function of an offspring $\mathrm{Y}_{i,n}$ denoted, according to the model used, $\tilde{\mathrm{f}}(\mathrm{Y}_{i,n})$ or $\mathrm{f}(\mathrm{Y}_{i,n})$ is then defined as

$$\mathrm{f}(\mathrm{Y}_{i,n}) = \|\mathrm{Y}_{i,n}\| + \sigma_\epsilon \|\mathrm{Y}_{i,n}\| \mathcal{N}_{i,n} \,, \tag{4.7}$$

for the model **pf**, and

$$\tilde{\mathrm{f}}(\mathrm{Y}_{i,n}) = \|\mathrm{Y}_{i,n}\| + \sigma_\epsilon \|\mathrm{X}_n\| \mathcal{N}_{i,n} \,, \tag{4.8}$$

for the model **apf** where, for $n \in \mathbb{N}$ and i an integer in $[1, \lambda]$, the random variables $\mathcal{N}_{i,n}$ are independent realizations of the (noise) random variable $\mathcal{N}$. In Eq. 4.8, we have used a tilde for the notation of the fitness function of the offspring for the model **apf**, which is denoted without a tilde for the model **pf**. In the sequel, we will use the same convention, i.e., use tilde for quantities relative to the model **apf**.

• **Selection step:**

In this step, only the best offspring (according to its objective function value) is kept as the new parent X_{n+1}. This means that X_{n+1} equals $\mathrm{Y}_{*,n}$ which verifies $\mathrm{f}(\mathrm{Y}_{*,n}) =$

$\min\{\mathrm{f}(\mathrm{Y}_{i,n})\ ,\ i=1,\ldots,\lambda\}$ if model **pf** is used and $\tilde{\mathrm{f}}(\mathrm{Y}_{*,n}) = \min\{\tilde{\mathrm{f}}(\mathrm{Y}_{i,n})\ ,\ i=1,\ldots,\lambda\}$ if model **apf** is used. For this chosen offspring the random vector (r.vec.) $\mathrm{N}^{(d)}_{*,n}$ and the random variable (r. var.) $\mathcal{N}_{*,n}$ are then implicitely defined by

$$\left\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}^{(d)}_{*,n}\right\|(1+\sigma_\epsilon\mathcal{N}_{*,n}) = \min_{1\le i\le\lambda}\left\{\left\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}^{(d)}_{i,n}\right\|(1+\sigma_\epsilon\mathcal{N}_{i,n})\right\} \tag{4.9}$$

if the model **pf** is used. For the model **apf** the previous equation writes

$$\left\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}^{(d)}_{*,n}\right\| + \sigma_\epsilon\|\mathrm{X}_n\|\mathcal{N}_{*,n} = \min_{1\le i\le\lambda}\left\{\left\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}^{(d)}_{i,n}\right\| + \sigma_\epsilon\|\mathrm{X}_n\|\mathcal{N}_{i,n}\right\}. \tag{4.10}$$

In other words, the random vector $\mathrm{N}^{(d)}_{*,n}$ and the random variable $\mathcal{N}_{*,n}$ are the instance that gave the best offspring. According to this three steps, the mathematical formulation of the algorithm is as follows: let $\mathrm{X}_0 \in \mathbb{R}^d$ be the first parent randomly chosen with the condition $P(\mathrm{X}_0 = 0) = 0$. Then an iteration of the scale-invariant $(1,\lambda)$-ES algorithm designed for the minimization of the function defined in Eq. 4.6 writes for $n \ge 0$ as:

$$\mathrm{X}_{n+1} = \mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}^{(d)}_{*,n}\,, \tag{4.11}$$

where $\mathrm{N}^{(d)}_{*,n}$ is defined in Eq. 4.9 and Eq. 4.10 according to the model used.

In section 4.4, we investigate the stability of the sequence X_n for the models **apf** and **pf** and derive the convergence theorem (Theorem 4.8). In section 4.5, we compute the limit for d going to infinity of the so-called normalized convergence rate derived from the expectation given in Theorem 4.8 using normalizations of the progress rate theory including Arnold and Beyer [25, 8] normalizations for the noise.

4.3 Definitions and preliminary results

In the sequel, e_1 will denote the unitary vector in $\mathbb{R}^d$ $(1,0,\ldots,0)$ and $\mathrm{Pr}\,(E)$ the probability of an event E. Moreover, let $\lambda \in \mathbb{N}^*$, $(M_i)_{1\le i\le\lambda}$ be λ random variables (or vectors) and R be a random variable or a real valued function. The argmin of the variables $R(M_i)$ $(i \in \{1,\ldots,\lambda\})$ is the random variable (or vector) M_* which lies in the set $\{M_i,\ i=1,\ldots,\lambda\}$ and which verifies $R(M_*) = \min_{\{1\le i\le\lambda\}}\{R(M_i)\}$.will also use the following definition.

Definition 4.1.

1. We define the maps H (relative to the model **pf**) and $\tilde{\mathrm{H}}$ (relative to the model **apf**) on $\mathbb{N}^*\times\mathbb{R}^d\times[0,+\infty[\times[0,+\infty[$ into $\mathbb{R}^+$ as the following:

   $$\mathrm{H}(\lambda,x,\sigma,\sigma_\epsilon) = \lambda\int_{\mathbb{R}}\mathrm{Pr}^{\,\lambda-1}\left[\|\mathrm{e}_1+\sigma x\|(1+\sigma_\epsilon y)\le\|\mathrm{e}_1+\sigma\mathrm{N}^{(d)}\|(1+\sigma_\epsilon\mathcal{N})\right]p_{\mathcal{N}}(y)dy\,,$$

   and

   $$\tilde{\mathrm{H}}(\lambda,x,\sigma,\sigma_\epsilon) = \lambda\int_{\mathbb{R}}\mathrm{Pr}^{\,\lambda-1}\left[\|\mathrm{e}_1+\sigma x\|+\sigma_\epsilon y\le\|\mathrm{e}_1+\sigma\mathrm{N}^{(d)}\|+\sigma_\epsilon\mathcal{N}\right]p_{\mathcal{N}}(y)dy\,.$$

2. Let $(\mathrm{N}_i^{(d)})_{i\in[1,\lambda]}$ (resp. $(\mathcal{N}_i)_{i\in[1,\lambda]}$) be λ independent samplings of $\mathrm{N}^{(d)}$ (resp. $\mathcal{N}$). We define the random vector $(\mathrm{N}_*^{(d)},\mathcal{N}_*)$ as the argmin of the variables $\{\|\mathrm{e}_1+\sigma\mathrm{N}_i^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_i)\,,\ i=1,\ldots,\lambda\}$ if model **pf** and as the argmin of the variables $\{\|\mathrm{e}_1+\sigma\mathrm{N}_i^{(d)}\|+\sigma_\epsilon\mathcal{N}_i,\ i=1,\ldots,\lambda\}$ if model **apf**.

In this context, we have the following lemma.

Lemma 4.2. Let H and $\tilde{\mathrm{H}}$ be the functions introduced in Definition. 4.1 and $\mathrm{N}_*^{(d)}$ the random vector introduced in the same Definition. Then the probability density function of the random vector $\mathrm{N}_*^{(d)}$ is defined, for a given $(\lambda,\sigma,\sigma_\epsilon)\in\mathbb{N}^*\times[0,+\infty[\times[0,+\infty[$, as

$$\frac{1}{(2\pi)^{d/2}}e^{-\frac{\|x\|^2}{2}}\mathrm{H}(\lambda,x,\sigma,\sigma_\epsilon),\ x\in\mathbb{R}^d \tag{4.12}$$

if model **pf** and

$$\frac{1}{(2\pi)^{d/2}}e^{-\frac{\|x\|^2}{2}}\tilde{\mathrm{H}}(\lambda,x,\sigma,\sigma_\epsilon),\ x\in\mathbb{R}^d \tag{4.13}$$

if model **apf**. Moreover, we introduce the functions F and $\tilde{\mathrm{F}}$ mapping $[0,+\infty[\times[0,+\infty[$ into $\mathbb{R}$ as follows:

$$\begin{aligned}\mathrm{F}(\sigma,\sigma_\epsilon)&:=E\left[\ln(\|\mathrm{e}_1+\sigma\mathrm{N}_*^{(d)}\|)\right]\\&=\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\ln(\|e_1+\sigma x\|)e^{-\frac{\|x\|^2}{2}}\mathrm{H}(\lambda,x,\sigma,\sigma_\epsilon)dx\end{aligned} \tag{4.14}$$

where $\mathrm{N}_*^{(d)}$ is defined according to model **pf** and

$$\begin{aligned}\tilde{\mathrm{F}}(\sigma,\sigma_\epsilon)&:=E\left[\ln(\|\mathrm{e}_1+\sigma\mathrm{N}_*^{(d)}\|)\right]\\&=\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\ln(\|e_1+\sigma x\|)e^{-\frac{\|x\|^2}{2}}\tilde{\mathrm{H}}(\lambda,x,\sigma,\sigma_\epsilon)dx\end{aligned} \tag{4.15}$$

where $\mathrm{N}_*^{(d)}$ is defined according to model **apf**. Then the functions F and $\tilde{\mathrm{F}}$ are well defined, continuous on $[0,+\infty[\times[0,+\infty[$ (endowed with the usual compact topology).

In order to take advantage of the fact that the random vector $\mathrm{N}^{(d)}$ has a spherical distribution, the following definition will be useful in the sequel.

Definition 4.3. Let $(\mathcal{N}_{\{n,i\}})_{n,i}\in\mathbb{R}^d$ a sequence of independent random vectors on $\mathbb{R}^d$ following the same distribution $\mathrm{N}^{(d)}$. Let also $(\mathcal{N}_{i,n})_{i,n}$ (i an integer in $[1,\lambda]$) be a sequence of independent identically distributed random variables (i.i.d.) with common law $\mathcal{N}$. We define the random vector $(\mathrm{U}_{*,n}^{(d)},\mathcal{V}_{*,n})$ as the argmin of the variables $\{\|\mathrm{e}_1+\sigma\mathrm{N}_{i,n}^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_{i,n})\,,\ i=1,\ldots,\lambda\}$ if model **pf** and as the argmin of the variables $\{\|\mathrm{e}_1+\sigma\mathrm{N}_{i,n}^{(d)}\|+\sigma_\epsilon\mathcal{N}_{i,n},\ i=1,\ldots,\lambda\}$ if model **apf**. Let σ a positive constant. We define the random sequence $(Z_n)_{n\geq0}$ as follows

$$Z_n:=\ln(\|\mathrm{e}_1+\sigma\mathrm{U}_{*,n}^{(d)}\|)-\mathrm{F}_\times(\sigma,\sigma_\epsilon)$$

where $\mathrm{F}_\times(\sigma,\sigma_\epsilon)$ is defined by Eq. 4.14 if model **pf** and by Eq. 4.15 if model **apf**.

Note that $\mathrm{U}_{*,n}^{(d)}$ is distributed as $\mathrm{N}_{*}^{(d)}$ introduced in Definition 4.1.

Remark 4.3.1. Note that in Definition 4.3, we have used the notation "$\mathrm{F}_{\times}(\sigma,\sigma_\epsilon)$" to refer to the quantity $\mathrm{F}(\sigma,\sigma_\epsilon)$ for the model **pf** and to the quantity $\tilde{\mathrm{F}}(\sigma,\sigma_\epsilon)$ for the model **apf**. In the sequel, we will use the same convention, i.e., the notation $A_\times$ will refer to a quantity A relative to the model **pf** and to a quantity $\tilde{A}$ relative to the model **apf**.

4.4 Log-Linear behavior of the scale-invariant $(1,\lambda)$-ES minimizing the objective function (Eq. 4.6)

The proof of the log-linear convergence for ES relies on the application of the Strong Law of Large Numbers (LLN) for independent or orthogonal random variables or for Markov chains.The following proposition is a key (classical) idea for the study of the stability of the sequence $\left(\ln\left(\|\mathrm{X}_n\|\right)\right)_n$ where $(\mathrm{X}_n)_n$ is defined by Eq. 4.11.

Proposition 4.4. Let $(\mathrm{X}_n)_n$ be the sequence of random vectors valued in $\mathbb{R}^d$ satisfying the recurrence relation Eq. 4.11. Then for all indices n, we have

$$\frac{1}{n}\ln\left(\frac{\|X_n\|}{\|X_0\|}\right)=\frac{1}{n}\sum_{k=0}^{n-1}\ln\left(\left\|\frac{X_k}{\|X_k\|}+\sigma\mathrm{N}_{*,k}^{(d)}\right\|\right)\ a.s. \tag{4.16}$$

where the random vectors $(\mathrm{N}_{*,n}^{(d)})_n$ satisfy Eq. 4.9 if the model is **pf** and Eq. 4.10 if the model is **apf**.

To compute the limit of the right hand side of Eq. 4.16, we will apply the following LLN for orthogonal random variables derived from [93, p. 458].

Theorem 4.5 (LLN for Orthogonal Random Variables)**.** Let $(Y_n)_{n\geq 0}$ be a sequence of identically distributed real random variables with finite variance and orthogonal, i.e., for all indices i, j, with $i\neq j$ one has $E(Y_i)=0$, $E(Y_i^2)<+\infty$ and $E(Y_iY_j)=0$. Then

$$\lim_n \frac{1}{n}\sum_{k=0}^{n-1}Y_k=0\quad a.s.$$

This theorem will be applied for the random variables $(Y_n)_{n\geq 0}$ that we introduce in the following definition.

Definition 4.6. Let $(\mathrm{X}_n)_n$ be the sequence of random vectors defined in Eq. 4.11, σ and σ_ϵ be striclty positive constants. Let also $F_\times$ be the function equal to the function F given in Lemma 4.2 and $(\mathrm{N}_{*,n}^{(d)})_n$ be the sequence of random variables given in in Eq. 4.9 if model **pf** is used; and $F_\times$ be the function equal to the function $\tilde{\mathrm{F}}$ given in Lemma 4.2 and $(\mathrm{N}_{*,n}^{(d)})_n$ be the sequence of random variables given in in Eq. 4.10 if model **apf** is used. We introduce the sequence of random variables $(Y_n)_n$ as the following: for $n\geq 0$,

$$Y_n:=\ln\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathrm{N}_{*,n}^{(d)}\right\|\right)-F_\times(\sigma,\sigma_\epsilon)\,. \tag{4.17}$$

In the following proposition, we show that the sequence $(Y_n)_n$ introduced in Definition 4.6 satisfies the assumptions of Theorem 4.5.

Proposition 4.7. Let $(Y_n)_n$ be the sequence of random variables in Definition 4.6. The followings hold:

1. For $n \geq 0$, $E(Y_n) = 0$ and $E(|Y_n|^2) < +\infty$.
2. The random variables Z_n $(n \geq 0)$ introduced in Definition 4.3 are identically distributed and for every $n \geq 0$, Y_n and Z_n follow the same distribution.
3. The sequence of random variables $(Y_n)_{n\geq 0}$ is orthogonal, i.e., for all indices i, j, with $i \neq j$ one has $E(Y_i) = 0$, $E(Y_i^2) < +\infty$ and $E(Y_i Y_j) = 0$.

Then the following theorem holds as a consequence of Theorem 4.5, Proposition 4.7 and Proposition 4.4.

Theorem 4.8 (Log-linear behavior of the scale-invariant $(1, \lambda)$-ES minimizing the objective function (Eq. 4.6))**.** The scale-invariant $(1, \lambda)$-ES minimizing the noisy sphere function defined in Eq. 4.6 converges (or diverges) log-linearly in the sense that for σ and σ_ϵ strictly positive the sequence $(\mathrm{X}_n)_n$ of random vectors given by the recurrence relation Eq. 4.11 verifies the following equations

$$\begin{aligned} \lim_n \frac{1}{n} \ln\left(\|\mathrm{X}_n\|\right) &= \mathrm{F}(\sigma, \sigma_\epsilon) \qquad \text{if model } \mathbf{pf} \text{ is used,} \\ \lim_n \frac{1}{n} \ln\left(\|\mathrm{X}_n\|\right) &= \tilde{\mathrm{F}}(\sigma, \sigma_\epsilon) \quad \text{if model } \mathbf{apf} \text{ is used,} \end{aligned} \tag{4.18}$$

almost surely, with F and $\tilde{\mathrm{F}}$ defined in Eq. 4.14 and Eq. 4.15.

Theorem 4.8 states that the convergence (or divergence) rate of the scale-invariant $(1, \lambda)$-ES minimizing the noisy sphere function given in Eq. 4.6 (or equivalently the convergence (or divergence) rate of the sequence $\left(\ln\left(\|\mathrm{X}_n\|\right)\right)_n$) is $\mathrm{F}(\sigma, \sigma_\epsilon)$ if model **pf** and $\tilde{\mathrm{F}}(\sigma, \sigma_\epsilon)$ if model **apf**. According to Eq. 4.3, the log-linear behavior holds if the convergence (or divergence) rates $\mathrm{F}(\sigma, \sigma_\epsilon)$ and $\tilde{\mathrm{F}}(\sigma, \sigma_\epsilon)$ are non zero. If $\mathrm{F}_\times(\sigma, \sigma_\epsilon) < 0$, the sequence $(\|\mathrm{X}_n\|)_n$ converges log-linearly to the optimum and if $\mathrm{F}_\times(\sigma, \sigma_\epsilon) > 0$ the algorithm diverges log-linearly. Fortunately, these quantities can be numerically computed using Monte Carlo simulations and Figures 4.2, 4.3 and 4.4 (see Section 4.6), which have been performed using a Gaussian noise, show that for almost all parameter settings of the algorithm they are not equal to zero. Therefore, the log-linear behavior of the algorithm holds. These figures give also the sign of the convergence (or divergence) rates $\mathrm{F}_\times(\sigma, \sigma_\epsilon)$. Moreover, the sign of these rates (multiplied by the search space dimension d and using some normalizations) is investigated when the search space dimension goes to infinity in the specific case of Gaussian noise (see Section 4.5).
An interesting question that arises now is how this convergence speed given by a possible negative value of $\mathrm{F}_\times(\sigma, \sigma_\epsilon)$ varies as a function of the dimension. In the context

of progress rate theory, this question was addressed (for noisy and non noisy cases) [25] by computing the limit when the dimension goes to infinity of the so-called normalized progress rate. The normalized progress rate corresponds to the expected progress made by an ES algorithm in a single step multiplied by the dimension d of the search space i.e., $d\left[E\left(\frac{\|\mathrm{X}_n\|-\|\mathrm{X}_{n+1}\|}{\|\mathrm{X}_n\|}|\,\mathrm{X}_n\right)\right]$. These computations have been done using the objective function with a Gaussian noise defined in Eq. 4.5, the model **apf** and the scale-invariant rule defined in Eq. 4.4 with $\sigma = \frac{\sigma^*}{d}$. Using these expressions, the normalized progress rate simplifies to $d\left(1 - E\left[\|\mathrm{e}_1 + \frac{\sigma^*}{d}\mathrm{N}_*^{(d)}\|\right]\right)$ where $\mathrm{N}_*^{(d)}$ is given in Definition 4.1.It is worth noticing that the quantity $E\left[\|\mathrm{e}_1 + \sigma\mathrm{N}_*^{(d)}\|\right]$ is the common ratio of the geometric sequence $E(\|\mathrm{X}_n\|)$ where (X_n) is defined by Eq. 4.11 which then converges to zero iff $E\left[\|\mathrm{e}_1 + \sigma\mathrm{N}_*^{(d)}\|\right] < 1$. Therefore, as already pointed in [17] in the non noisy case, the progress rate determines if the algorithm converges or not in expectation. The computed limit of the normalized progress rate shows that the progress rate varies asymptotically linearly as a function of the inverse of the search space dimension.
In the next section, and using normalizations of σ and σ_ϵ as a function of d, we rigorously compute the limit of the normalized convergence rate w.r.t to the dimension d of the quantity $d \times \mathrm{F}_\times(\sigma(d), \sigma_\epsilon(d))$ that we will refer to as the normalized convergence rate.

4.5 Approximation of the convergence rate when the search space dimension goes to infinity

In non noisy cases, it has been theoretically proven in [17] that the convergence rate of ES varies asymptotically linearly as a function of inverse of the dimension of the search space. This result is not specific to ES but holds also for more general cases: it is true for any rank-based algorithm [137], or any Hit-and-Run direct search method [75]. In this section, the goal is to extend the result of asymptotic linear complexity of the convergence rate derived in the non noisy case to the noisy case. Moreover, we show rigorously that the approximation of the model **pf** by the model **apf** that has been done in [8] is reliable for infinite dimension of the search space. This is done by investigating the limit of the normalized convergence rate. For this sake, we adopt the expression of the scale-invariant mutation used in the context of progress rate theory i.e., $\sigma = \frac{\sigma^*}{d}$ and the normalizations introduced in [8, 25] for the noise strength i.e., $\sigma_\epsilon = \frac{\sigma_\epsilon^*}{d}$[8] where $\sigma^* > 0$ and $\sigma_\epsilon^* > 0$ are respectively the normalized step-size mutation and the normalized noise strength. Theorem 4.9 summarizes the result of the limit, when d goes to infinity, of the normalized convergence rate for any noise distribution. In Theorem 4.10, we give the simplified expression of the limit of the normalized convergence rate in the specific case of Gaussian noise. The main difficulty in establishing the proof of Theorem 4.9 is the verification of the technical condition of uniform integrability. This condition was not verified in [25].

[8] In the general case where the ideal objective function equals $\|x\|^\alpha$, $(\alpha > 0)$, Beyer [25] stated that the normalization should be $\sigma_\epsilon = \frac{\alpha\sigma_\epsilon^*}{d}$.

Theorem 4.9. Consider the function F defined in Lemma 4.2. Let σ^* and σ^*_ϵ be two strictly positive constants. For $\sigma(d) = \frac{\sigma^*}{d}$, $\sigma_\epsilon(d) = \frac{\sigma^*_\epsilon}{d}$, the following holds

$$\lim_{d\to\infty} d \times \tilde{\mathrm{F}}(\sigma(d), \sigma_\epsilon(d)) = \lim_{d\to\infty} d \times \mathrm{F}(\sigma(d), \sigma_\epsilon(d)) = \mathcal{A}(\sigma^*, \sigma^*_\epsilon, \lambda) \times \sigma^* + \frac{\sigma^{*2}}{2}$$
$$\text{with } \mathcal{A}(\sigma^*, \sigma^*_\epsilon, \lambda) := \int_{\mathbb{R}} x \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \left(\lambda \int_{\mathbb{R}} \Pr{}^{\lambda-1} \left[\sigma^* x + \sigma^*_\epsilon y \leq \sigma^* \mathrm{N} + \sigma^*_\epsilon \mathcal{N}\right] p_{\mathcal{N}}(y) dy \right), \tag{4.19}$$

where N is the standard normal distribution with mean zero and variance one. Moreover $\mathcal{A}(\sigma^*, \sigma^*_\epsilon, \lambda) \leq 0$ for any $(\sigma^*, \sigma^*_\epsilon, \lambda) \in \mathbb{R}^*_+ \times \mathbb{R}^*_+ \times \mathbb{N}^*$.

Theorem 4.9 states that the convergence rate of the $(1,\lambda)$-ES varies linearly as a function of the inverse of the search space dimension for noisy sphere functions. Besides, this theorem is true for any absolutely continuous noise distribution. Therefore, it applies to the particular case of Gaussian noise and then confirms the reliability of the approximation, when the search space dimension goes to infinity, of the original model **pf** by the model **apf** made in [25, 8]. The quantities $d\ \mathrm{F}(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d})$ and $d\ \tilde{\mathrm{F}}(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d})$ have the same limit: the models **pf** and **apf** are similar when the dimension goes to infinity which confirm the reliability of such an approximation.

Specific case of Gaussian noise: Suppose that the random variable $\mathcal{N}$ modeling the noise follows the standard normal distribution[9]. In this case, the asymptotic expression of the normalized convergence rate is given by the following theorem. Note that for establishing the proof, we used the same techniques that have been used in [25] to derive the limit of the normalized progress rate.

Theorem 4.10. Consider the functions F and $\tilde{\mathrm{F}}$ defined in Lemma 4.2 for the models **pf** and **apf** respectively. Assume that the r.var $\mathcal{N}$ follows the standard normal distribution. For $\lambda \geq 1$, we denote by $c(1,\lambda)$ the expectation of λ independent random variables which follow the same standard normal distribution, then $c(1,\lambda) = \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u\, e^{-\frac{1}{2}u^2} \left[\phi(u)\right]^{\lambda-1} du$ where ϕ is the distribution function of the standard normal distribution. For $\sigma(d) = \frac{\sigma^*}{d}$, $\sigma_\epsilon(d) = \frac{\sigma^*_\epsilon}{d}$ where σ^* and σ^*_ϵ are strictly positive constants, the following holds

$$\lim_{d\to\infty} d\ \mathrm{F}_\times(\sigma(d), \sigma_\epsilon(d)) = -c(1,\lambda)\sigma^* \frac{1}{\sqrt{1 + \left(\frac{\sigma^*_\epsilon}{\sigma^*}\right)^2}} + \frac{\sigma^{*2}}{2}. \tag{4.20}$$

where $\mathrm{F}_\times$ stands for F if model **pf** and $\tilde{\mathrm{F}}$ if model **apf**.

[9] The standard normal distribution is the normal distribution with a mean of zero and a variance of one.

The right hand side of Eq. 4.20 generalizes the limit of the normalized convergence rate computed in [17] for the non-noisy sphere functions and corresponding to $\sigma_\epsilon^* = 0$ in Eq. 4.20. Besides, the limit of the normalized convergence rate is equal to the opposite of the limit of the normalized progress rate computed by Beyer in [25]. This result was expected due to the mathematical approximation of $\ln(x)$ by $x - 1$ when x is close to 1. For the same reason, the limit of the normalized convergence rate, computed in [17] for non-noisy sphere functions, was found to be equal to the opposite of the limit of the normalized progress rate computed in [25]. We can also see from Eq. 4.20 that, for fixed σ^*, the normalized convergence rate is an increasing function of σ_ϵ^* . Besides, $c(1, \lambda)$ is an increasing function of λ as it corresponds to the expectation of the maximum of λ independent distributed random variables with a common law the standard normal distribution. Thus, the normalized convergence rate is a decreasing function of λ.

4.6 Study of the specific case of Gaussian noise

In this section, the noise distribution $\mathcal{N}$ is supposed to be Gaussian. Moreover, σ and σ_ϵ are respectively set equal to $\frac{\sigma^*}{d}$ and $\frac{\sigma_\epsilon^*}{d}$ where σ^* and σ_ϵ^* are strictly positive constant. The object of this section is to study the convergence and divergence cases of the algorithm in the case of finite search space dimension and when the search space dimension goes to infinity.

Convergence and divergence in the limit case of infinite search space dimension : It is easy to see from Theorem 4.10 that, if $\sigma^{*2} + \sigma_\epsilon^{*2} < 4c^2(1, \lambda)$ then $\lim_{d\to\infty} \mathrm{F}_\times \left(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d}\right) < 0$ and then the algorithm converges if the dimension of the search space d is sufficiently large. Otherwise, if $\sigma^{*2} + \sigma_\epsilon^{*2} > 4c^2(1, \lambda)$ then $\mathrm{F}_\times \left(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d}\right) > 0$ and the algorithm diverges when d is sufficiently large. Then if $\sigma_\epsilon^* < 2c(1, \lambda)$ (including in particular the non-noisy case $\sigma_\epsilon^* = 0$), the algorithm converges for some values of σ^* and sufficiently large values of d. But if $\sigma_\epsilon^* > 2c(1, \lambda)$, then the algorithm diverges for any value of σ^* if d is sufficiently large. This means that one has to choose λ sufficiently large such that $\sigma_\epsilon^* < 2c(1, \lambda)$ to ensure that the algorithm converges (provided that d is sufficiently large). However, as the function $\lambda \mapsto c(1, \lambda)$ verifies $c(1, \lambda) \sim \sqrt{2\ln(\lambda)}$ [4] , it increases very slowly. This leads, for sufficiently large values of σ_ϵ^* , to huge values of minimal numbers of offspring needed for convergence as already pointed in [25] and shown in Fig 4.1. As an example, for $\sigma_\epsilon^* = 8$, the minimal number of offspring necessary for satisfying the convergence condition $\sigma_\epsilon^* < 2c(1, \lambda)$ is 18477. Another way to satisfy the convergence condition (for large values of d) is to decrease the noise level σ_ϵ^* by using reevaluation of offspring. Reevaluation means that the objective function of an offspring y will be equal, for the model **pf** for example, to $\frac{1}{n}\sum_{k=1}^{N} \mathrm{f}_k(y)$ with $\mathrm{f}_k(y) = \|y\| + \sigma_\epsilon \|y\| \mathcal{N}_k$ where $\mathcal{N}_k$ are independent realizations of the noise $\mathcal{N}$. The reevaluation using the computation of the objective function of an offspring as the average over N evaluations, leads to

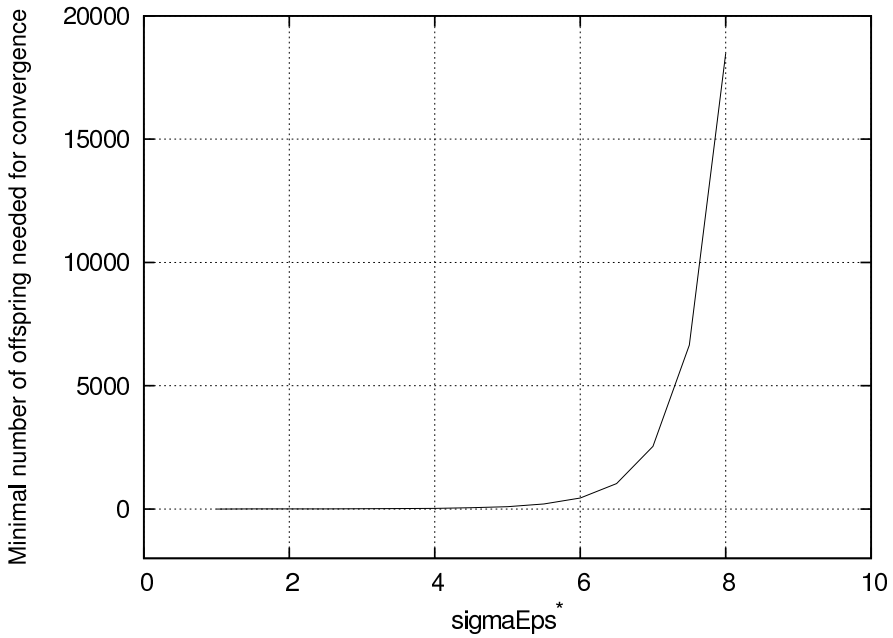


Figure 4.1: Minimal number of offspring λ needed to converge as a function of σ_ϵ^* in the case of infinite dimension.

Table 4.1: Minimal number of evaluations needed per generation for different σ_ϵ^* values and corresponding numbers of evaluations and offspring.

σ_ϵ^*	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8
$N \times \lambda$	2	3	4	6	10	12	15	20	24	30	35	40	48	54	60
N	1	1	1	1	1, 2	2, 3	3	4, 5	4, 6	5, 6	5, 7	8, 10	8, 12	9	10, 12
λ	2	3	4	6	10, 5	6, 4	5	5, 4	6, 4	6, 5	7, 5	5, 4	6, 4	6	6, 5

a decrease of the noise level from σ_ϵ^* to $\sigma_\epsilon^*/\sqrt{N}$. Then a "large" noise level value for which a great number of offspring is needed to converge decrease to a "small" value for which a reasonable number of offspring is sufficient for convergence. This happens at the expense of an additional evaluation cost due to reevaluations of the offspring. We computed, for different values of σ_ϵ^*, the minimal number of evaluations needed (for convergence) per generation and saw the corresponding (optimal) number of evaluations $N \geq 1$ by offspring. Note that the case $N = 1$ means that no reevaluation is used. Results are shown in Table 4.1. This table shows that as the normalized noise strength σ_ϵ^* increases one has to use more and more reevaluations of the offspring. Table 4.1 does not show the gain in the cost of the number of evaluations that can be performed by using reevaluation. The minimal costs of evaluations needed for convergence as a function of the number of reevaluations of an offspring for different normalized noise strengths σ_ϵ^* is shown in Table 4.2. According to Table 4.2, it is better (in term of evaluation cost per generation), for sufficiently large values of σ_ϵ^*, to reevaluate the offspring fitness than to increase the number of offspring λ. This holds only for ES with single parents. For comma ES us-

Table 4.2: Minimal number of evaluations $N \times \lambda$ needed per generation for different σ_ϵ^* values and different number of evaluations N.

σ_ϵ^*	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8
$N \times \lambda = \lambda$	**2**	**3**	**4**	**6**	**10**	16	28	51	98	203	444	1031	2541	6649	18477
$N \times \lambda = 2 \times \lambda$			6	8	**10**	**12**	16	24	34	48	74	114	184	304	516
$N \times \lambda = 3 \times \lambda$				9	12	**12**	**15**	21	27	36	45	63	87	123	177
$N \times \lambda = 4 \times \lambda$					12	16	16	**20**	**24**	32	40	48	64	84	112
$N \times \lambda = 5 \times \lambda$					15	15	20	**20**	25	**30**	**35**	45	55	70	85
$N \times \lambda = 6 \times \lambda$						18	18	24	**24**	**30**	36	42	54	60	78
$N \times \lambda = 7 \times \lambda$							21	28	28	35	**35**	42	49	56	70
$N \times \lambda = 8 \times \lambda$								24	32	32	40	**40**	**48**	56	64
$N \times \lambda = 9 \times \lambda$								27	27	36	36	45	54	**54**	63
$N \times \lambda = 10 \times \lambda$								30	30	40	40	**40**	50	60	**60**
$N \times \lambda = 11 \times \lambda$									33	33	44	44	55	55	66
$N \times \lambda = 12 \times \lambda$									36	36	48	48	**48**	60	**60**
$N \times \lambda = 13 \times \lambda$										39	39	52	52	65	65
$N \times \lambda = 14 \times \lambda$										42	42	56	56	56	70
$N \times \lambda = 15 \times \lambda$										45	45	45	60	60	75
$N \times \lambda = 16 \times \lambda$											48	48	64	64	64
$N \times \lambda = 17 \times \lambda$											51	51	68	68	68
$N \times \lambda = 18 \times \lambda$												54	54	72	72
$N \times \lambda = 19 \times \lambda$												57	54	76	76
$N \times \lambda = 20 \times \lambda$												60	60	60	80
$N \times \lambda = 21 \times \lambda$													63	63	84
$N \times \lambda = 22 \times \lambda$													66	66	88
$N \times \lambda = 23 \times \lambda$													69	69	69
$N \times \lambda = 24 \times \lambda$													72	72	72
$N \times \lambda = 25 \times \lambda$														75	75
$N \times \lambda = 26 \times \lambda$														78	78
$N \times \lambda = 27 \times \lambda$														81	81
$N \times \lambda = 28 \times \lambda$															84
$N \times \lambda = 29 \times \lambda$															87
$N \times \lambda = 30 \times \lambda$															90

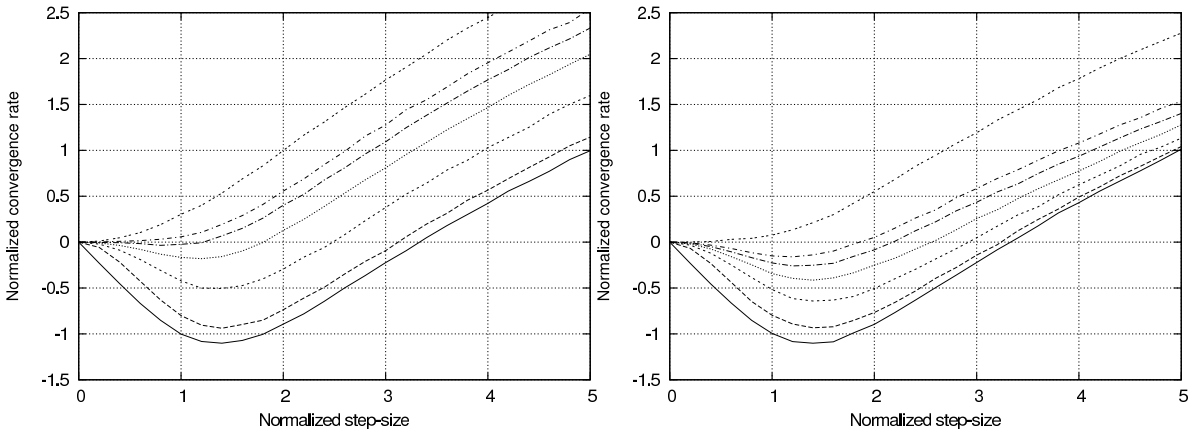


Figure 4.2: $d = 3, \lambda = 5$. Monte Carlo simulations of the normalized convergence rate as a function of the normalized step-size σ^* for the following σ_ϵ^* values : 0, 0.6 , 1.2, 1.8, 2.4, 3.0, 10.0 (from bottom to top). Plots in the left correspond to the normalized convergence rate of the model **pf** (i.e., $d \times \mathrm{F}(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d})$ where F is defined in Eq. 4.14) and plots in the right correspond to the normalized convergence rate of the model **apf** (i.e., $d \times \tilde{\mathrm{F}}(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d})$ where $\tilde{\mathrm{F}}$ is defined in Eq. 4.15).

ing recombination of many parents (the so-called $(\mu/\mu, \lambda)$-ES), the progress rate formula derived in [25] suggests that it is preferable to increase the number of offspring than to reevaluate them.

Convergence and divergence for finite dimensions For $d < +\infty$, if the normalized convergence rate $d\ \tilde{\mathrm{F}}\left(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d}\right)$ (or $d\ \mathrm{F}\left(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d}\right)$) is strictly negative, the algorithm converges. If it is strictly positive, the algorithm diverges. We plot, using Monte Carlo simulations, the expectations $d\ \tilde{\mathrm{F}}\left(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d}\right)$ and $d\ \mathrm{F}\left(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d}\right)$ as a function of σ^* for different values of σ_ϵ^*. Figures 4.2, 4.3 and 4.4 represent these plots for the models **pf** and **apf** respectively for dimensions 3, 10 and 30.

Finite and infinite normalized convergence rates Using the explicit expression of the limit of the normalized convergence rate given in Eq. 4.20 for Gaussian noise, we plotted, for $\sigma_\epsilon^* = 1.2$ (Fig 4.5) and $\sigma_\epsilon^* = 3$ (Fig 4.6), the limit of the normalized convergence rate when the dimension d goes to infinity with normalized convergence rates for dimensions 3, 10 and 30 and models **pf** and **apf** as a function of the normalized step-size mutation σ^*.

These plots use $\lambda = 5$ and confirm results in Theorem 4.10. In fact, the curves are getting closer to the limit expression of the convergence rate given in Eq. 4.20 as the dimension increases. This holds for the two models **pf** and **apf**. Moreover, these curves reveals that the limit expression of the normalized convergence rate is an upped bound for normalized convergence rates of finite dimensions. This shows that the study of the limit of the convergence rate is safe as whenever this limit is strictly negative (and the "limit"

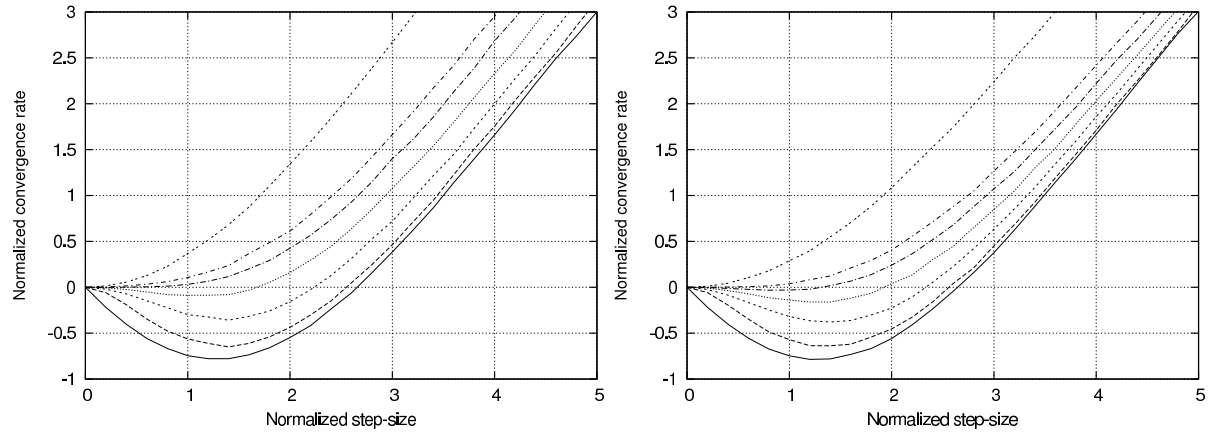


Figure 4.3: $d = 10, \lambda = 5$. Monte Carlo simulations of the normalized convergence rate as a function of the normalized step-size σ^* for the following σ_ϵ^* values : 0, 0.6 , 1.2, 1.8, 2.4, 3.0, 10.0 (from bottom to top). Plots in the left correspond to the normalized convergence rate of the model **pf** (i.e., $d \times \mathrm{F}(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d})$ where F is defined in Eq. 4.14) and plots in the right correspond to the normalized convergence rate of the model **apf** (i.e., $d \times \tilde{\mathrm{F}}(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d})$ where $\tilde{\mathrm{F}}$ is defined in Eq. 4.15).

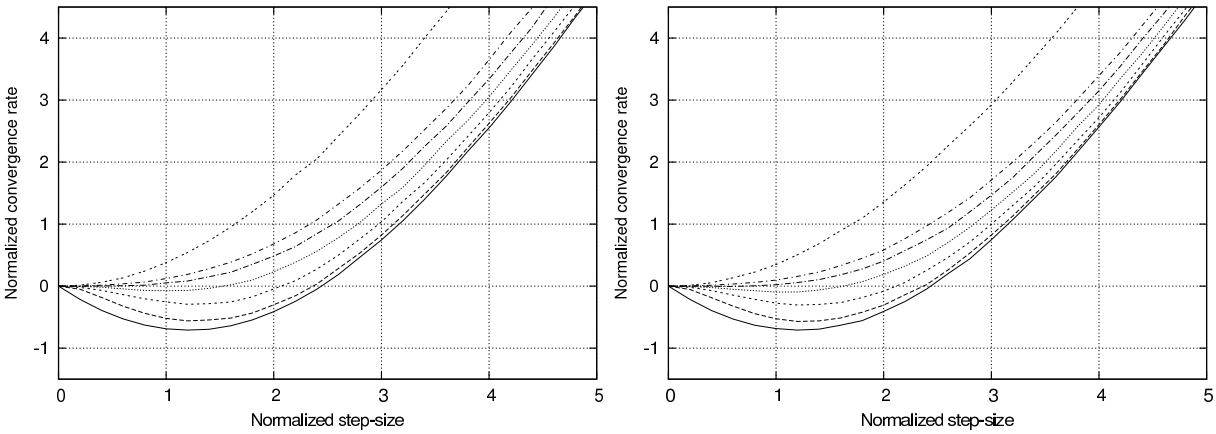


Figure 4.4: $d = 30, \lambda = 5$. Monte Carlo simulations of the normalized convergence rate as a function of the normalized step-size σ^* for the following σ_ϵ^* values : 0, 0.6 , 1.2, 1.8, 2.4, 3.0, 10.0 (from bottom to top). Plots in the left correspond to the normalized convergence rate of the model **pf** (i.e., $d \times \mathrm{F}(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d})$ where F is defined in Eq. 4.14) and plots in the right correspond to the normalized convergence rate of the model **apf** (i.e., $d \times \tilde{\mathrm{F}}(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d})$ where $\tilde{\mathrm{F}}$ is defined in Eq. 4.15).

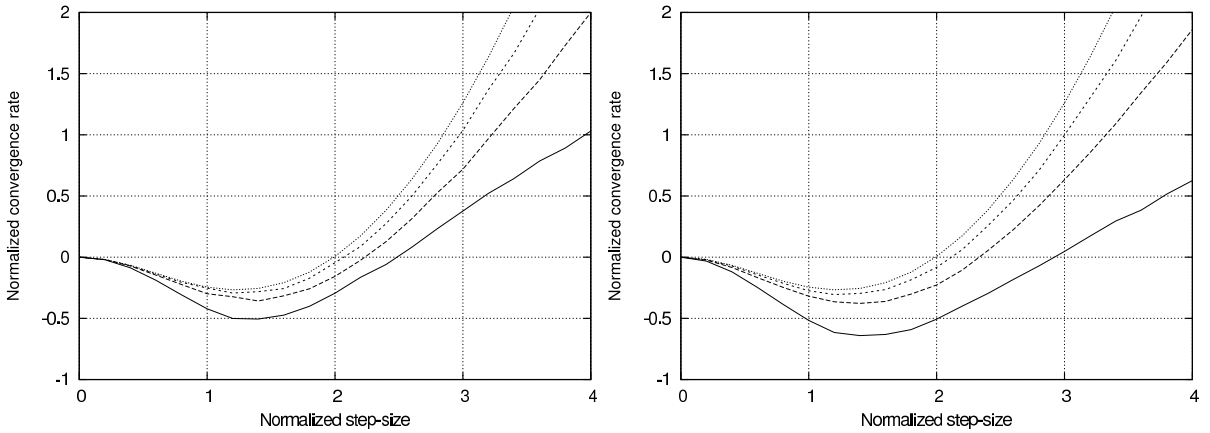


Figure 4.5: Normalized convergence rates for dimensions 3, 10 and 30 and the limit expression of the convergence rate ($d = +\infty$) as a function of σ^* for $\sigma^*_\epsilon = 1.2$, $\lambda = 5$ and models **pf** (left) and **apf** (right). From bottom to top, the curves correspond to dimensions 3, 10, 30 and the limit $d = +\infty$.

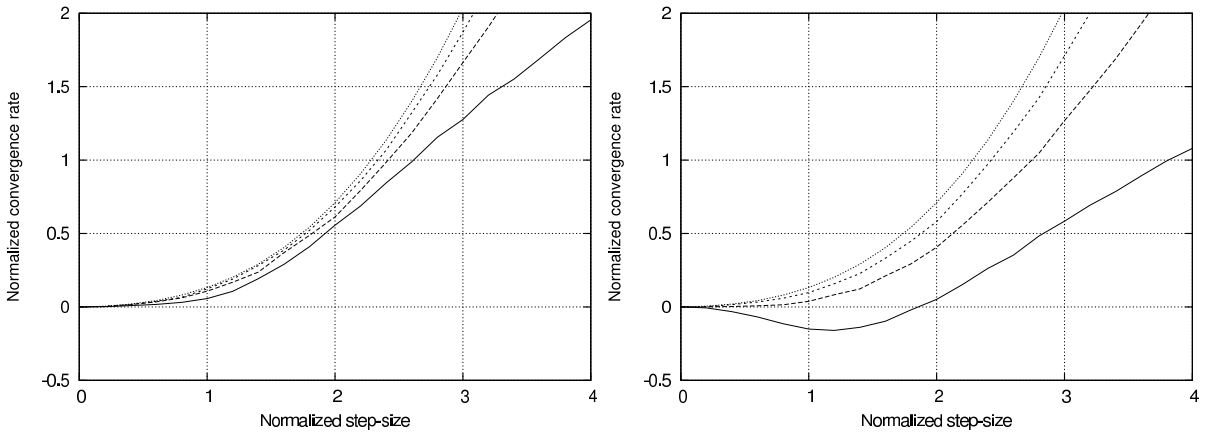


Figure 4.6: Normalized convergence rates for dimensions 3, 10 and 30 and the limit expression of the convergence rate ($d = +\infty$) as a function of σ^* for $\sigma^*_\epsilon = 3$, $\lambda = 5$ and models **pf** (left) and **apf** (right). From bottom to top, the curves correspond to dimensions 3, 10, 30 and the limit $d = +\infty$

algorithm converges), not only the convergence holds for sufficiently large dimensions but for all dimensions. In the case of $\sigma_\epsilon^* = 3$, Eq. 4.20 implies that the algorithm diverges for sufficiently high values of d as $\sigma_\epsilon^* - 2 * c(1,5)$ is strictly positive. However, for small dimensions, the algorithm can converge for some σ^* values as shown in Fig 4.6 (right) in the lower curve corresponding to $d = 3$. This represents a limit for the usefulness of infinite dimension results as infinite dimension study predicts divergence and the plot corrseponding to dimension 3 in Fig 4.6 (right) shows that the algorithm converges for the same settings of the algorithm and of the normalized noise strength. Another fact revealed by the comparison of finite dimension curves corresponding to the model **pf** (left) to those corresponding to the model **apf** (right) is that, for the same parameters values (i.e., σ^*, σ_ϵ^*, λ and d), the signs of the convergence rates are sometimes different. This means that, while a convergence is predicted for one of the two models, a divergence occurs for the other model. This is a limitation of the use, when the dimension is finite, of the approximation of the model **pf** by the model **apf**.

Optimal convergence rates, optimal step-sizes and limit values for convergence for different noise levels We plotted, using $\lambda = 5$ and the model **pf**, as a function of the normalized noise strength σ_ϵ^* the following quantities:

- optimal normalized convergence rates (Fig 4.7)
- optimal normalized step-size mutations (Fig 4.8 (Left))
- upper values of the normalized step-size mutation for which the algorithm converges (Fig 4.8 (Right))

The plots show that, for a given σ_ϵ^* these values decrease as the dimension increases and have as limit the values corresponding to $d = +\infty$. It is woth noticing that in Figures 4.7 and 4.8, the curves relative to infinite dimension can be found in [114, Fig. 14-2 and 14-3].

4.7 Discussion and conclusion

In this chapter we have analyzed the convergence of the scale-invariant $(1, \lambda)$-ES for the noisy sphere function. Two models for the noise have been analyzed: the model **pf**, where the noise is scaled proportionally to the location of the individual or to the non-noisy part of the objective function and the model **apf**, introduced as an approximation of the model **pf** in [25, 8], where the noise is scaled proportionally to the norm of the parent and therefore to the step-size.

We prove rigorously that comma ES are more robust than plus ES in presence of noise: In Chapter 3, it is shown that the algorithm cannot converge (at least in expectation), if the noise is Gaussian. However, we have shown in this chapter that convergence holds almost surely (also in expectation) for Gaussian noise but with small standard deviation (or noise strength). Moreover there is a robustness in the technique used for the proof: the convergence in presence of noise is obtained using the same tools used for the analysis of convergence of ES on non-noisy functions.

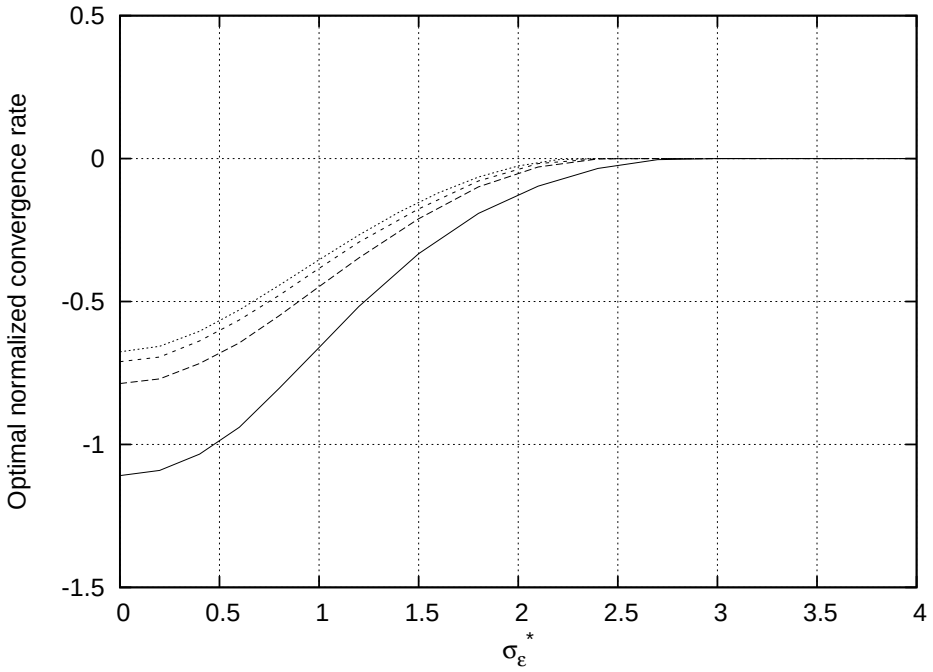


Figure 4.7: Optimal normalized convergence rate as a function of the normalized noise strength σ_ϵ^* for dimensions 3,10,30 and the limit of infinite dimension (from bottom to top).

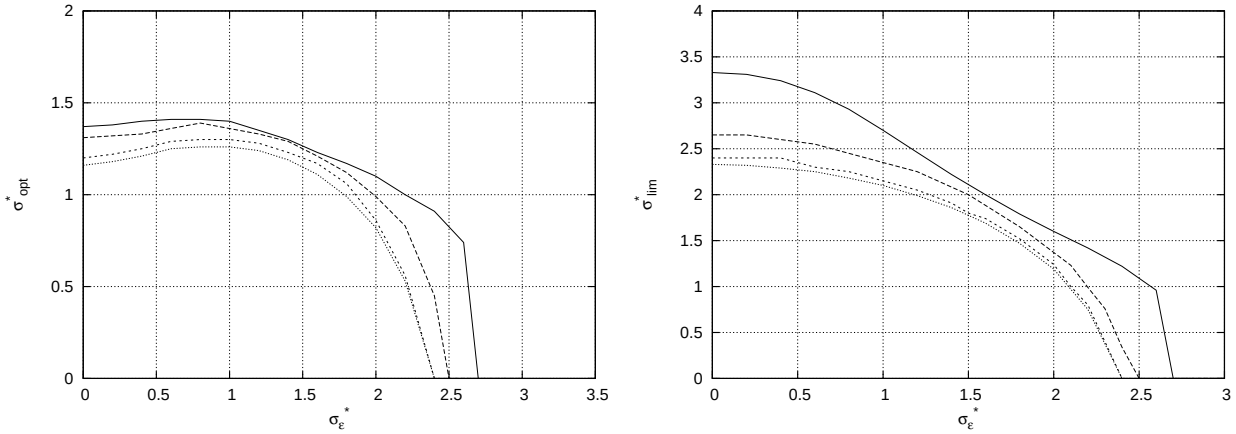


Figure 4.8: **Left:** Optimal normalized step size σ^* as a function of the normalized noise strength σ_ϵ^* for dimensions 3,10,30 and the limit of infinite dimension (form top to bottom (considering the values corresponding to $\sigma_\epsilon^* = 0$)). **Right:** Normalized step size σ^* for which the convergence rate equals 0 as a function of the normalized noise strength σ_ϵ^* for dimensions 3,10,30 and the limit of infinite dimension (from top to bottom).

The convergence rate obtained for finite dimension is expressed as the expectation of a random variable. Though it is difficult to have a theoretical estimation of this convergence rate without making an assumption (that the dimension is large for instance), our study shows that it is fairly easy to simulate the convergence rate with a Monte-Carlo method.

We derive rigorously the limit of the normalized convergence rate when the dimension d goes to infinity and meet the results obtained with the progress rate approach [25], bridging therefore the gap between finite approximation results and infinite approximations results. As already observed in [25], the computed expression is a generalization of the normalized progress rate (or normalized convergence rate in our case) in the case of non-noisy comma ES but this computation allowed us to prove: 1) the similarity of the two models for infinite dimensions; 2) that the convergence rate of the algorithm changes asymptotically linearly with the inverse of the search space dimension. In the particular case of Gaussian noise, the limit of the normalized convergence rate has been explicitly derived (the same expression has been previously derived in [25] for the progress rate) and we investigate the use of re-sampling versus increasing the number of offspring to make the algorithm converge when noise levels are large. Moreover, the specific study of the Gaussian noise case: 1) show the usefulness of infinite dimension studies where normalized convergence rate can be quantified explicitly, to learn about the behavior of the algorithm for finite dimensions studies; 2) the limits of adopting, for finite dimensions, infinite dimension results and for approximating the model **pf** by the model **apf**.

Appendix

Proof of Lemma 4.2

Let $(\mathrm{N}_i^{(d)})_{i\in[1,\lambda]}$ (resp. $(\mathcal{N}_i)_{i\in[1,\lambda]}$) be λ independent samplings of $\mathrm{N}^{(d)}$ (resp. $\mathcal{N}$). The random vector $(\mathrm{N}_*^{(d)},\mathcal{N}_*)$ verifies, according to Definition 4.1,

$$\|\mathrm{e}_1+\sigma\mathrm{N}_*^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_*)=\min_{1\le i\le\lambda}\{\|\mathrm{e}_1+\sigma\mathrm{N}_i^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_i)\} \tag{4.21}$$

for the model **pf**, and

$$\|\mathrm{e}_1+\sigma\mathrm{N}_*^{(d)}\|+\sigma_\epsilon\mathcal{N}_*=\min_{1\le i\le\lambda}\{\|\mathrm{e}_1+\sigma\mathrm{N}_i^{(d)}\|+\sigma_\epsilon\mathcal{N}_i\} \tag{4.22}$$

for the model **apf**. First, we give interest to the probability density function of the random vecotr $\mathrm{N}_*^{(d)}$ in the specific case of the model **pf**. The same reasoning holds for the model **apf**. Let $A\in\mathfrak{B}(\mathbb{R}^d)$[10]. According to Eq. 4.21, we have:

$$P(\mathrm{N}_*^{(d)}\in A)=\cup_{i=1}^{\lambda}P(\mathrm{N}_i^{(d)}\in A;\cap_{\{1\le j\le\lambda;j\neq i\}}\|\mathrm{e}_1+\sigma\mathrm{N}_i^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_i)\le\|\mathrm{e}_1+\sigma\mathrm{N}_j^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_j))$$

The random variables $(\mathrm{N}_i^{(d)})_{i\in[1,\lambda]}$ and $(\mathcal{N}_i)_{i\in[1,\lambda]}$ play the same role. Therefore, we have,

$$P(\mathrm{N}_*^{(d)}\in A)=\lambda P(\mathrm{N}_1^{(d)}\in A;\cap_{\{2\le j\le\lambda\}}\|\mathrm{e}_1+\sigma\mathrm{N}_1^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_1)\le\|\mathrm{e}_1+\sigma\mathrm{N}_j^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_j))$$

This can be rewritten as

$$P(\mathrm{N}_*^{(d)}\in A)=\frac{\lambda}{(2\pi)^{\frac{d}{2}}}\int_A e^{-\frac{\|x\|^2}{2}}P(\cap_{2\le j\le\lambda}\|\mathrm{e}_1+\sigma x\|\,(1+\sigma_\epsilon\mathcal{N}_1)\le\|\mathrm{e}_1+\sigma\mathrm{N}_j^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_j))dx$$

This gives

$$\begin{aligned}P(\mathrm{N}_*^{(d)}\in A)&=\frac{\lambda}{(2\pi)^{\frac{d}{2}}}\times\\&\int_A\int_{\mathbb{R}}e^{-\frac{\|x\|^2}{2}}f_{\mathcal{N}}(y)\left(P(\cap_{2\le j\le\lambda}\|\mathrm{e}_1+\sigma x\|\,(1+\sigma_\epsilon y)\le\|\mathrm{e}_1+\sigma\mathrm{N}_j^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_j))\right)dxdy\end{aligned} \tag{4.23}$$

The random vectors $(\mathrm{N}_i^{(d)},\mathcal{N}_i)_{i\in[2,\lambda]}$ are independent identically distributed. Therefore, for fixed $(x,y)\in\mathbb{R}^d\times\mathbb{R}$, we have

$$\begin{aligned}&P(\cap_{2\le j\le\lambda}\|\mathrm{e}_1+\sigma x\|\,(1+\sigma_\epsilon y)\le\|\mathrm{e}_1+\sigma\mathrm{N}_j^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_j)\\&=\cap_{2\le j\le\lambda}P(\|\mathrm{e}_1+\sigma x\|\,(1+\sigma_\epsilon y)\le\|\mathrm{e}_1+\sigma\mathrm{N}_j^{(d)}\|\,(1+\sigma_\epsilon\mathcal{N}_j))\\&=P^{\lambda-1}(\|\mathrm{e}_1+\sigma x\|\,(1+\sigma_\epsilon y)\le\|\mathrm{e}_1+\sigma\mathrm{N}^d\|\,(1+\sigma_\epsilon\mathcal{N}))\,.\end{aligned}$$

[10] $\mathfrak{B}(\mathbb{R}^d)$ is the Borel σ-algebra on $\mathbb{R}^d$.

Combining the last equation with Eq. 4.23, one gets

$$P(\mathrm{N}_*^{(d)} \in A) = \frac{\lambda}{(2\pi)^{\frac{d}{2}}} \times$$
$$\int_A \int_{\mathbb{R}} e^{-\frac{\|x\|^2}{2}} f_{\mathcal{N}}(y) \left(P^{\lambda-1}(\|\mathrm{e}_1 + \sigma x\| (1 + \sigma_\epsilon y) \leq \|\mathrm{e}_1 + \sigma \mathrm{N}^d\| (1 + \sigma_\epsilon \mathcal{N})) \right) dxdy \quad (4.24)$$

This gives

$$P(\mathrm{N}_*^{(d)} \in A) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_A e^{-\frac{\|x\|^2}{2}} \mathrm{H}(\lambda, x, \sigma, \sigma_\epsilon) dx \,, \quad (4.25)$$

where H is given in Definition 4.1. This ends the proof for the probability density function of the random vector $\mathrm{N}_*^{(d)}$.
Now, we define the quantities $\mathrm{F}_\times^-(\sigma, \sigma_\epsilon) := E\left[\ln^-(\|\mathrm{e}_1 + \sigma \mathrm{N}_*^{(d)}\|)\right]$ and $\mathrm{F}_\times^+(\sigma, \sigma_\epsilon) := E\left[\ln^+(\|\mathrm{e}_1 + \sigma \mathrm{N}_*^{(d)}\|)\right]$ where $\mathrm{F}_\times^-(\sigma, \sigma_\epsilon)$ (resp. $\mathrm{F}_\times^+(\sigma, \sigma_\epsilon)$) stands for $\mathrm{F}^-(\sigma, \sigma_\epsilon)$ (resp. $\mathrm{F}^+(\sigma, \sigma_\epsilon)$) with $\mathrm{N}_*^{(d)}$ given by Eq. 4.21 if model **pf**, and for $\tilde{\mathrm{F}}^-(\sigma, \sigma_\epsilon)$ (resp. $\tilde{\mathrm{F}}^+(\sigma, \sigma_\epsilon)$) with $\mathrm{N}_*^{(d)}$ given by Eq. 4.22 if model **apf**. Note that we have used the notation "$\mathrm{F}_\times^-$" to refer to the quantity F^- for the model **pf** and to the quantity $\tilde{\mathrm{F}}^-$ for the model **apf**. In the sequel, we will use the same convention, i.e., the notation $A_\times$ will refer to a quantity A relative to the model **pf** and to a quantity $\tilde{A}$ relative to the model **apf**.

The quantities $\mathrm{F}_\times^-$ and $\mathrm{F}_\times^+$ exist but could be infinite. Let $g_\times^+, g_\times^- : \mathbb{N}^* \times \mathbb{R}^d \times [0, +\infty[\times[0, +\infty[$ be defined for $(\lambda, x, \sigma, \sigma_\epsilon)$ in $\mathbb{N}^* \times \mathbb{R}^d \times [0, +\infty[\times[0, +\infty[$ by

$$g_\times^+(\lambda, x, \sigma, \sigma_\epsilon) = \frac{1}{(2\pi)^{d/2}} \ln^+(\|e_1 + \sigma x\|^2) e^{-\frac{\|x\|^2}{2}} H_\times(\lambda, x, \sigma, \sigma_\epsilon)$$

and

$$g_\times^-(\lambda, x, \sigma, \sigma_\epsilon) = \frac{1}{(2\pi)^{d/2}} \ln^-(\|e_1 + \sigma x\|^2) e^{-\frac{\|x\|^2}{2}} H_\times(\lambda, x, \sigma, \sigma_\epsilon).$$

We notice that for $d \geq 2$, $g_\times^+(\lambda, (x_1, x_2, \ldots, x_d), \sigma, \sigma_\epsilon) = g_\times^+(\lambda, (x_1, \epsilon_2 x_2, \ldots, \epsilon_d x_d), \sigma, \sigma_\epsilon)$ (which is also true for $g_\times^-$) for all $(\epsilon_2, \ldots, \epsilon_d)$ in $\{-1, +1\}^{d-1}$ and $(x_1, x_2, \ldots, x_d)$ in $\mathbb{R}^d$ then we can restrict the integration giving $F_\times(\sigma, \sigma_\epsilon)$ to the domain $\mathcal{D} := \mathbb{R}^* \times]0, +\infty[^{d-1}$, more precisely one has (for $d \geq 2$)

$$\mathrm{F}_\times^-(\sigma, \sigma_\epsilon) = 2^{d-2} \int_{\mathcal{D}} g_\times^-(\lambda, x, \sigma, \sigma_\epsilon) dx$$

and

$$\mathrm{F}_\times^+(\sigma, \sigma_\epsilon) = 2^{d-2} \int_{\mathcal{D}} g_\times^+(\lambda, x, \sigma, \sigma_\epsilon) dx \,.$$

Changing to spherical coordinates (with $d \geq 2$) we obtain after partial integration

$$\mathrm{F}_\times^-(\sigma, \sigma_\epsilon) = \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{1}{W_{d-2} \Gamma\left(\frac{d}{2}\right)}$$
$$\int_0^{+\infty} \int_0^{\frac{\pi}{2}} \ln^-\left(|\sigma r - e^{i\theta}|\right) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta) K_\times(\lambda, r, \theta, \sigma, \sigma_\epsilon) \, dr \, d\theta \,,$$

and

$$\mathrm{F}_{\times}^{+}(\sigma,\sigma_\epsilon) = \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{1}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty}\int_0^{\pi} \ln^{+}\left(|\sigma r - e^{i\theta}|\right) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta) K_{\times}(\lambda, r, \theta, \sigma, \sigma_\epsilon)\, dr\, d\theta\,,$$

where for $n \in \mathbb{N}$, $W_n = \int_0^{\pi/2} \sin^n\theta \; d\theta$ is the classical Wallis integral and for $z \in \mathbf{C}$ such that $Re(z) > 0, \Gamma(z) = \int_0^{+\infty} e^{-u} u^{z-1} du$ is the Gamma function and $K_{\times}$ is the function defined on $\mathbb{N}^* \times [0,+\infty[\times[0,\pi] \times [0,+\infty[\times[0,+\infty[$ by

$$K(\lambda, r, \theta, \sigma, \sigma_\epsilon) = \lambda \int_{\mathbb{R}} \Pr^{\lambda-1}\left[|\sigma r - e^{i\theta}|\,(1+\sigma_\epsilon y) \leq \|\mathrm{e}_1 + \sigma \mathrm{N}^{(d)}\|\,(1+\sigma_\epsilon \mathcal{N})\right] p_{\mathcal{N}}(y) dy,$$

for the model **pf** and

$$\tilde{K}(\lambda, r, \theta, \sigma, \sigma_\epsilon) = \lambda \int_{\mathbb{R}} \Pr^{\lambda-1}\left[|\sigma r - e^{i\theta}| + \sigma_\epsilon y \leq \|\mathrm{e}_1 + \sigma \mathrm{N}^{(d)}\| + \sigma_\epsilon \mathcal{N}\right] p_{\mathcal{N}}(y) dy,$$

for the model **apf**.

The integrand $h_{\times}^{-} : (r,\theta,\sigma,\sigma_\epsilon) \mapsto \ln^{-}(|\sigma r - e^{i\theta}|) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta) K_{\times}(\lambda, r, \theta, \sigma, \sigma_\epsilon)$ defined on the set $]0,+\infty[\times[0,\pi/2] \times [0,+\infty[\times[0,+\infty[$ is continuous for almost all $(r,\theta,\sigma,\sigma_\epsilon)$ in $]0,+\infty[\times[0,\pi/2]\times[0,+\infty[\times[0,+\infty[$. In particular, for almost all (r,θ) in $]0,+\infty[\times[0,\pi/2]$, the map $(\sigma,\sigma_\epsilon) \mapsto h_{\times}^{-}(r,\theta,\sigma,\sigma_\epsilon)$ is continuous. Moreover, the function $K_{\times}$ is dominated by λ and $|\sigma r - e^{i\theta}| \geq \sin\theta$ for all (r,θ) in $]0,+\infty[\times[0,\pi/2]$. Then $h_{\times}^{-}$ is dominated by $h_1 : (r,\theta) \mapsto \ln^{-}(\sin\theta) r^{d-1} e^{-r^2/2}$ i.e., $h_{\times}^{-}(r,\theta,\sigma,\sigma_\epsilon) \quad h_1(r,\theta)$ for all $(r,\theta,\sigma,\sigma_\epsilon)$ in $]0,+\infty[\times[0,\pi/2] \times [0,+\infty[\times[0,+\infty[$. Since h_1 is integrable, the mapping $\mathrm{F}_{\times}^{-}$ is finite and continuous w.r.t. the variables σ and σ_ϵ on $[0,+\infty[\times[0,+\infty[$ thanks to the Lebesgue dominated convergence theorem. Besides, we have

$$\mathrm{F}_{\times}^{+}(\sigma,\sigma_\epsilon) \leq \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{1}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty}\int_0^{\pi} \sigma r^d e^{-\frac{r^2}{2}}\, dr\, d\theta \; < +\infty.$$

Then $\mathrm{F}_{\times}^{+}$ and $\mathrm{F}_{\times}^{-}$ are finite meaning that the map $\mathrm{F}_{\times}$ is well defined. Now we have to look at the continuity of $\mathrm{F}_{\times}^{+}$. The integrand

$$h_{\times}^{+} : (r,\theta,\sigma,\sigma_\epsilon) \mapsto \ln^{+}(|\sigma r - e^{i\theta}|) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta) K_{\times}(\lambda, r, \theta, \sigma, \sigma_\epsilon)$$

defined on the set $]0,+\infty[\times[0,\pi] \times [0,+\infty[\times[0,+\infty[$ verifies that for almost all (r,θ) in $]0,+\infty[\times[0,\pi]$, the map $(\sigma,\sigma_\epsilon) \mapsto h_{\times}^{+}(r,\theta,\sigma,\sigma_\epsilon)$ is continuous on every set $[0,S]\times[0,+\infty[$ with $0 < S < +\infty$. Moreover, $h_{\times}^{+}$ is dominated by $h_2 : r \mapsto S r^d e^{-r^2/2}$ for $(r,\theta,\sigma,\sigma_\epsilon)$ in $]0,+\infty[\times[0,\pi] \times [0,S] \times [0,+\infty[$. Since h_2 is integrable, the continuity of $\mathrm{F}_{\times}^{+}$ w.r.t. the variables σ and σ_ϵ on $[0,S]\times[0,+\infty[$ follows from the Lebesgue dominated convergence theorem. This is true for any $[0,S]\times[0,+\infty[$ with $0 < S < +\infty$ then the continuity of $\mathrm{F}_{\times}^{+}$ holds also on $[0,+\infty[\times[0,+\infty[$. For the remaining case $d = 1$, the integrand in $\mathrm{F}_{\times}^{+}(\sigma,\sigma_\epsilon)$ will be dominated by $S x e^{-\frac{x^2}{2}}$ for $(x,\sigma,\sigma_\epsilon) \in \mathbb{R}\times[0,S]\times[0,+\infty[$ which gives the continuity of $\mathrm{F}_{\times}^{+}(\sigma,\sigma_\epsilon)$ on $[0,+\infty[\times[0,+\infty[$. For $\mathrm{F}_{\times}^{-}$, after a change of variables $y = \sigma x$, the integrand in $\mathrm{F}_{\times}^{-}(\sigma,\sigma_\epsilon)$ will be dominated by $\frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \frac{\ln(|1+y|)}{y}$ for $(y,\sigma,\sigma_\epsilon) \in]-2,0]\times[0,+\infty[\times[0,+\infty[$. □

Proof of Proposition 4.4

At each iteration n, Eq. 4.11 gives

$$\|\mathrm{X}_{n+1}\| = \|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_{*,n}^{(d)}\|\,,$$

where $(\mathrm{N}_{*,n}^{(d)})_n$ is defined in Eq. 4.9 or in Eq. 4.10 according to the model considered. In the beginning, we show inductively that, for all $n \geq 0$, $\|\mathrm{X}_n\| > 0$ almost surely:
1) By definition $P(\|X_0\| > 0) = 1$. 2) Suppose that $P(\|\mathrm{X}_n\| > 0) = 1$ for $n \geq 0$; then, by Eq. 4.11, the i^{th} offspring has a strictly positive non-noisy objective function (i.e., $P(\|\mathrm{Y}_{i,n}\| > 0) = 1$ for all i in $[1, \lambda]$) as the multivariate normal distribution is absolutely continuous w.r.t. to the Lebesgue measure and in particular $P(\|\mathrm{X}_{n+1}\| > 0) = 1$. This gives that for all $n \geq 0$, $\|\mathrm{X}_n\| > 0$ almost surely and we can write

$$\|\mathrm{X}_{n+1}\| = \|\mathrm{X}_n\|\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_{*,n}^{(d)}\right\| \; a.s.$$

Taking the logarithm of the previous equation, we get

$$\ln\left(\|\mathrm{X}_{n+1}\|\right) = \ln\left(\|\mathrm{X}_n\|\right) + \ln\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_{*,n}^{(d)}\right\|\right) \; a.s.$$

and after summing such equalities we obtain

$$\ln\left(\|X_n\|\right) - \ln\left(\|X_0\|\right) = \sum_{k=0}^{n-1} \ln\left(\left\|\frac{X_k}{\|X_k\|} + \sigma\mathrm{N}_{*,k}^{(d)}\right\|\right) \; a.s.$$

□

Proof of Proposition 4.7

We will detail the proof for the model **apf**. Thus in the remainder of this proof the random vectors $\mathrm{N}_{*,n}^{(d)}$ and $\mathrm{N}_{*}^{(d)}$ are relative to the **apf** model (i.e., respectively defined in Eq. 4.10 and Definition 4.1). The same reasoning holds for the model **pf**. For X_n fixed, let $\tilde{L}_n : \mathbb{N}^* \times \mathbb{R}^d \times [0, +\infty[\times[0, +\infty[\mapsto \mathbb{R}^+$ be the function defined by

$$\tilde{L}_n(\lambda, x, \sigma, \sigma_\epsilon) = \lambda \int_{\mathbb{R}} \Pr{}^{\lambda-1}\left[\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma x\right\| + \sigma_\epsilon y \leq \left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}^{(d)}\right\| + \sigma_\epsilon\mathcal{N}\right] p_{\mathcal{N}}(y)dy\,, \tag{4.26}$$

for $(\lambda, x, \sigma, \sigma_\epsilon) \in \mathbb{N}^* \times \mathbb{R}^d \times [0, +\infty[\times[0, +\infty[$. Similarly to the proof of Lemma 4.2, we have

$$P(\mathrm{N}_{*,n}^{(d)} \in A|\mathrm{X}_n) = \int_A \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{\|x\|^2}{2}} \tilde{L}_n(\lambda, x, \sigma, \sigma_\epsilon)dx\,. \tag{4.27}$$

Therefore, the probability density function of the random vector $\mathrm{N}_{*,n}^{(d)}$ conditionnally to X_n is obtained by multiplying the probability density function of $\mathrm{N}^{(d)}$ by the function $\tilde{L}_n$

given in Eq. 4.26.
The isotropy of the standard d-dimensional normal distribution gives

$$\tilde{L}_n(\lambda, x, \sigma, \sigma_\epsilon) = \lambda \int_{\mathbb{R}} \Pr^{\lambda-1}\left[\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma x\right\| + \sigma_\epsilon y \leq \left\|\mathrm{e}_1 + \sigma \mathrm{N}^{(d)}\right\| + \sigma_\epsilon \mathcal{N}\right] p_{\mathcal{N}}(y)dy\,.$$

Let us compute $E\Big(\ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}^{(d)}_{*,n}\right\|\right)\Big)$ and $E\Big(\ln^+\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}^{(d)}_{*,n}\right\|\right)\Big)$. We have

$$E\Big(\ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}^{(d)}_{*,n}\right\|\right)|\mathrm{X}_n\Big) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma x\right\|\right) e^{-\frac{\|x\|^2}{2}} \tilde{L}_n(\lambda, x, \sigma, \sigma_\epsilon)dx\,.$$

Using again the isotropy of the standard d-dimensional normal distribution, one gets

$$E\Big(\ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}^{(d)}_{*,n}\right\|\right)|\mathrm{X}_n\Big) = E\left[\ln^-(\|\mathrm{e}_1 + \sigma \mathrm{N}^{(d)}_*\|)\right] < +\infty\,. \tag{4.28}$$

Similarly, we have

$$E\Big(\ln^+\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}^{(d)}_{*,n}\right\|\right)|\mathrm{X}_n\Big) = E\left[\ln^+(\|\mathrm{e}_1 + \sigma \mathrm{N}^{(d)}_*\|)\right] < +\infty\,. \tag{4.29}$$

Hence $E\left[\ln\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}^{(d)}_{*,n}\right\|\right)\right] = \tilde{\mathrm{F}}(\sigma, \sigma_\epsilon) < +\infty$, and so $E(Y_n) = 0$.
Let $F_2 : [0,\infty[\times[0,+\infty[\to [0,+\infty]$ be defined, for $(t_1, t_2) \in [0,+\infty[\times[0,+\infty[$, by

$$\tilde{G}(t_1, t_2) := \frac{\lambda}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left[\ln(\|\mathrm{e}_1 + t_1 x\|)\right]^2 e^{-\frac{\|x\|^2}{2}} \tilde{\mathrm{H}}(\lambda, x, t_1, t_2)dx\,,$$

where $\tilde{\mathrm{H}}$ is the function defined in Definition 4.1. Similarly to the proof of Lemma 4.2, we prove that $\tilde{G}$ has finite values. Now, from the definitions of F and F_2 one has

$$E(|Y_n|^2) = \tilde{G}(\sigma, \sigma_\epsilon) - (\tilde{\mathrm{F}}(\sigma, \sigma_\epsilon))^2 < +\infty\,. \tag{4.30}$$

This ends the proof of the first point. The random vectors Y_n and Z_n have the same distribution if their characteristic functions are identical. But successively

$$\begin{aligned}
E(e^{itY_n} \,|\, \mathrm{X}_n) &= e^{-it\tilde{\mathrm{F}}(\sigma,\sigma_\epsilon)} E\Big(e^{it \ln\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}^{(d)}_{*,n}\right\|\right)} \,|\, \mathrm{X}_n\Big) \\
&= \frac{e^{-it\tilde{\mathrm{F}}(\sigma,\sigma_\epsilon)}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{it \ln\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma x\right\|\right)} e^{-\|x\|^2/2} \tilde{L}_n(\lambda, x, \sigma, \sigma_\epsilon)dx \\
&= \frac{e^{-it\tilde{\mathrm{F}}(\sigma,\sigma_\epsilon)}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{it \ln(\|e_1 + \sigma x\|)} e^{-\|x\|^2/2} \tilde{\mathrm{H}}(\lambda, x, \sigma, \sigma_\epsilon)dx \\
&= E(e^{itZ_n})\,.
\end{aligned}$$

Therefore $E(e^{itY_n}) = E(E(e^{itY_n} \,|\, \mathrm{X}_n)) = E(e^{itZ_n})$. To finish the proof we show the orthogonality property of the sequence (Y_n). Let n and m be indices such that $n < m$. The random vector Y_n is $\sigma(\mathrm{X}_n, \mathrm{N}^{(d)}_{*,n})$-measurable, so that

$$E(Y_m Y_n \,|\, \mathrm{X}_n, \mathrm{X}_m, \mathrm{N}^{(d)}_{*,n}) = Y_n E(Y_m | \mathrm{X}_n, \mathrm{X}_m, \mathrm{N}^{(d)}_{*,n})\,.$$

The random variable Y_m depends only on the random vectors $\mathrm{N}_{*,m}^{(d)}$ and X_m such that $E(Y_m|\mathrm{X}_n, \mathrm{X}_m, \mathrm{N}_{*,n}^{(d)})$ reduces to $E(Y_m|\mathrm{X}_m)$ and we get

$$\begin{aligned} E(Y_m|\mathrm{X}_m) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left(\ln \left(\left\| \frac{\mathrm{X}_m}{\|\mathrm{X}_m\|} + \sigma x \right\| \right) \right) e^{-\frac{\|x\|^2}{2}} \tilde{L}_m(\lambda, x, \sigma, \sigma_\epsilon) dx - \tilde{\mathrm{F}}(\sigma, \sigma_\epsilon) \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left(\ln(\|e_1 + \sigma x\|) \right) e^{-\frac{\|x\|^2}{2}} \tilde{\mathrm{H}}(\lambda, x, \sigma, \sigma_\epsilon) dx - \tilde{\mathrm{F}}(\sigma, \sigma_\epsilon) = 0 \,, \end{aligned}$$

that implies $E(Y_m Y_n) = 0$. □

Proof of Theorem 4.8

In Proposition 4.7, we show that the random variables $(Y_n)_n$ introduced in Definition 4.6 satisfy the assmuptions of Theorem 4.5. Therefore, the LLN for orthognal random variables applies for the sequence $(Y_n)_n$ in the sense that $\frac{1}{n}\sum_{k=1}^{n} \ln \left(\left\| \frac{\mathrm{X}_k}{\|\mathrm{X}_k\|} + \sigma \mathrm{N}_{*,k}^{(d)} \right\| \right)$ converges almost surely to $F_\times(\sigma, \sigma_\epsilon)$ when n goes to infinity. Then, by Proposition 4.4, we have $\frac{1}{n} \ln \left(\frac{\|\mathrm{X}_n\|}{\|\mathrm{X}_0\|} \right)$ converges almost surely to $F_\times(\sigma, \sigma_\epsilon)$ when n goes to infinity.

Proof of Theorem 4.9

We recall here that the multivariate normal distribution on $\mathbb{R}^d$ with mean $(0, \ldots, 0)$ and covariance matrix the identity I_d, $\mathrm{N}(0, I_d)$, is simply denoted N^d. In the one dimension case, i.e., $d = 1$, it will be simply denoted N. Moreover, for $d \geq 1$, χ_d^2 denotes the chi-square distribution with d degrees of freedom. To prove the theorem, we need the following proposition.

Proposition needed to establish Theorem 4.9

Proposition 4.11. Consider the function F defined in Lemma 4.2. Let σ^* and σ_ϵ^* be two strictly positive constants. The functions H and $\tilde{\mathrm{H}}$ introduced in Definition 4.1 are redefined as mapping $\mathbb{N}^* \times \mathbb{R} \times [0, +\infty[$ into $\mathbb{R}^+$ with

$$\begin{aligned} \mathrm{H}(d, x, u) = \lambda \int_{\mathbb{R}} p_{\mathcal{N}}(y) \times \\ \Pr{}^{\lambda-1} \left[\sqrt{\left(1 + \frac{\sigma^*}{d} x\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 u} \left(1 + \frac{\sigma_\epsilon^*}{d} y\right) \leq \|\mathrm{e}_1 + \frac{\sigma^*}{d} \mathrm{N}^{(d)}\| \left(1 + \frac{\sigma_\epsilon^*}{d} \mathcal{N}\right) \right] dy \,, \end{aligned}$$

and

$$\begin{aligned} \tilde{\mathrm{H}}(d, x, u) = \lambda \int_{\mathbb{R}} p_{\mathcal{N}}(y) \times \\ \Pr{}^{\lambda-1} \left[\sqrt{\left(1 + \frac{\sigma^*}{d} x\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 u} + \frac{\sigma_\epsilon^*}{d} y \leq \|\mathrm{e}_1 + \frac{\sigma^*}{d} \mathrm{N}^{(d)}\| + \frac{\sigma_\epsilon^*}{d} \mathcal{N} \right] dy \,, \end{aligned}$$

for $x \in \mathbb{R}$, $u \in [0, +\infty[$ and $d \in \mathbb{N}^*$. The following holds

$$d \times F_{\times}\left(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d}\right) = E\left[\frac{d}{2}\ln\left(\left(1+\frac{\sigma^*}{d}\mathrm{N}\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 \chi^2_{d-1}\right) \mathrm{H}_{\times}\left(d, \mathrm{N}, \chi^2_{d-1}\right)\right] \quad (4.31)$$

and the family $\left\{\frac{d}{2}\ln\left(\left(1+\frac{\sigma^*}{d}\mathrm{N}\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 \chi^2_{d-1}\right) \mathrm{H}_{\times}\left(d, \mathrm{N}, \chi^2_{d-1}\right)\right\}_{d\geq 1}$, where $\mathrm{H}_{\times}$ stands for H or $\tilde{\mathrm{H}}$, is uniformly integrable.

Proof :
Th proof is given for the model **apf**. The result for the model **pf** is obtained using the same proof. Let us rewrite $\tilde{\mathrm{F}}(\sigma(d), \sigma_\epsilon(d))$ in Eq. 4.15 using $\sigma(d) = \frac{\sigma^*}{d}$, $\sigma_\epsilon(d) = \frac{\sigma^*_\epsilon}{d}$:

$$d \times \tilde{\mathrm{F}}\left(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d}\right) = \frac{\lambda}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{d}{2}\ln(\|e_1 + \frac{\sigma^*}{d}x\|^2) e^{-\frac{\|x\|^2}{2}} \left(\int_{\mathbb{R}} \Pr^{\lambda-1}\left[\left(\|\mathrm{e}_1 + \frac{\sigma^*}{d}x\| + \frac{\sigma^*_\epsilon}{d}y \leq \|\mathrm{e}_1 + \frac{\sigma^*}{d}\mathrm{N}^{(d)}\| + \frac{\sigma^*_\epsilon}{d}\mathcal{N}\right)\right] p_{\mathcal{N}}(y)dy\right) dx\,. \quad (4.32)$$

In the remainder of this proof, the positive quantities σ^*, σ^*_ϵ and λ are fixed. Let $\tilde{\mathrm{H}}$ be the measurable function defined on $\mathbb{N}^* \times \mathbb{R}^d$ by:

$$\tilde{\mathrm{H}}(d, x) = \lambda \int_{\mathbb{R}} \Pr^{\lambda-1}\left[\|\mathrm{e}_1 + \frac{\sigma^*}{d}x\| + \frac{\sigma^*_\epsilon}{d}y \leq \|\mathrm{e}_1 + \frac{\sigma^*}{d}\mathrm{N}^{(d)}\| + \frac{\sigma^*_\epsilon}{d}\mathcal{N}\right] p_{\mathcal{N}}(y)dy\,.$$

The probability of an event E is upper bounded by 1. Therefore, the function $\tilde{\mathrm{H}}$ is upper bounded by λ and $d \times \tilde{\mathrm{F}}(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d})$ can be rewritten as

$$d \times \tilde{\mathrm{F}}\left(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d}\right) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{d}{2}\ln(\|e_1 + \frac{\sigma^*}{d}x\|^2) e^{-\frac{\|x\|^2}{2}} \tilde{\mathrm{H}}(d, x)dx\,. \quad (4.33)$$

Let us apply the change of variables $x_1 = t$, $x_2 = \sqrt{r}\cos(\theta_1)$, $x_3 = \sqrt{r}\sin(\theta_1)\cos(\theta_2)$, $x_4 = \sqrt{r}\sin(\theta_1)\sin(\theta_2)\cos(\theta_3)$, $\ldots$, $x_{d-2} = \sqrt{r}\sin(\theta_1)\ldots\sin(\theta_{d-3})\cos(\theta_{d-2})$ and $x_d = \sqrt{r}\sin(\theta_1)\ldots\sin(\theta_{d-3})\sin(\theta_{d-2})$. Then, for $d \geq 2$, $d \times \tilde{\mathrm{F}}(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d})$ writes as

$$d \times \tilde{\mathrm{F}}\left(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d}\right) = \frac{d}{2\sqrt{2\pi}} \frac{1}{2^{\frac{d-1}{2}}\Gamma(\frac{d-1}{2})} \int_{\mathbb{R}} \int_{[0,+\infty[} \ln\left[\left(1+\frac{\sigma^*}{d}t\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 u\right] u^{\frac{d-1}{2}-1} e^{-\frac{x^2+u}{2}} \tilde{\mathrm{H}}(d, t, u)dtdu\,,$$

where for $t \in \mathbb{R}$, $u \in [0, +\infty[$

$$\tilde{\mathrm{H}}(d, t, u) = \lambda \int_{\mathbb{R}} \Pr^{\lambda-1}\left[\sqrt{\left(1+\frac{\sigma^*}{d}t\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 u} + \frac{\sigma^*_\epsilon}{d}y \leq \|\mathrm{e}_1 + \frac{\sigma^*}{d}\mathrm{N}^{(d)}\| + \frac{\sigma^*_\epsilon}{d}\mathcal{N}\right] p_{\mathcal{N}}(y)dy\,.$$

This means that we have

$$d \times \tilde{\mathrm{F}}\left(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d}\right) = E\left[\frac{d}{2}\ln\left(\left(1+\frac{\sigma^*}{d}\mathrm{N}\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 \chi^2_{d-1}\right)\tilde{\mathrm{H}}\left(d, \mathrm{N}, \chi^2_{d-1}\right)\right] \tag{4.34}$$

where χ^2_{d-1} denote the chi-square distribution with $d-1$ degrees of freedom and

$$\tilde{\mathrm{H}}\left(d, \mathrm{N}, \chi^2_{d-1}\right) = \lambda \int_{\mathbb{R}} p_{\mathcal{N}}(y) \times$$
$$\Pr{}^{\lambda-1}\left[\sqrt{\left(1+\frac{\sigma^*}{d}\mathrm{N}\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 \chi^2_{d-1}} + \frac{\sigma^*_\epsilon}{d} y \leq \|\mathrm{e}_1 + \frac{\sigma^*}{d}\mathrm{N}^{(d)}\| + \frac{\sigma^*_\epsilon}{d}\mathcal{N}|\mathrm{N}, \chi^2_{d-1}\right] dy\,,$$

For fixed $\sigma^* > 0$, let $((\tilde{\mathrm{K}})_d)_{d\geq 1}$ be the sequence of random variables defined as

$$\tilde{\mathrm{K}}_d\left(d, \mathrm{N}, \chi^2_{d-1}\right) := \frac{d}{2}\ln\left(\left(1+\frac{\sigma^*}{d}\mathrm{N}\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 \chi^2_{d-1}\right)\tilde{\mathrm{H}}\left(d, \mathrm{N}, \chi^2_{d-1}\right)$$

Therefore, we get $d \times \tilde{\mathrm{F}}\left(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d}\right) = E\left(\tilde{\mathrm{K}}_d\right)$. Let $\tilde{\mathrm{K}}^+_d$ and $\tilde{\mathrm{K}}^-_d$ be respectively the positive and negative part of the function $\tilde{\mathrm{K}}_d$ such that $\tilde{\mathrm{K}}_d = \tilde{\mathrm{K}}^+_d - \tilde{\mathrm{K}}^-_d$. We have to show that the families of positive random variables $((\tilde{\mathrm{K}})^+_d)_{d\geq 1}$ and $((\tilde{\mathrm{K}})^-_d)_{d\geq 1}$ are uniformly integrable. First, we give interest to the family $((\tilde{\mathrm{K}})^+_d)_{d\geq 1}$. We have

$$\begin{aligned}
(\tilde{\mathrm{K}})^+_d &\leq \frac{\lambda}{2} d \ln^+\left(\left(1+\frac{\sigma^*}{d}\mathrm{N}\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 \chi^2_{d-1}\right) \\
&= \frac{\lambda}{2} d \ln^+\left(1 + 2\frac{\sigma^*}{d}\mathrm{N} + \left(\frac{\sigma^*}{d}\right)^2 \left(\mathrm{N}^2 + \chi^2_{d-1}\right)\right) \\
&\leq \frac{\lambda}{2} d \ln^+\left(1 + 2\frac{\sigma^*}{d}|\mathrm{N}| + \left(\frac{\sigma^*}{d}\right)^2 \left(\mathrm{N}^2 + \chi^2_{d-1}\right)\right) \\
&\leq \frac{\lambda}{2} d \left(2\frac{\sigma^*}{d}|\mathrm{N}| + \left(\frac{\sigma^*}{d}\right)^2 \left(\mathrm{N}^2 + \chi^2_{d-1}\right)\right) \\
&= \lambda\left(\sigma^*|\mathrm{N}| + \frac{(\sigma^*)^2}{2d}\left(\mathrm{N}^2 + \chi^2_{d-1}\right)\right)
\end{aligned} \tag{4.35}$$

According to the last inequality, we have to show that the families $|\mathrm{N}|$ and $(\frac{\mathrm{N}^2+\chi^2_{d-1}}{d})_{d\geq 1}$ are uniformly integrable. The family $|\mathrm{N}|$ contains a unique integrable random variable therefore it is uniformly integrable. The random variable $(\frac{\mathrm{N}^2+\chi^2_{d-1}}{d})_d$ converges (by the Law of Large Numbers) almost surely and therefore in probability to 1. Moreover the sequence of postive real values $E\left[\frac{|\mathrm{N}^2+\chi^2_{d-1}|}{d}\right]_d = 1$ converges to $E\left[|1|\right]$ which gives, by the so-called L^r convergence theorem from [93], that $(\frac{\mathrm{N}^2+\chi^2_{d-1}}{d})_{d\geq 1}$ converges to 1 in the sense

of the norm L^1. Finally, the family $(\frac{\mathrm{N}^2+\chi^2_{d-1}}{d})_{d\geq 1}$ converges in L^1 therefore it is uniformly integrable.
Let us now give interest to the family $((\tilde{\mathrm{K}})^-_d)_{d\geq 2}$. We have

$$\begin{aligned}
(\tilde{\mathrm{K}})^-_d &\leq \frac{\lambda}{2} d \ln^- \left(\left(1+\frac{\sigma^*}{d}\mathrm{N}\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 \chi^2_{d-1} \right) \\
&= \frac{\lambda}{2} d \ln^- \left(\left(1+\frac{\sigma^*}{d}\mathrm{N}\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 \chi^2_{d-1} \right) \mathbb{1}_{\{\mathrm{N}<0\}} \\
&\leq \frac{\lambda}{2} d \ln^- \left(1 - \frac{\mathrm{N}^2}{\mathrm{N}^2+\chi^2_{d-1}} \right) \mathbb{1}_{\{\mathrm{N}<0\}} \\
&= \frac{\lambda}{2} \ln^- \left[\left(1 - \frac{\mathrm{N}^2}{\mathrm{N}^2+\chi^2_{d-1}} \right)^d \right] \mathbb{1}_{\{\mathrm{N}<0\}} \\
&= \frac{\lambda}{2} \ln \left[\left(\frac{1}{1-\frac{\mathrm{N}^2}{\mathrm{N}^2+\chi^2_{d-1}}} \right)^d \right] \mathbb{1}_{\{\mathrm{N}<0\}} \\
&\leq 4\lambda \left(\frac{1}{1-\frac{\mathrm{N}^2}{\mathrm{N}^2+\chi^2_{d-1}}} \right)^{\frac{d}{8}} \mathbb{1}_{\{\mathrm{N}<0\}}
\end{aligned} \tag{4.36}$$

Let us show that the family $(G_d)_{d\geq 2} := \left(\left(\frac{1}{1-\frac{\mathrm{N}^2}{\mathrm{N}^2+\chi^2_{d-1}}} \right)^{\frac{d}{8}} \mathbb{1}_{\{\mathrm{N}<0\}} \right)_{d\geq 2}$ is uniformly integrable. A criterium that can be used to show the uniform integrability of $(G_d)_{d\geq 2}$ is to show that the family

$$\left(E\left[G^2(d)\right]\right)_{d\geq 1} = \left(E\left[\left(\frac{1}{1-\frac{\mathrm{N}^2}{\mathrm{N}^2+\chi^2_{d-1}}} \right)^{\frac{d}{4}} \mathbb{1}_{\{\mathrm{N}<0\}} \right] \right)_{d\geq 1}$$

is uniformly bounded. The expectation $(E\left[G^2(d)\right])$ can be rewritten as:

$$E\left[G^2(d)\right] = \frac{1}{2} E\left[\left(\frac{1}{1-\frac{(\mathrm{N}_1(0,I_d))^2}{\|\mathrm{N}(0,I_d)\|^2}} \right)^{\frac{d}{4}} \right] = \frac{1}{2(2\pi)^{\frac{d}{2}}} \int_{R^d} \left(\frac{1}{1-\frac{(x_1)^2}{\|x\|^2}} \right)^{\frac{d}{4}} e^{-\frac{\|x\|^2}{2}} dx\,,$$

Changing to spherical coordinates (with $d \geq 2$), one gets

$$\begin{aligned}
E\left[G^2(d)\right] &= \frac{1}{2W_{d-2}} \int_0^{\pi/2} \left(\frac{1}{\sin(\theta)} \right)^{\frac{d}{2}} \sin^{d-2}(\theta)\, d\theta \\
&= \frac{1}{2W_{d-2}} \int_0^{\pi/2} \sin^{\frac{d}{2}-2}(\theta)\, d\theta = \frac{W_{\frac{d}{2}-2}}{2W_{d-2}}\,.
\end{aligned}$$

Suppose now that $\frac{d}{2}$ is an integer. Then $\exists p \geq 1$ such that $d = 2p$. As $\lim_{n\to\infty} \sqrt{n} W_n = \sqrt{\pi/2}$ then

$$\lim_{d\to\infty} \frac{W_{\frac{d}{2}-2}}{W_{d-2}} = \lim_{p\to\infty} \frac{W_{p-2}}{W_{2p-2}} = \lim_{p\to\infty} \frac{\sqrt{2p-2}}{\sqrt{p-2}} \frac{\lim_{p\to\infty} \sqrt{p-2} W_{p-2}}{\lim_{p\to\infty} \sqrt{2p-2} W_{2p-2}} = \sqrt{2}\,.$$

If $\frac{d}{2}$ is odd, then $\frac{d-1}{2}$ is an integer and $W_{\frac{d}{2}-2} \leq W_{\frac{d-1}{2}-2}$ and we have also

$$\lim_{d\to\infty} \frac{W_{\frac{d-1}{2}-2}}{W_{d-2}} = \sqrt{2}\,.$$

Then for $d \geq d_0$, $E\left[G^2(d)\right] \leq \frac{\sqrt{2}+1}{2}$. Consequently, the family $\{E\left[G^2(d)\right]\}_{d\geq d_0}$ is uniformly bounded which means that the family $(\tilde{\mathrm{K}}^-)_{d\geq d_0}$ is uniformly integrable and therefore the family

$$\left\{ d \ln\left(\left(1 + \frac{\sigma^*}{d}\mathrm{N}\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 \chi^2_{d-1} \right) \tilde{\mathrm{H}}\left(d, \mathrm{N}, \chi^2_{d-1}\right) \right\}_{d\geq 1}$$

is uniformly integrable.

□

proof of the Theorem First, we show the Theorem for the model **apf**. Let g denote the measurable function defined on $\mathbb{N}^* \times \mathbb{R} \times [0, +\infty[$ for $(d, x, u) \in \mathbb{N}^* \times \mathbb{R} \times [0, +\infty[$ by

$$g(d, x, u) = \left(1 + 2\frac{\sigma^*}{d}x\right)^2 + \left(\frac{\sigma^*}{d}\right)^2 u\,.$$

Let $\bar{\mathrm{N}}$, $\bar{\chi}^2_{d-1}$ and $\bar{\mathcal{N}}$ be random variables respectively distributed as N, χ^2_{d-1} and $\mathcal{N}$. Using the definition of the function $\tilde{\mathrm{H}}$ introduced in Proposition 4.11, one can write

$$\tilde{\mathrm{H}}(d, \mathrm{N}, \chi^2_{d-1}) = \lambda E_{\mathcal{N}}\left[E^{\lambda-1}_{\bar{\mathrm{N}}, \bar{\chi}^2_{d-1}, \bar{\mathcal{N}}}\left(\mathbb{1}_{\left\{\sqrt{g(d,\mathrm{N},\chi^2_{d-1})} + \frac{\sigma^*_\epsilon}{d}\mathcal{N} \leq \sqrt{g(d,\bar{\mathrm{N}},\bar{\chi}^2_{d-1})} + \frac{\sigma^*_\epsilon}{d}\bar{\mathcal{N}}\right\}} | \mathrm{N}, \chi^2_{d-1}, \mathcal{N} \right) \right]\,.$$

The indicator function in the previous equation is upper bounded by 1. Therefore, $\tilde{\mathrm{H}}(\mathrm{N}, \chi^2_{d-1}) \leq \lambda$ and we have

$$d \times \tilde{\mathrm{F}}\left(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d}\right) = E_{\mathrm{N}^{(d)}, \chi^2_{d-1}}\left[\left(\frac{d}{2} \ln(g(d, \mathrm{N}, \chi^2_{d-1})) \right) H(d, \mathrm{N}, \chi^2_{d-1}) \right]\,.$$

By Proposition 4.11, the family $\left\{ \left(\frac{d}{2}\ln(g(d, \mathrm{N}, \chi^2_{d-1}))\right) \tilde{\mathrm{H}}(d, \mathrm{N}, \chi^2_{d-1}) \right\}_{d\geq 1}$ is uniformly integrable then

$$\lim_{d\to\infty} d \times \tilde{\mathrm{F}}\left(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d}\right) = E_{\mathrm{N}^{(d)}, \chi^2_{d-1}}\left[\lim_{d\to\infty} \left(\frac{d}{2} \ln(g(d, \mathrm{N}, \chi^2_{d-1})) \right) \tilde{\mathrm{H}}\left(d, \mathrm{N}, \chi^2_{d-1}\right) \right]\,.$$

Then we have to compute $\lim_{d\to\infty} \frac{d}{2} \ln\left(g\left(d, \mathrm{N}, \chi^2_{d-1}\right)\right)$ and $\lim_{d\to\infty} \tilde{\mathrm{H}}(d, \mathrm{N}, \chi^2_{d-1})$.

$$\frac{d}{2} \ln\left(g(d, \mathrm{N}, \chi^2_{d-1})\right) = \frac{d}{2} \ln\left(1 + 2\frac{\sigma^*}{d}\mathrm{N} + \left(\frac{\sigma^*}{d}\right)^2 \left(\mathrm{N}^2 + \chi^2_{d-1}\right) \right)\,.$$

We have $\ln(1+x) \sim x$ when $x \to 0$. Moreover, $\mathrm{N}^2 + \chi^2_{d-1}$ can be rewritten as the sum of d idenpendent random variables following the distribution of N^2. Therefore, by the LLN for independent identically distributed random variable, we have $\lim_{d\to\infty} \frac{1}{d}\left(\mathrm{N}^2 + \chi^2_{d-1}\right) = E(\mathrm{N}^2) = 1$ almost surely. Consequently, one gets:

$$\begin{aligned}\lim_{d\to\infty} \frac{d}{2} \ln\left(g(d, \mathrm{N}, \chi^2_{d-1})\right) &= \lim_{d\to\infty} d\left(\frac{\sigma^*}{d}\mathrm{N} + \left(\frac{\sigma^*}{2d}\right)^2 \left(\mathrm{N}^2 + \chi^2_{d-1}\right)\right) \\ &= \sigma^* \mathrm{N} + \frac{\sigma^{*2}}{2}\,.\end{aligned}$$

Now, Let us compute the limit of $\tilde{\mathrm{H}}(d, \mathrm{N}, \chi^2_{d-1})$ when d goes to infinity. First, we notice that the acceptance event

$$\left(\sqrt{g(d, \mathrm{N}, \chi^2_{d-1})} + \frac{\sigma^*_\epsilon}{d}\mathcal{N} \leq \sqrt{g(d, \bar{\mathrm{N}}, \bar{\chi}^2_{d-1})} + \frac{\sigma^*_\epsilon}{d}\bar{\mathcal{N}}\right)$$

can be rewritten as

$$\left(d\left[\sqrt{g(d, \mathrm{N}, \chi^2_{d-1})} - 1 + \frac{\sigma^*_\epsilon}{d}\mathcal{N}\right] \leq d\left[\sqrt{g(d, \bar{\mathrm{N}}, \bar{\chi}^2_{d-1})} - 1 + \frac{\sigma^*_\epsilon}{d}\bar{\mathcal{N}}\right]\right)\,.$$

We denote by $\tilde{h}\left(d, \mathrm{N}, \chi^2_{d-1}, \bar{\mathrm{N}}, \bar{\chi}^2_{d-1}\right)$ the quantity

$$\mathbb{1}_{\left\{d\left(\sqrt{g(d,\mathrm{N},\chi^2_{d-1})}-1+\frac{\sigma^*_\epsilon}{d}\mathcal{N}\right)\leq d\left(\sqrt{g(d,\bar{\mathrm{N}},\bar{\chi}^2_{d-1})}-1+\frac{\sigma^*_\epsilon}{d}\bar{\mathcal{N}}\right)\right\}}\,.$$

Then $\tilde{\mathrm{H}}(d, \mathrm{N}, \chi^2_{d-1})$ becomes

$$\tilde{\mathrm{H}}(d, \mathrm{N}, \chi^2_{d-1}) = \lambda E_{\mathcal{N}}\left[E^{\lambda-1}_{\bar{\mathrm{N}},\bar{\chi}^2_{d-1},\bar{\mathcal{N}}}\left(\tilde{h}\left(d, \mathrm{N}, \chi^2_{d-1}, \bar{\mathrm{N}}, \bar{\chi}^2_{d-1}\right)|\mathrm{N}, \chi^2_{d-1}, \mathcal{N}\right)\right]\,.$$

As $\tilde{\mathrm{H}}(d, \mathrm{N}, \chi^2_{d-1}) \leq \lambda$ then by the dominated convergence theorem, we have

$$\lim_{d\to\infty} \tilde{\mathrm{H}}(d, \mathrm{N}, \chi^2_{d-1}) = \lambda E_{\mathcal{N}}\left[E^{\lambda-1}_{\bar{\mathrm{N}},\bar{\chi}^2_{d-1},\bar{\mathcal{N}}}\left(\lim_{d\to\infty} \tilde{h}\left(d, \mathrm{N}, \chi^2_{d-1}, \bar{\mathrm{N}}, \bar{\chi}^2_{d-1}\right)|\mathrm{N}, \chi^2_{d-1}, \mathcal{N}\right)\right]\,.$$

Now, by the (almost sure) continuity of the indicator function, we have

$$\lim_{d\to\infty} \tilde{h}\left(d, \mathrm{N}, \chi^2_{d-1}, \bar{\mathrm{N}}, \bar{\chi}^2_{d-1}\right) = \mathbb{1}_{\left\{\sigma^*\mathrm{N}+\sigma^*_\epsilon\mathcal{N}\leq\sigma^*\bar{\mathrm{N}}+\sigma^*_\epsilon\bar{\mathcal{N}}\right\}}$$

almost surely. Then

$$\lim_{d\to\infty} \tilde{\mathrm{H}}(d, \mathrm{N}, \chi^2_{d-1}) = \lambda E_{\mathcal{N}}\left[E^{\lambda-1}_{\bar{\mathrm{N}},\bar{\mathcal{N}}}\left(\mathbb{1}_{\left\{\sigma^*\mathrm{N}+\sigma^*_\epsilon\mathcal{N}\leq\sigma^*\bar{\mathrm{N}}+\sigma^*_\epsilon\bar{\mathcal{N}}\right\}}|\mathrm{N}, \mathcal{N}\right)\right]\,.$$

Collecting the information above, one gets

$$\lim_{d\to\infty} d \times \tilde{\mathrm{F}}\left(\frac{\sigma^*}{d}, \frac{\sigma^*_\epsilon}{d}\right) = \frac{\sigma^{*2}}{2} + \mathcal{A}(\sigma^*, \sigma^*_\epsilon, \lambda) \times \sigma^*\,, \tag{4.37}$$

where $\mathcal{A}(\sigma^*, \sigma_\epsilon^*, \lambda) := \lambda E_{\mathrm{N}} \left[\mathrm{N}\ E_{\mathcal{N}} \left[E^{\lambda-1}_{\bar{\mathrm{N}},\bar{\mathcal{N}}} \left(\mathbb{1}_{\left\{\sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N} \leq \sigma^* \bar{\mathrm{N}} + \sigma_\epsilon^* \bar{\mathcal{N}}\right\}} | \mathrm{N}, \mathcal{N} \right) \right] \right]$.
For the model **pf**, we similarly get (replacing $\tilde{h}$ for the model **apf** by its analogue h for the model **pf**)

$$\lim_{d\to\infty} d \times \mathrm{F}\left(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d}\right) = \frac{\sigma^{*2}}{2} + \lambda\sigma^* E_{\mathrm{N}^{(d)},\chi^2_{d-1}} \left[\mathrm{N} E_{\mathcal{N}} \left[E^{\lambda-1}_{\bar{\mathrm{N}},\bar{\chi}^2_{d-1},\bar{\mathcal{N}}} \left(\lim_{d\to\infty} h\left(d, \mathrm{N}, \chi^2_{d-1}, \bar{\mathrm{N}}, \bar{\chi}^2_{d-1}\right) | \mathrm{N}, \chi^2_{d-1}, \mathcal{N} \right) \right] \right] .$$

where $h\left(d, \mathrm{N}, \chi^2_{d-1}, \bar{\mathrm{N}}, \bar{\chi}^2_{d-1}\right)$ is given by

$$\mathbb{1}_{\left\{d\left(\sqrt{g(d,\mathrm{N},\chi^2_{d-1})}\left(1+\frac{\sigma_\epsilon^*}{d}\mathcal{N}\right)-1\right)\leq d\left(\sqrt{g(d,\bar{\mathrm{N}},\bar{\chi}^2_{d-1})}\left(1+\frac{\sigma_\epsilon^*}{d}\bar{\mathcal{N}}\right)-1\right)\right\}} .$$

As

$$\lim_{d\to\infty} \tilde{h}\left(d, \mathrm{N}, \chi^2_{d-1}, \bar{\mathrm{N}}, \bar{\chi}^2_{d-1}\right) = \lim_{d\to\infty} h\left(d, \mathrm{N}, \chi^2_{d-1}, \bar{\mathrm{N}}, \bar{\chi}^2_{d-1}\right)$$

then

$$\lim_{d\to\infty} d \times \mathrm{F}\left(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d}\right) = \lim_{d\to\infty} d \times \tilde{\mathrm{F}}\left(\frac{\sigma^*}{d}, \frac{\sigma_\epsilon^*}{d}\right) .$$

To end the proof, we have to show that the quantity $\mathcal{A}(\sigma^*, \sigma_\epsilon^*, \lambda)$ defined in Eq. 4.37 is negative for all $(\sigma^*, \sigma_\epsilon^*, \lambda) \in \mathbb{R}_+^* \times \mathbb{R}_+^* \times \mathbb{N}^*$. The quantity $\mathcal{A}(\sigma^*, \sigma_\epsilon^*, \lambda)$ can be rewritten as

$$\mathcal{A}(\sigma^*, \sigma_\epsilon^*, \lambda) = \int_{\mathbb{R}} x \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \left(\lambda \int_{\mathbb{R}} \Pr{}^{\lambda-1} \left[\sigma^* x + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}\right] p_{\mathcal{N}}(y) dy \right) .$$

Let $x, y \in \mathbb{R}$. If $x \geq 0$ and $\sigma^* x + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}$ then $\sigma^*(-x) + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}$. Therefore, for $x \geq 0$,

$$\Pr{}^{\lambda-1} \left[\sigma^* x + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}\right] \leq \Pr{}^{\lambda-1} \left[\sigma^*(-x) + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}\right] . \tag{4.38}$$

The quantity $\mathcal{A}(\sigma^*, \sigma_\epsilon^*, \lambda)$ can be rewritten as

$$\begin{aligned}\mathcal{A}(\sigma^*, \sigma_\epsilon^*, \lambda) := & \int_{\mathbb{R}^+} x \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \left(\lambda \int_{\mathbb{R}} \Pr{}^{\lambda-1} \left[\sigma^* x + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}\right] p_{\mathcal{N}}(y) dy \right) \\ & + \int_{\mathbb{R}^-} x \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \left(\lambda \int_{\mathbb{R}} \Pr{}^{\lambda-1} \left[\sigma^* x + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}\right] p_{\mathcal{N}}(y) dy \right)\end{aligned}$$

Applying a change of variables, one gets

$$\begin{aligned}\mathcal{A}(\sigma^*, \sigma_\epsilon^*, \lambda) := & \int_{\mathbb{R}^+} x \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \left(\lambda \int_{\mathbb{R}} \Pr{}^{\lambda-1} \left[\sigma^* x + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}\right] p_{\mathcal{N}}(y) dy \right) \\ & - \int_{\mathbb{R}^+} x \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \left(\lambda \int_{\mathbb{R}} \Pr{}^{\lambda-1} \left[\sigma^*(-x) + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}\right] p_{\mathcal{N}}(y) dy \right)\end{aligned}$$

This gives

$$\mathcal{A}(\sigma^*,\sigma_\epsilon^*,\lambda) := \int_{\mathbb{R}^+} \lambda x \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \int_{\mathbb{R}} p_{\mathcal{N}}(y)dy \\ \left(\Pr{}^{\lambda-1}\left[\sigma^* x + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}\right] - \Pr{}^{\lambda-1}\left[\sigma^*(-x) + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}\right]\right) .$$

The result follow from Eq. 4.38. □

Proof of Theorem 4.10

Let us recompute the quantity $\mathcal{A}(\sigma^*,\sigma_\epsilon^*,\lambda)$ in Eq. 4.19. We have

$$\mathcal{A}(\sigma^*,\sigma_\epsilon^*,\lambda) = \frac{\lambda}{\sqrt{2\pi}} \int_{\mathbb{R}} x_1 e^{-\frac{x_1^2}{2}} dx_1 \left(\int_{\mathbb{R}} \Pr{}^{\lambda-1}\left[\sigma^* x_1 + \sigma_\epsilon^* y \leq \sigma^* \mathrm{N} + \sigma_\epsilon^* \mathcal{N}\right] p_{\mathcal{N}}(y)dy \right) .$$

In the case where the noise $\mathcal{N}$ is Gaussian, the random variable $\sigma^*\mathrm{N} + \sigma_\epsilon^*\mathcal{N}$ is a Gaussian variable with mean 0 and variance $\sigma^{*2} + \sigma_\epsilon^{*2}$. Then

$$\mathcal{A}(\sigma^*,\sigma_\epsilon^*,\lambda) = \frac{\lambda}{\sqrt{2\pi}} \int_{\mathbb{R}} x_1 e^{-\frac{x_1^2}{2}} dx_1 \left(\int_{\mathbb{R}} \Pr{}^{\lambda-1}\left[\frac{\sigma^* x_1 + \sigma_\epsilon^* y}{\sqrt{\sigma^{*2} + \sigma_\epsilon^{*2}}} \leq \mathrm{N}\right] \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \right) .$$

Applying the change of variables $t = x_1$ and $s = \frac{\sigma^* x_1 + \sigma_\epsilon^* y}{\sqrt{\sigma^{*2}+\sigma_\epsilon^{*2}}}$, we get

$$\mathcal{A}(\sigma^*,\sigma_\epsilon^*,\lambda) = \sqrt{1+\left(\frac{\sigma^*}{\sigma_\epsilon^*}\right)^2} \frac{\lambda}{2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} t e^{-\frac{t^2}{2}} e^{-\frac{1}{2}\left(\sqrt{1+\left(\frac{\sigma^*}{\sigma_\epsilon^*}\right)^2} s - \frac{\sigma^*}{\sigma_\epsilon^*} t\right)^2} dt \left[1-\phi(s)\right]^{\lambda-1} ds .$$

Now, from the appendix of [25] (Eq.A.8), we know that for $(a,b) \in \mathbb{R}^2$,

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} t e^{-\frac{t^2}{2}} e^{-\frac{1}{2}(at+b)^2} dt = \frac{-ab \exp\left(-\frac{1}{2}\frac{b^2}{1+a^2}\right)}{\sqrt{1+a^2}\,(1+a^2)} . \tag{4.39}$$

Using Eq. 4.39, we get

$$\mathcal{A}(\sigma^*,\sigma_\epsilon^*,\lambda) = \sqrt{1+\left(\frac{\sigma^*}{\sigma_\epsilon^*}\right)^2} \frac{\lambda}{\sqrt{2\pi}} \int_{\mathbb{R}} \left[\frac{\frac{\sigma^*}{\sigma_\epsilon^*} s \exp\left(-\frac{1}{2}s^2\right)}{1+\left(\frac{\sigma^*}{\sigma_\epsilon^*}\right)^2} \right] \left[1-\phi(s)\right]^{\lambda-1} ds .$$

Thus

$$\mathcal{A}(\sigma^*,\sigma_\epsilon^*,\lambda) = \frac{\lambda}{\sqrt{2\pi}} \frac{1}{\sqrt{1+\left(\frac{\sigma_\epsilon^*}{\sigma^*}\right)^2}} \int_{\mathbb{R}} s\, e^{-\frac{1}{2}s^2} \left[1-\phi(s)\right]^{\lambda-1} ds .$$

Using the symmetry property stating that for any s in $\mathbb{R}$, $1-\phi(s) = \phi(-s)$, one has

$$\mathcal{A}(\sigma^*,\sigma_\epsilon^*,\lambda) = \frac{\lambda}{\sqrt{2\pi}} \frac{1}{\sqrt{1+\left(\frac{\sigma_\epsilon^*}{\sigma^*}\right)^2}} \int_{-\infty}^{+\infty} s\, e^{-\frac{1}{2}s^2} \left[\phi(-s)\right]^{\lambda-1} ds .$$

After substituting $u = -s$, one gets

$$\mathcal{A}(\sigma^*, \sigma_\epsilon^*, \lambda) = -\left[\frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u\ e^{-\frac{1}{2}u^2} \left[\phi(u)\right]^{\lambda-1} du\right] \frac{1}{\sqrt{1 + \left(\frac{\sigma_\epsilon^*}{\sigma^*}\right)^2}}\,.$$

Consequently

$$\mathcal{A}(\sigma^*, \sigma_\epsilon^*, \lambda) = -c(1, \lambda) \frac{1}{\sqrt{1 + \left(\frac{\sigma_\epsilon^*}{\sigma^*}\right)^2}}\,.$$

□

Application

Chapter 5

Identification of the Isotherm Function in Chromatography Using CMA-ES

The main material contained in this chapter is the paper [78] published in the Proceedings of the *2007 IEEE Congress on Evolutionary Computation* conference. The work presented here has been funded by the CNRS program ACI NIM (Nouvelles Interfaces des Mathématiques – New Frontiers for Mathematics) *Chromalgema*, coordinated by F. James (University of Orléans), and is a joint work with François James and Marie Postel (University Pierre et Marie Curie – Paris 6).

The goal is to solve an identification problem arising from a model of analytic chromatography, a technique used by chemical engineers. Chromatography aims at separating the m components of a mixture (that can be a gas or a liquid) by injecting the mixture in a column of length L filled by a porous medium (generally a solid, but sometimes a liquid). Pushed by a continuous injection of an inert medium, the different components of the mixture moves through the column at different speeds, due to their different affinities with the porous medium in the column. The different components of the mixture reach the end of the column at different times. In a perfectly linear world, and if the column was long enough, each component would have its own propagation speed, and the separation would be perfect. However, because the propagation speed of each component depends on the concentrations of the other components, the model is non-linear and the components are not perfectly separated, whatever the column length. It is however very important to be able to predict when this or that component will be highly concentrated at the end of the column. The 'output' concentration vector (one concentration per component) at the end of the column is called a *chromatogram* and will be denoted $\mathbf{c}(t, L)$ $(t \in [0, T])$.

Writing the mass balance of the system leads to a system of Partial Differential Equations [140] that has been shown to be a non-linear hyperbolic system [141]. The unknown are the concentrations $\mathbf{c}(t, z), t \in [0, T], z \in [0, L]$ and the 'flux' $\mathbf{F}$ of this system involves what chemists call the *isotherm function* of the process (because the temperature is fixed during the whole process). Solving the direct problem, i.e. computing the output chromatogram from the initial conditions and the concentrations that are injected in the column during the whole experiment, thus amounts to solving the system of PDEs (5.2), with flux given by Equation (5.3).

Because this system is hyperbolic, it is well-known that it has a unique solution, and

many numerical schemata can be used in order to numerically approximate its solution. Also, because all eigenvalues of this system are positive [141], the standard Godunov scheme here amounts to a simple forward finite difference discretization, and the resulting discrete system is numerically stable under the so-called CFL condition given by Eq. 5.5.

The goal The art of chromatography separation requires knowing when to gather the output of the column to reach a desired level of purity of the products. This can be easily computed provided the numerical model described above gives a good prediction of the chromatogram. However, the accuracy of the prediction given by the numerical solution of system (5.2) highly depends on the validity of the isotherm function for the actual chemical system at hand – and isotherm functions are not precisely known by chemists in the case of multiple components. Moreover, there are very few data points that would allow the engineers to fit an approximate model, and acquiring a new data point requires several months of tedious experiment. On the other hand, it is much easier to experiment with a given chromatographic column, recording both the input concentrations and the corresponding output chromatograms. It should hence be possible to identify the isotherm function from those data by solving the inverse problem: find the isotherm function $\mathbf{H}$ such that the numerical solution of system (5.2) with the given input fits the experimental chromatogram as accurately as possible.

More formally, this problem can be turned into a minimization problem: given an experimental chromatogram $\mathbf{c}_{exp}(t)$, $t \in [0, T]$, find the isotherm function $\mathbf{H}$ such that the solution of the direct system given in Eq. 5.2 minimizes the cost function $\mathcal{J}$ computed as the least square difference between the computed chromatogram $\mathbf{c}_{\mathbf{H}}(t, L)$ and the experimental one $\mathbf{c}_{exp}(t)$:

$$\mathcal{J}(\mathbf{H}) = \int_0^T \|\mathbf{c}_{\mathbf{H}}(t, L) - \mathbf{c}_{exp}(t)\|^2 dt \tag{5.1}$$

Chemical scientists have introduced several parametric models for isotherm functions (see Section 5.3.2 for a presentation of some models). The resulting optimization problem hence amounts to parametric optimization. This parametric optimization problem has alreay been adressed using gradient-based approaches [73, 74]. However, the function to optimize is not convex, and experiments performed in [73] suggest that the function is multi-modal. An additional difficulty induced by the computation of the fitness function is that the CFL stability condition can be violated during the optimization, leading to infeasible individuals (in the sense that no value can be computed for the $\mathcal{J}$ function) without any easy way to a priori predict for a given set of parameter whether this will happen or not. Finally, the different variables of the problem have very different scales.

Implementation and results The minimization of the cost function $\mathcal{J}$, as a function of the parameters of some parametric model for the isotherm function, is addressed using the Covariance Matrix Adaptation-Evolution Strategy (CMA-ES, see Section 5.4.2). The implmentation that has been used here is that described in [16] and written in Scilab, that has been interfaced with the C++ code developed during the ACI *Chromalgema* for

the fitness function [107]. This approach has been tested on the real data set provided in [73], and results compared with those of the gradient based approach provided on the same publication. Note that in [73], the gradient based approach adopted is the conjugate gradient method of the discretized cost function. The gradient of the cost function $\mathcal{J}$ with respect to parameters of the isotherm function is obtained as follows: A discretized expression of the (parametric) cost function $\mathcal{J}(\alpha_1, \ldots, \alpha_m)$ where $\alpha_1, \ldots, \alpha_m$ are the parameters to identify is computed. Then the gradient of the discretized cost function with respect to the parameters to identify is computed and used as an estimator of the gradient of the continuous formulation of the optimization problem in a conjugate gradient approach. Our study shows that randomized search methods can perform better that the gradient-based on this problem. In fact, CMA-ES is more robust as it always converges to the same point, independently of the starting point – and this was clearly not the case for the gradient approach. Moreover, CMA-ES is more efficient in solving the problem at hand as it proposed more accurate solutions for two different configurations of the parameters to identify. In particular, CMA-ES was able to handle the complete identification problem, whereas the gradient approach required that some parameter values are pre-determined using some experimental values. Another fact that has been learned during this case study is that the two approaches (CMA-ES and gradient) have very similar computation times: this is quite unusual as deterministic methods are in general much faster than population based randomized search methods.

A common drawback of both the gradient-based and CMA-ES approaches is the poor fit of the identified chromatogam with the (sparse) data points that the chemists had gathered for the isotherm function – though the chromatograms were nicely fitted. This suggests to use a multi-objective approach, fitting both the chromatogram through solving the direct problem, and directly fitting the isotherm using the few data available points.

Identification of the Isotherm function in Chromatography Using CMA-ES

Mohamed Jebalia[1], Anne Auger[1], Marc Schoenauer[1], François James[2] and Marie Postel[3]

[1] TAO Team, INRIA Futurs
Université Paris Sud, LRI, 91405 Orsay cedex, France
{jebalia,auger,marc}@lri.fr

[2] MAPMO, UMR-CNRS 6628
BP 6759 - 45067 Orléans Cedex 2, France
francois.james@math.cnrs.fr

[3] Laboratoire Jacques-Louis Lions
UPMC, Boîte courrier 187, 75252 Paris cedex 05, France
postel@ann.jussieu.fr

Proceedings of the 2007 IEEE Congress on Evolutionary Computation, pp 4289-4296.

Erratum :
In Section 5.4.2, the sentence " An important property of CMA-ES is its invariance to linear transformations of the search space." should be replaced by " An important property of CMA-ES is its invariance to orthogonal transformations of the search space."

Identification of the Isotherm function in Chromatography Using CMA-ES

Mohamed Jebalia[1], Anne Auger[1], Marc Schoenauer[1], François James[2] and Marie Postel[3]

[1] TAO Team, INRIA Futurs
Université Paris Sud, LRI, 91405 Orsay cedex, France
{jebalia,auger,marc}@lri.fr

[2] Mathématiques, Applications et Physique Mathématique d'Orléans - France
francois.james@math.cnrs.fr

[3] Laboratoire Jacques-Louis Lions - UPMC Paris - France
postel@ann.jussieu.fr

Abstract

This paper deals with the identification of the flux for a system of conservation laws in the specific example of analytic chromatography. The fundamental equations of chromatographic process are highly non linear. The state-of-the-art Evolution Strategy, CMA-ES (the Covariance Matrix Adaptation Evolution Strategy), is used to identify the parameters of the so-called isotherm function. The approach was validated on different configurations of simulated data using either one, two or three components mixtures. CMA-ES is then applied to real data cases and its results are compared to those of a gradient-based strategy.

5.1 Introduction

The chromatography process is a powerful tool to separate or analyze mixtures [50]. It is widely used in chemical industry (pharmaceutical, perfume and oil industry, etc) to produce relatively high quantities of very pure components. This is achieved by taking advantage of the selective absorption of the different components in a solid porous medium. The moving fluid mixture is percolated through the motionless medium in a column. The various components of the mixture propagate in the column at different speeds, because of their different affinities with the solid medium. The art of chromatography separation requires predicting the different proportions of every component of the mixture at the

end of the column (called *the chromatogram*) during the experiment. In the ideal (linear) case, every component has its own fixed propagation speed, that does not depend on the other components. In this case, if the column is sufficiently long, pure components come out at the end of the column at different times: they are perfectly separated. But in the real world, the speed of a component heavily depends on every other component in the mixture. Hence, the fundamental Partial Differential Equations of the chromatographic process, derived from the mass balance, are highly non linear. The process is governed by a nonlinear function of the mixture concentrations, the so-called *Isotherm Function*. This function computes the amount of absorbed quantity of each component w.r.t. all other components.

Mathematically speaking, thermodynamical properties of the isotherm ensure that the resulting system of PDEs is hyperbolic, and standard numerical tools for hyperbolic systems can hence be applied; if the isotherm is known: The precise knowledge of the isotherm is crucial, both from the theoretical viewpoint of physico-chemical modeling and regarding the more practical preoccupation of accurately controlling the experiment to improve separation. Specific chromatographic techniques can be used to directly identify the isotherm, but gathering a few points requires several months of careful experiments. Another possible approach to isotherm identification consists in solving the inverse problem numerically: find the isotherm such that numerical simulations result in chromatograms that are as close as possible to the actual experimental outputs.

This paper introduces an evolutionary method to tackle the identification of the isotherm function from experimental chromatograms. The goal of the identification is to minimize the difference between the actual experimental chromatogram and the chromatogram that results from the numerical simulation of the chromatographic process. Chemical scientists have introduced several parametric models for isotherm functions (see [50] for all details of the most important models). The resulting optimization problem hence amounts to parametric optimization, that is addressed here using the state-of-the-art Evolution Strategy, CMA-ES. Section 5.2 introduces the direct problem and Section 5.3 the optimization (or inverse) problem. Section 5.4.1 reviews previous approaches to the problem based on gradient optimization algorithms [74, 73]. Section 5.4.2 details the CMA-ES method and the implementation used here. Finally, Section 5.5 presents experimental results: first, simulated data are used to validate the proposed approach; second, real data are used to compare the evolutionary approach with a gradient-based method.

5.2 Physical problem and model

Chromatography aims at separating the components of a mixture based on the selective absorption of chemical species by a solid porous medium. The fluid mixture moves down through a column of length L, considered here to be one-dimensional. The various components of the mixture propagate in the column at different speeds, because of their different behavior when interacting with the porous medium. At a given time $t \in \mathbb{R}^+$, for a given $z \in [0, L]$ the concentration of m species is a real vector of $\mathbb{R}^m$ denoted $\mathbf{c}(t, z)$.

The evolution of $\mathbf{c}$ is governed by the following partial differential equation:

$$\begin{cases} \partial_z \mathbf{c} + \partial_t \mathbf{F}(\mathbf{c}) = 0, \\ \mathbf{c}(0, z) = \mathbf{c}_0(z), \\ \mathbf{c}(t, 0) = \mathbf{c}_{inj}(t). \end{cases} \tag{5.2}$$

where $\mathbf{c}_0 : \mathbb{R} \to \mathbb{R}^m$ is the initial concentration, $\mathbf{c}_{inj} : \mathbb{R} \to \mathbb{R}^m$ the injected concentration at the entrance of the column and $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^m$ is the flux function that can be expressed in the following way

$$\mathbf{F}(\mathbf{c}) = \frac{1}{u}\left(\mathbf{c} + \frac{1-\epsilon}{\epsilon}\mathbf{H}(\mathbf{c})\right) \tag{5.3}$$

where $\mathbf{H} : \mathbb{R}^m \to \mathbb{R}^m$ is the so-called isotherm function, $\epsilon \in (0, 1)$ and $u \in \mathbb{R}^+$ [73]. The Jacobian matrix of $\mathbf{F}$ being diagonalizable with strictly positive eigenvalues, the system (5.2) is strictly hyperbolic and thus admits an unique solution as soon as $\mathbf{F}$ is continuously differentiable, and the initial and injection conditions are piecewise continuous. The solution of Eq. 5.2 can be approximated using any finite difference method that is suitable for hyperbolic systems [48]. A uniform grid in space and time of size $(K+1) \times (N+1)$ is defined: Let Δz (resp. Δt) such that $K\Delta z = L$ (resp. $N\Delta t = T$). Then an approximation of the solution of Eq. 5.2 can be computed with the Godunov scheme:

$$\mathbf{c}_{k+1}^n = \mathbf{c}_k^n - \frac{\Delta z}{\Delta t}(\mathbf{F}(\mathbf{c}_k^n) - \mathbf{F}(\mathbf{c}_k^{n-1})) \tag{5.4}$$

where $\mathbf{c}_k^n$ is an approximation of the mean value of the solution $\mathbf{c}$ at point $(k\Delta z, n\Delta t)$[11]. For a fixed value of $\frac{\Delta z}{\Delta t}$, the solution of Eq. 5.4 converges to the solution of Eq. 5.2 as Δt and Δz converge to zero. The numerical scheme given in Eq. 5.4 is numerically stable under the so-called CFL condition stating that the largest absolute value of the eigenvalues of the Jacobian matrix of $\mathbf{F}$ is upper-bounded by a constant

$$\frac{\Delta z}{\Delta t} \max_c \mathrm{Sp}(|\mathbf{F}'(c)|) \leq \mathrm{CFL} < 1. \tag{5.5}$$

5.3 The Optimization Problem

5.3.1 Goal

The goal is to identify the isotherm function from experimental chromatograms: given initial data c_0, injection data c_{inj}, and the corresponding experimental chromatogram c_{exp} (that can be either the result of a simulation using a known isotherm function, or the result of actual experiments by chemical scientists), find the isotherm function $\mathbf{H}$ such that the numerical solution of Eq. 5.2 using the same initial and injection conditions results in a chromatogram as close as possible to the experimental one c_{exp}.

[11] Mean value over the volume defined by the corresponding cell of the grid.

Ideally, the goal is to find $\mathbf{H}$ such that the following system of PDEs has a unique solution $\mathbf{c}(t, z)$:

$$\begin{cases} \partial_z \mathbf{c} + \partial_t \mathbf{F}(\mathbf{c}) = 0, \\ \mathbf{c}(0, z) = \mathbf{c}_0(z), \\ \mathbf{c}(t, 0) = \mathbf{c}_{inj}(t), \\ \mathbf{c}(t, L) = \mathbf{c}_{exp}(t). \end{cases} \tag{5.6}$$

However, because in most real-world cases this system will not have an exact solution, it is turned into a minimization problem. For a given isotherm function $\mathbf{H}$, solve system 5.2 and define the cost function $\mathcal{J}$ as the least square difference between the computed chromatogram $\mathbf{c}_{\mathbf{H}}(t, L)$ and the experimental one $\mathbf{c}_{exp}(t)$:

$$\mathcal{J}(\mathbf{H}) = \int_0^T \|\mathbf{c}_{\mathbf{H}}(t, L) - \mathbf{c}_{exp}(t)\|^2 dt \tag{5.7}$$

If many experimental chromatograms are provided, the cost function is the sum of such functions $\mathcal{J}$ computed for each experimental chromatogram.

5.3.2 Search Space

When tackling a function identification problem, the first issue to address is the parametric vs non-parametric choice [120]: parametric models for the target function result in parametric optimization problems that are generally easier to tackle – but a bad choice of the model can hinder the optimization. On the other hand, non-parametric models are a priori less biased, but search algorithms are also less efficient on large unstructured search space.

Early trials to solve the chromatography inverse problem using a non-parametric model (recurrent neural-network) have brought a proof-of-concept to such approach [43], but have also demonstrated its limits: only limited precision could be reached, and the approach poorly scaled up with the number of components of the mixture.

Fortunately, chemists provide a whole zoology of parametrized models for the isotherm function $\mathbf{H}$, and using such models, the identification problem amounts to parametric optimization. For $i \in \{1, \dots, m\}$, denote $\mathbf{H}_i$ the component i of the function $\mathbf{H}$. The main models for the isotherm function that will be used here are the following:

- The **Langmuir** isotherm [89] assumes that the different components are in competition to occupy each site of the porous medium. This gives, for all $i = 1, \dots, m$

$$\mathbf{H}_i(c) = \frac{\mathbf{N}^*}{1 + \sum_{l=1}^{m} \mathbf{K}_l \mathbf{c}_l} \mathbf{K}_i \mathbf{c}_i. \tag{5.8}$$

  There are $m + 1$ positive parameters: the *Langmuir coefficients* $(\mathbf{K}_i)_{i \in [1,m]}$, homogeneous to the inverse of a concentration, and the *saturation coefficient* $\mathbf{N}^*$ that corresponds to some limit concentration.

- The **Bi-Langmuir** isotherm generalizes the Langmuir isotherm by assuming two different kinds of sites on the absorbing medium. The resulting equations are, for all $i = 1, \ldots, m$

$$\mathbf{H}_i(c) = \sum_{s\in\{1,2\}} \frac{\mathbf{N}_s^*}{1 + \sum_{l=1}^{m} \mathbf{K}_{l,s}\mathbf{c}_l} \mathbf{K}_{i,s}\mathbf{c}_i. \tag{5.9}$$

  This isotherm function here depends on $2(m + 1)$ parameters: the generalized Langmuir coefficients $(\mathbf{K}_{i,s})_{i\in[1,m],s=1,2}$ and the generalized saturation coefficients $(\mathbf{N}_s^*)_{s=1,2}$.

- The **Lattice** isotherm [141] is a generalization of Langmuir isotherm that also considers interactions among the different sites of the porous medium. Depending on the degree d of interactions (number of interacting sites grouped together), this model depends, additionally to the Langmuir coefficients $(\mathbf{K}_i)_{i\in[1,m]}$ and the saturation coefficient $\mathbf{N}^*$, on interaction energies $(\mathbf{E}_{ij})_{i,j\in[0,d],2\leq i+j\leq d}$ resulting in $\prod_{i=1}^{m} \frac{d+i}{i}$ parameters. For instance, for one component ($m = 1$) and degree 2, this gives:

$$\mathbf{H}_1(\mathbf{c}) = \frac{\mathbf{N}^*}{2} \frac{\mathbf{K}_1\ \mathbf{c} + e^{-\frac{\mathbf{E}_{11}}{RT}} (\mathbf{K}_1\ \mathbf{c})^2}{1 + 2\mathbf{K}_1\ \mathbf{c} + e^{-\frac{\mathbf{E}_{11}}{RT}} (\mathbf{K}_1\ \mathbf{c})^2}, \tag{5.10}$$

  where T is the absolute temperature and R is the universal gas constant. Note that in all cases, a Lattice isotherm with 0 energies simplifies to the Langmuir isotherm with the same Langmuir and saturation coefficients up to a factor $\frac{1}{2}$.

5.4 Approach Description

5.4.1 Motivations

Previous works on parametric optimization of the chromatography inverse problem have used gradient-based approaches [74, 73]. In [74], the gradient of $\mathcal{J}$ is obtained by writing and solving numerically the adjoint problem, while direct differentiation of the discretized equation have also been investigated in [73]. However the fitness function to optimize is not necessarily convex and no results are provided for differentiability. Moreover, experiments performed in [73] suggest that the function is multimodal, since the gradient algorithm converges to different local optima depending on the starting point. Evolutionary algorithms (EAs) are stochastic global optimization algorithms, less prone to get stuck in local optima than gradient methods, and do not rely on convexity assumptions. Thus they seem a good choice to tackle this problem. Among EAs, Evolution Strategies have been specifically designed for continuous optimization. The next section introduces the state of the art EA for continuous optimization, the covariance matrix adaptation ES (CMA-ES).

5.4.2 The CMA Evolution Strategy

CMA-ES is a stochastic optimization algorithm specifically designed for continuous optimization [61, 59, 57, 16]. At each iteration g, a population of points of an n-dimensional

continuous search space (subset of $\mathbb{R}^n$), is sampled according to a multi-variate normal distribution. Evaluation of the fitness of the different points is then performed, and parameters of the multi-variate normal distribution are updated.

More precisely, let $\langle \vec{x} \rangle_{\mathrm{W}}^{(g)}$ denotes the mean value of the (normally) sampling distribution at iteration g. Its covariance matrix is usually factorized in two terms: $\sigma^{(g)} \in \mathbb{R}^+$, also called the *step-size*, and $\boldsymbol{C}^{(g)}$, a definite positive $n \times n$ matrix, that is abusively called the covariance matrix. The independent sampling of the λ offspring can then be written:

$$\vec{x}_k^{(g+1)} = \langle \vec{x} \rangle_{\mathrm{W}}^{(g)} + \mathcal{N}_k\left(0, (\sigma^{(g)})^2 \boldsymbol{C}^{(g)}\right) \text{ for } k = 1, \ldots, \lambda$$

where $\mathcal{N}_k(0, M)$ denote independent realizations of the multi-variate normal distribution of covariance matrix M.

The μ best offspring are recombined into

$$\langle \vec{x} \rangle_{\mathrm{W}}^{(g+1)} = \sum_{i=1}^{\mu} w_i \vec{x}_{i:\lambda}^{(g+1)} \quad , \tag{5.11}$$

where the positive weights $w_i \in \mathbb{R}$ are set according to individual ranks and sum to one. The index $i{:}\lambda$ denotes the i-th best offspring. Eq. 5.11 can be rewritten as

$$\langle \vec{x} \rangle_{\mathrm{W}}^{(g+1)} = \langle \vec{x} \rangle_{\mathrm{W}}^{(g)} + \sum_{i=1}^{\mu} w_i \mathcal{N}_{i:\lambda}\left(0, (\sigma^{(g)})^2 \boldsymbol{C}^{(g)}\right) \quad , \tag{5.12}$$

The covariance matrix $\boldsymbol{C}^{(g)}$ is a positive definite symmetric matrix. Therefore it can be decomposed in

$$\boldsymbol{C}^{(g)} = \boldsymbol{B}^{(g)} \boldsymbol{D}^{(g)} \boldsymbol{D}^{(g)} \left(\boldsymbol{B}^{(g)}\right)^T \quad ,$$

where $\boldsymbol{B}^{(g)}$ is an orthogonal matrix, *i.e.* $\boldsymbol{B}^{(g)} \left(\boldsymbol{B}^{(g)}\right)^T = I_d$ and $\boldsymbol{D}^{(g)}$ a diagonal matrix whose diagonal contains the square root of the eigenvalues of $\boldsymbol{C}^{(g)}$.

The so-called strategy parameters of the algorithm, the covariance matrix $\boldsymbol{C}^{(g)}$ and the step-size $\sigma^{(g)}$, are updated so as to increase the probability to reproduce good steps. The so-called rank-one update for $\boldsymbol{C}^{(g)}$ [61] takes place as follows. First, an evolution path is computed:

$$\vec{p}_c^{(g+1)} = (1 - c_c)\vec{p}_c^{(g)} + \frac{\sqrt{c_c(2 - c_c)\mu_{\mathrm{eff}}}}{\sigma^{(g)}} \left(\langle \vec{x} \rangle_{\mathrm{W}}^{(g+1)} - \langle \vec{x} \rangle_{\mathrm{W}}^{(g)}\right)$$

where $c_c \in]0, 1]$ is the cumulation coefficient and μ_{eff} is a strictly positive coefficient. This evolution path can be seen as the descent direction for the algorithm.

Second the covariance matrix $\boldsymbol{C}^{(g)}$ is "elongated" in the direction of the evolution path, *i.e.* the rank-one matrix $\vec{p}_c^{(g+1)} \left(\vec{p}_c^{(g+1)}\right)^T$ is added to $\boldsymbol{C}^{(g)}$:

$$\boldsymbol{C}^{(g+1)} = (1 - c_{\mathrm{cov}})\boldsymbol{C}^{(g)} + c_{\mathrm{cov}} \vec{p}_c^{(g+1)} \left(\vec{p}_c^{(g+1)}\right)^T$$

where $c_{\mathrm{cov}} \in]0, 1[$. The complete update rule for the covariance matrix is a combination of the rank-one update previously described and the rank-mu update presented in [59].

The update rule for the step-size $\sigma^{(g)}$ is called the path length control. First, another evolution path is computed:

$$\vec{p}_\sigma^{(g+1)} = (1-c_\sigma)\vec{p}_\sigma^{(g)} + \frac{\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}}{\sigma^{(g)}} \times \boldsymbol{B}^{(g)}\boldsymbol{D}^{(g)^{-1}}\boldsymbol{B}^{(g)^T}\left(\langle\vec{x}\rangle_{\text{W}}^{(g+1)} - \langle\vec{x}\rangle_{\text{W}}^{(g)}\right) \quad (5.13)$$

where $c_\sigma \in]0,1]$. The length of this vector is compared to the length that this vector would have had under random selection, *i.e.* in a scenario where no information is gained from the fitness function and one is willing to keep the step-size constant. Under random selection the vector $\vec{p}_\sigma^{(g)}$ is distributed as $\mathcal{N}(0, I_d)$. Therefore, the step-size is increased if the length of $\vec{p}_\sigma^{(g)}$ is larger than $\mathbf{E}(\parallel \mathcal{N}(0, I_d) \parallel)$ and decreased if it is shorter. Formally, the update rule reads:

$$\sigma^{(g+1)} = \sigma^{(g)} \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\parallel \vec{p}_\sigma^{(g+1)} \parallel}{\mathbf{E}(\parallel \mathcal{N}(0, I_d) \parallel)} - 1\right)\right) \quad (5.14)$$

where $d_\sigma > 0$ is a damping factor.

The default parameters for CMA-ES were carefully derived in [57], Eqs. 6-8. The only problem-dependent parameters are $\langle\vec{x}\rangle_{\text{W}}^{(0)}$ and $\sigma^{(0)}$, and, to some extend, the offspring size λ: its default value is $\lfloor 4+3\log(n)\rfloor$ (the μ default value is $\lfloor\frac{\lambda}{2}\rfloor$), but increasing λ increases the probability to converge towards the global optimum when minimizing multimodal fitness functions [57].

This fact was systematically exploited in [16], where a "CMA-ES restart" algorithm is proposed, in which the population size is increased after each restart. Different restart criteria are used:

1. *RestartTolFun*: Stop if the range of the best objective function values of the recent generation is below than a TolFun value.

2. *RestartTolX*: Stop if the standard deviation of the normal distribution is smaller than a TolX value and $\sigma\vec{p}_c$ is smaller than TolX in all components.

3. *RestartOnNoEffectAxis*: Stop if adding a 0.1 standard deviation vector in a principal axis direction of $\boldsymbol{C}^{(g)}$ does not change $\langle\vec{x}\rangle_{\text{W}}^{(g)}$.

4. *RestartCondCov*: Stop if the condition number of the covariance matrix exceeds a fixed value.

The resulting algorithm (the CMA-ES restart, simply denoted CMA-ES in the remainder of this paper) is a quasi parameter free algorithm that performed best for the CEC 2005 special session on parametric optimization [2].

An important property of CMA-ES is its invariance to linear transformations of the search space. Moreover, because of the rank-based selection, CMA-ES is invariant to any monotonous transformation of the fitness function: optimizing f or $h \circ f$ is equivalent, for any rank-preserving function $h : \mathbb{R} \to \mathbb{R}$. In particular, convexity has no impact on the actual behavior of CMA-ES.

5.4.3 CMA-ES Implementation

This section describes the specific implementation of CMA-ES to identify n isotherm coefficients. For the sake of clarity we will use a single index in the definition of the coefficients of the isotherm, *i.e* we will identify $\mathbf{K}_a$, $\mathbf{N}_b^*$ and $\mathbf{E}_c$ for $a \in [1, A]$, $b \in [1, B]$ and $c \in [1, C]$ where A, B and C are integers summing up to n.

Fitness function and CFL condition The goal is to minimize the fitness function defined in Section 5.3.1. In the case where identification is done using only one experimental chromatogram, the fitness function is the function $\mathcal{J}$ defined in Eq. 5.7 as the least squared difference between an experimental chromatogram $\mathbf{c}_{exp}(t)$ obtained using experimental conditions $\mathbf{c}_0$, $\mathbf{c}_{inj}$ and a numerical approximation of the solution of system (5.2) for a candidate isotherm function $\mathbf{H}$ using the same experimental conditions. The numerical simulation of a solution of Eq. 5.2 is computed with a Godunov scheme written in C++ (see [107] for the details of the implementation).

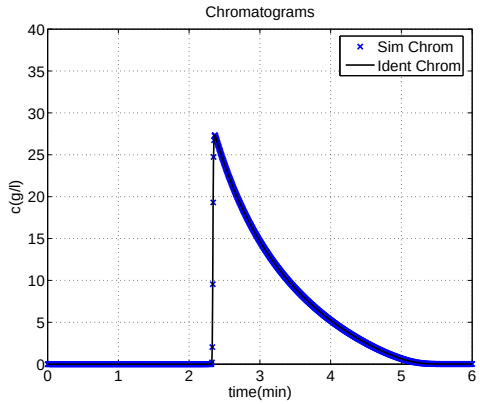
In order to validate the CMA-ES approach, first "experimental" chromatograms were in fact computed using numerical simulations of Eq. 5.2 with different experimental conditions. Let $\mathbf{F}_{sim}$ denotes the flux function used to simulate the experimental chromatogram. For the simulation of an approximated solution of Eq. 5.2, a time step Δt and a CFL coefficient strictly smaller than one (typically 0.8) are fixed beforehand. The quantity $\max \mathrm{Sp}(|\mathbf{F}'_{sim}(c)|)$ is then estimated using a power method, and the space step Δz can then be set such that Eq. 5.5 is satisfied for $\mathbf{F}_{sim}$. The same Δt and Δz are then used during the optimization of $\mathcal{J}$.

When $\mathbf{c}_{exp}$ comes from real data, an initial value for the parameters to estimate, i.e. an initial guess given by the expert is used to set the CFL condition (5.5).

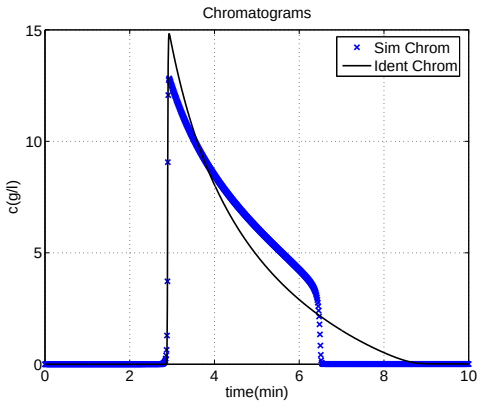
Using expert knowledge The choice of the type of isotherm function to be identified will be, in most cases, given by the chemists. Fig 5.1 illustrates the importance of this choice. In Fig 5.1-(a), the target chromatogram $\mathbf{c}_{exp}$ is computed using a Langmuir isotherm with one component ($m = 1$ and thus $n = 2$). In Fig 5.1-(b), the target chromatogram $\mathbf{c}_{exp}$ is computed using a Lattice of degree 3 with one component ($m = 1$ and thus $n = 4$). In both cases, the identification is done using a Langmuir model, with $n = 2$. It is clear from the figure that one is able to correctly identify the isotherm, and hence fit the "experimental" chromatogram when choosing the correct model (Fig 5.1 (a)) whereas the fit of the chromatogram is very poor when the model is not correct (Fig 5.1 (b)).

Another important issue when using CMA-ES is the initial choice for the covariance matrix: without any information, the algorithm starts with the identity matrix. However, this is a poor choice in case the different variables have very different possible order of magnitude, and the algorithm will spend some time adjusting its principal directions to those ranges.

In most cases of chromatographic identification, however, chemists provide orders of magnitudes, bounds and initial guesses for the different values of the unknown parameters.



(a) Simulation using a Langmuir isotherm, identification using a Langmuir model: the chromatogram is perfectly fit.



(b) Simulation using a Lattice isotherm, identification using a Langmuir model: poor fit of the chromatogram.

Figure 5.1: Importance of the choice of model (one component mixture)

Let $[(\mathbf{K}_a)_{min}, (\mathbf{K}_a)_{max}]$, $[(\mathbf{N}_b^*)_{min}, (\mathbf{N}_b^*)_{max}]$ and $[(\mathbf{E}_c)_{min}, (\mathbf{E}_c)_{max}]$ the ranges guessed by the chemists for respectively each $\mathbf{K}_a$, $\mathbf{N}_b^*$ and $\mathbf{E}_c$. All parameters are linearly scaled into those intervals from $[-1, 1]$, removing the need to modify the initial covariance matrix of CMA-ES.

Unfeasible solutions Two different situations can lead to *unfeasible* solutions:

First when one parameter at least, among parameters which have to be positive, becomes negative (remember that CMA-ES generates offspring using an unbounded normal distribution), the fitness function is arbitrarily set to 10^{20}.

Second when the CFL condition is violated, the simulation is numerically unstable, and generates absurd values. In this case, the simulation is stopped, and the fitness function is arbitrarily set to a value larger than 10^6. Note that a better solution would be to detect such violation before running the simulation, and to penalize the fitness by some amount that would be proportional to the actual violation. But it is numerically intractable to predict in advance if the CFL is going to be violated (see Eq. 5.5), and the numerical absurd values returned in case of numerical instability are not clearly correlated with the amount of violation either.

Initialization The initial mean $\langle \vec{x} \rangle_{\mathrm{W}}^{(0)}$ for CMA-ES is uniformly drawn in $[-1, 1]^n$, i.e., the parameters $\mathbf{K}_a$, $\mathbf{N}_b^*$ and $\mathbf{E}_c$ are uniformly drawn in the ranges given by the expert. The initial step-size σ_0 is set to 0.3. Besides we reject individuals of the population sampled outside the initial ranges. Unfeasible individuals are also rejected at initialization: at least one individual should be feasible to avoid random behavior of the algorithm. In both cases, rejection is done by resampling until a "good" individual is got or a maximal number of sampling individuals is reached. Initial numbers of offspring λ and parents μ

are set to the default values ($\lambda = \lfloor 4 + 3\log(\mathrm{n}) \rfloor$ and $\mu = \lfloor \lambda/2 \rfloor$).

Restarting and stopping criteria The algorithm stops if it reaches 5 restarts, or a given fitness value (typically a value between 10^{-9} and 10^{-15} for artificial problems, and adjusted for real data). Restart criteria (see Section 5.4.2) are RestartTolFun with TolFun= $10^{-12} \times \sigma^{(0)}$, RestartTolX with TolX= $10^{-12} \times \sigma^{(0)}$, RestartOnNoEffectAxis and RestartCondCov with a limit upper bound of 10^{14} for the condition number. The offspring size λ is doubled after each restart and μ is set equal to $\lfloor \lambda/2 \rfloor$.

5.5 Results

5.5.1 Validation using artificial data

A first series of validation runs was carried out using simulated chromatograms. Each identification uses one or many experimental chromatograms. Because the same discretization is used for both the identification and the generation of the "experimental" data, one solution is known (the same isotherm that was used to generate the data), and the best possible fitness is thus zero.

Several tests were run using different models for the isotherm, different numbers of components, and different numbers of time steps. In all cases, CMA-ES identified the correct parameters, *i.e.* the fitness function reaches values very close to zero. In most cases, CMA-ES did not need any restart to reach a precision of (10^{-14}), though this was necessary in a few cases. This happened when the whole population remained unfeasible during several generations, or when the algorithm was stuck in a local optimum. Figures 5.2, 5.3, 5.4 show typical evolutions during one run of the best fitness value with respect to the number of evaluations, for problems involving respectively 1, 2 or 3 components. Figure 5.4 is a case where restarting allowed the algorithm to escape a local optimum.

Specific tests were then run in order to study the influence of the expert guesses about both the ranges of the variables and the starting point of the algorithm possibly given by the chemical engineers: In CMA-ES, in a generation g, offspring are drawn from a Gaussian distribution centered on the mean $\langle \vec{x} \rangle_{\mathrm{W}}^{(g)}$. An expert guess for a good solution can hence be input as the mean of the first distribution $\langle \vec{x} \rangle_{\mathrm{W}}^{(0)}$ that will be used to generate the offspring of the first generation. The results are presented in Table 5.1. First 3 lines give the probabilities that a given run converges (i.e., reaches a fitness value of 10^{-12}), computed on 120 runs, and depending on the number of restarts (this probability of course increases with the number of restarts). The last line is the ratio between the average number of evaluations that were needed before convergence (averaged over the runs that did converge), and the probability of convergence: this ratio measures the performance of the different experimental settings, as discussed in details in [15].

The results displayed in Table 5.1 clearly demonstrate that a good guess of the range of the variables is the most prominent factor of success: even without any hint about the starting point, all runs did reach the required precision without any restart. However,

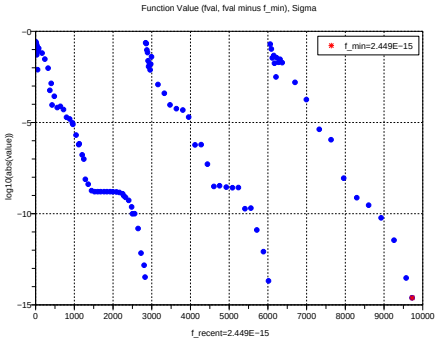


Figure 5.2: Single component mixture, 1000 time steps. Simulate a Lattice (5 parameters) and identify a Lattice of degree 4 (5 parameters): Best fitness versus number of evaluations. The first run gave a satisfactory solution but two restarts have been performed to reach a fitness value ($2.4\ 10^{-15}$) lower than 10^{-14}.

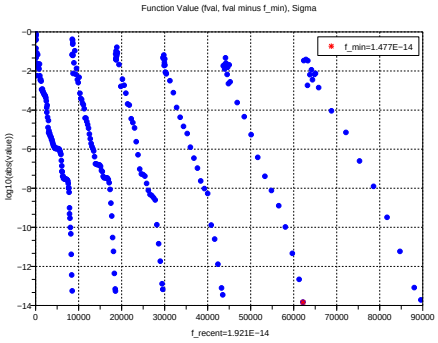


Figure 5.3: Binary component mixture, 500 time steps . Simulate a Langmuir (3 parameters) and identify a Lattice of degree 3 (10 parameters): Best fitness versus number of evaluations. The first run gave a satisfactory solution but the maximal number (here five) of restarts have been performed attempting to reach a fitness value of 10^{-14}, the best fitness value ($1.4\ 10^{-14}$) was reached in the fourth restart.

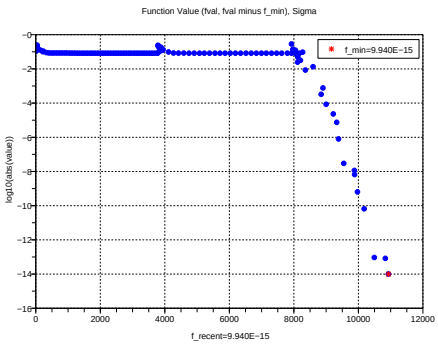


Figure 5.4: Ternary component mixture, 2000 time steps. Simulate a Langmuir (4 parameters) and identify a Langmuir (4 parameters): Best fitness versus number of evaluations. Two restarts were necessary: Before the second restart, CMA-ES is stuck in some local optima (fitness of order of 10^{-1}), in the second restart, the algorithm reaches a fitness value of 9.9 10^{-15}.

when no indication about the range is available, a good initial guess significantly improves the results, without reaching the perfect quality brought by tight bounds on the ranges: scaling is more important than rejecting unfeasible individuals at the beginning.

Computational cost The duration of an evaluation depends on the discretization of the numerical scheme (number of space- and time-steps), and on the number n of unknown parameters to identify. Several runs were precisely timed to assess the dependency of the computational cost on both factors. The simple Langmuir isotherm was used to both generate the data and identify the isotherm. Only computational costs of single evaluations are reported, as the number of evaluations per identification heavily depends on many parameters, including the possible expert guesses, and in any case is a random variable of unknown distribution. All runs in this paper were performed on a 1.8GHz Pentium computer running with a recent Linux system.

For one component ($m = 1$, $n = 2$), and 100, 500 and 1000 time steps, the averages of the durations of a single evaluation are respectively 0.0097, 0.22, and 0.9 seconds, fitting the theoretical quadratic increase with the number of time steps (though 3 sample points are too few to demonstrate anything!). This also holds for the number of space steps as the number of space steps is proportional to the number of time steps due to the CFL condition. For an identification with a 1-component Langmuir isotherm, the total cost of the identification is on average 540 seconds for a 1000 time steps discretization.

When looking at the dependency of the computational cost on the number of unknown parameters, things are not that clear from a theoretical point of view, because the cost of each computation of the isotherm function also depends on the number of components and on the number of experimental chromatograms to compare with. Experimentally, for, 2, 3 and 4 variables, the costs of a single evaluation are respectively 0.9, 1.04, and 2.2 seconds (for a 1000 time steps discretization). For an identification, the total time is roughly 15 to 25 minutes for 2 variables, 40 to 60 minutes for 3 variables, and 1 to 2

Table 5.1: On the usefulness of Expert Knowledge: target values for Langmuir isotherm are here $(\mathbf{K}_1, \mathbf{N}^*) = (0.0388, 107)$. Expert range is $[0.01, 0.05] \times [50, 150]$, wide range is $[0.001, 1] \times [50, 150]$. The expert guess for the starting point is a better initial mean (according to fitness value) than random. The first 3 lines give the probabilities (computed over 120 runs) to reach a 10^{-12} fitness value within the given number of restarts. The last line is the ratio of the number of evaluations needed for convergence (averaged over the runs that did converge) by the probability of convergence after two restarts (line 3).

Range	Expert range	Wide range	Wide range
Starting point	No guess	No guess	Expert guess
no restart	1	0.84	0.95
1 restart	1	0.92	0.97
2 restarts	1	0.95	0.97
Perf.	601	1015	905

hours for 4 variables.

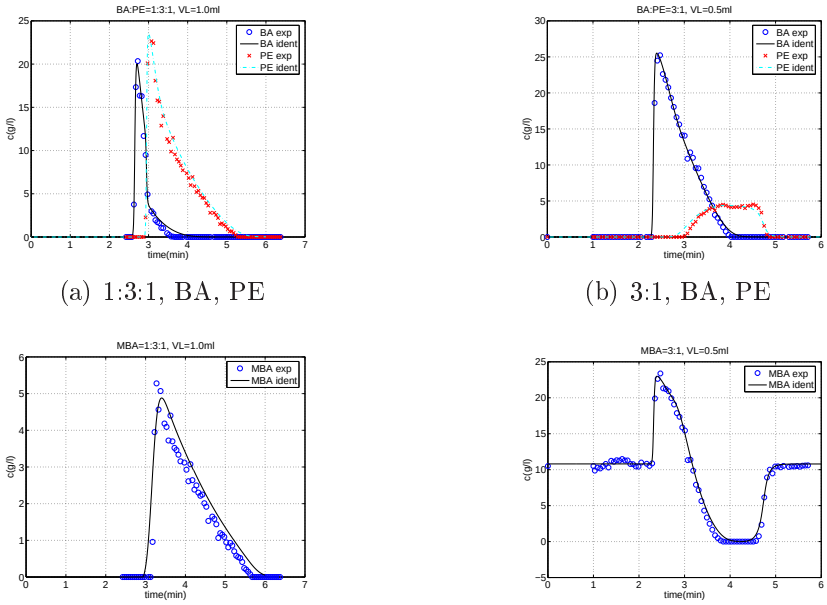
5.5.2 Experiments on real data

The CMA-ES based approach has also been tested on a set of data taken from [66]. The mixture was composed of 3 chemical species: the benzylalcohol (BA), the 2-phenylethanol (PE) and the 2-methylbenzylalcohol (MBA). Two real experiments have been performed with different proportions of injected mixtures, with respective proportions (1,3,1) and (3,1,0). Consequently, two real chromatograms have been provided. For this identification, Quiñones *et a.l.* [66] have used a *modified Langmuir* isotherm model in which each species has a different saturation coefficient $\mathbf{N}_i^*$:

$$\mathbf{H}_i(c) = \frac{\mathbf{N}_i^*}{1 + \sum_{l=1}^{3} \mathbf{K}_l \, c_l} \mathbf{K}_i \, \mathbf{c}_i, \;\; i = 1, \ldots, 3. \tag{5.15}$$

Six parameters are to be identified: $\mathbf{N}_i^*$ and $\mathbf{K}_i$, for $i = 1, \ldots, 3$. A change of variable has been made for those tests so that the unknown parameters are in fact $\mathbf{N}_i^*$ and $\mathbf{K}_i^{'}$, where $\mathbf{K}_i^{'} = \mathbf{K}_i \, \mathbf{N}_i$: those are the values that chemical engineers are able to experimentally measure.

Two series of numerical tests have been performed using a gradient-based method [73]: identification of the whole set of 6 parameters, and identification of the 3 saturation coefficients $\mathbf{N}_i^*$ only, after setting the Langmuir coefficients to the experimentally measured values $(\mathbf{K}_1^{'}, \mathbf{K}_2^{'}, \mathbf{K}_3^{'}) = (1.833, 3.108, 3.511)$. The initial ranges used for CMA-ES are $[60, 250] \times [60, 250] \times [60, 250]$ (resp. $[1.5, 2.5] \times [2.7, 3.7] \times [3, 4] \times [90, 200] \times [100, 200] \times [100, 210]$) when optimizing 3 parameters (resp. 6 parameters). Comparisons between the two experimental chromatograms and those resulting from CMA-ES identification for the two experiments are shown in Figure 5.5, for the 6-parameters case. The corresponding plots in the 3-parameters case are visually identical though the fitness value is slightly



(a) 1:3:1, BA, PE

(b) 3:1, BA, PE

(c) 1:3:1, MBA

(d) 3:1, MBA

Figure 5.5: Experimental chromatograms (markers) and identified chromatograms (continuous line) for the BA, BE and MBA species. Plots on the left/right correspond to an injection with proportions (1,3,1)/(3,1,0).

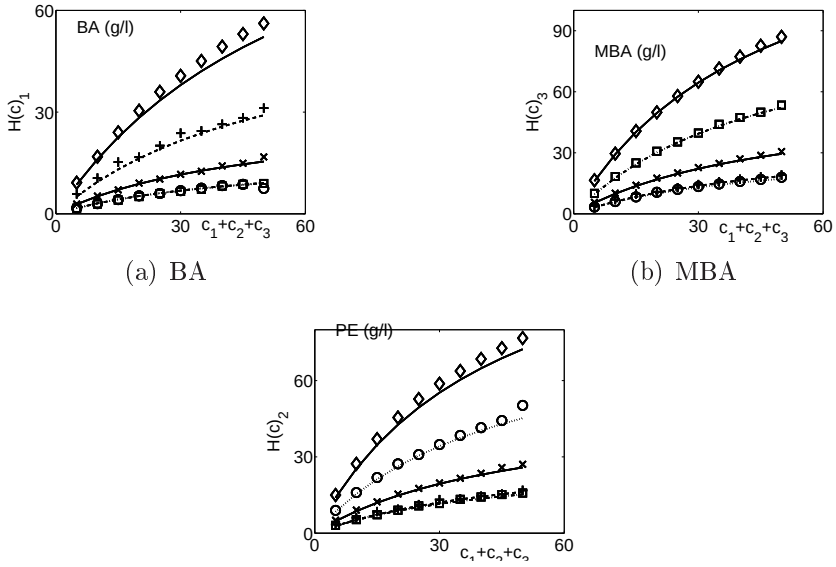
lower than in the 6-parameters case (see Tables 5.2 and 5.3). But another point of view on the results is given by the comparison between the identified isotherms and the (few) experimental values gathered by the chemical engineers. The usual way to present those isotherms in chemical publications is that of Figure 5.6: the absorbed quantity $\mathbf{H}(\mathbf{c})_i$ of each component $i = 1, 2, 3$ is displayed as a function of the total amount of mixture $(\mathbf{c}_1 + \mathbf{c}_2 + \mathbf{c}_3)$, for five different compositions of the mixture [73]. Identified (resp. experimental) isotherms are plotted in Figure 5.6 using continuous lines (resp. discrete markers), for the 6-parameters case. Here again the corresponding plots for the 3-parameters case are visually identical.

5.5.3 Comparison with a Gradient Method

CMA-ES results have then been compared with those of the gradient method from [73], using the same data case of ternary mixture taken from [66] and described in previous Section. Chromatograms found by CMA-ES are, according to the fitness (see Tables 5.2 and 5.3), closer to the experimental ones than those obtained with the gradient method. Moreover, contrary to the gradient algorithm, all 12 independent runs of CMA-ES converged to the same point. Thus, no variance is to be reported on Tables 5.2 and 5.3. Furthermore, there seems to be no need, when using CMA-ES, to fix the 3 Langmuir coefficients in order to find good results: when optimizing all 6 parameters, the gradient approach could not reach a value smaller than 0.01, whereas the best fitness found by CMA-ES in the same context is $8.32\,10^{-3}$ (Table 5.3).

Table 5.2: Comparing CMA-ES and gradient: the 3-parameters case. Solution (line 1) and associated fitness values (line 2) for the modified Langmuir model (Eq. 5.15). Line 3: For CMA-ES, "median (minimal)" number of fitness evaluations (out of 12 runs) needed to reach the corresending fitness value on line 2. For gradient, "number of fitness evaluations – number of gradient evaluations" for the best of the 10 runs described in [73].

	CMA-ES		Gradient
$\mathbf{N}_i^*$	(120.951,135.319,165.593)		(123.373,135.704,159.637)
Fitness $\times 10^3$	8.96	**8.78**	8.96
# Fit evals.	175 (70)	280 (203)	140 – 21



(a) BA (b) MBA

(c) PE

Figure 5.6: Isotherms associated to parameters values of Table 5.3 (continuous line) and experimental ones (markers) versus total amount of the mixture for different proportions of the component in the injected concentration [73].

Table 5.3: Comparing CMA-ES and gradient: the 6-parameters case. Solutions (lines 1 and 2) and associated fitness values (line 3) for the modified Langmuir model (Eq. 5.15).

	CMA-ES	Gradient
$K_i^{'}$	(1.861,3.120,3.563)	(1.780,3.009,3.470)
N_i^*	(118.732,134.860,162.498)	(129.986,141.07,168.495)
Fitness $\times 10^3$	**8.32**	10.7

Finally, when comparing the identified isotherms to the experimental ones (figure 5.6), the fit is clearly not very satisfying (similar deceptive results were obtained with the gradient method in [73]): Fitting both the isotherms and the chromatograms seem to be contradictory objectives. Two directions can lead to some improvements in this respect: modify the cost function $\mathcal{J}$ in order to take into account some least-square error on the isotherm as well as on the chromatograms; or use a multi-objective approach. Both modifications are easy to implement using Evolutionary Algorithms (a multi-objective version of CMA-ES was recently proposed [67]), while there are beyond what gradient-based methods can tackle. However, it might also be a sign that the modified Langmuir model that has been suggested for the isotherm function is not the correct one.

Comparison of convergence speeds Tables 5.2 and 5.3 also give an idea of the respective computational costs of both methods on the same real data. For the best run out of 10, the gradient algorithm reached its best fitness value after 21 iterations, requiring on average 7 evaluations per iteration for the embedded line search. Moreover, the computation of the gradient itself is costly – roughly estimated to 4 times that of the fitness function. Hence, the total cost of the gradient algorithm can be considered to be larger than 220 fitness evaluations. To reach the same fitness value ($8.96\,10^{-3}$), CMA-ES only needed 175 fitness evaluations (median value out of 12 runs). To converge to its best value ($8.78\,10^{-3}$, best run out of 12) CMA-ES needed 280 fitness evaluations. Those results show that the best run of the gradient algorithms needs roughly the same amount of functions evaluations than CMA-ES to converge. Regarding the robustness issue, note that CMA-ES always reached the same fitness value, while the 10 different runs of the gradient algorithm from 10 different starting points gave 10 different solutions: in order to assess the quality of the solution, more runs are needed for the gradient method than for CMA-ES!

5.6 Conclusions

This paper has introduced the use of CMA-ES for the parametric identification of isotherm functions in chromatography. Validation tests on simulated data were useful to adjust the (few) CMA-ES parameters, but also demonstrated the importance of expert knowledge: choice of the type of isotherm, ranges for the different parameters, and possibly some initial guess of a not-so-bad solution.

The proposed approach was also applied on real data and compared to previous work using gradient methods. On this data set, the best fitness found by CMA-ES is better than that found by the gradient approach. Moreover, the results obtained with CMA-ES are far more robust: (1) CMA-ES always converges to the same values of the isotherm parameters, independently of its starting point; (2) CMA-ES can handle the full problem that the gradient method failed to efficiently solve: there is no need when using CMA-ES to use experimental values of the Langmuir parameters in order to obtain a satisfactory fitness value. Note that the fitness function only takes into account the fit of the chromatograms, resulting in a poor fit on the isotherms. The results confirm the ones obtained with a

gradient approach, and suggest to either incorporate some measure of isotherm fit in the fitness, or to try some multi-objective method – probably the best way to go, as both objectives (chromatogram and isotherm fits) seem somehow contradictory.

Acknowledgments

This work was supported in part by MESR-CNRS ACI NIM Chromalgema. The authors would like to thank Nikolaus Hansen for the Scilab version of CMA-ES, and for his numerous useful comments.

Summary and Conclusion

The context of this thesis is the non linear continuous optimization using Evolution Strategies (ES). The work is composed of two parts. The first part is a theoretical and numerical study of the optimization using ES. In particular, we focus on the optimization of noisy objective functions which are frequently encountered in practice. In the second part, the state-of-the-art ES, the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) is applied to solve an identification problem relative to the chromatography technique used by chemical engineers.

Theoretical and numerical study

The study in Chapter 2 of the $(1 + 1)$-ES generalizes previous results relative to the behavior of the $(1, \lambda)$-ES [17]: The optimal convergence rate of ES is reached when the adaptation rule of the step-size is the artificial scale-invariant adaptation rule and the objective function is the spherical function. Therefore, these optimal settings (scale-invariant + spherical functions) can be used to assess the performances of algorithms using realistic adaptation rules and optimizing real world objective functions by comparing their performances with the optimal one.

In our study, we mainly investigate the rigorous proof with the numerical illustration of the convergence and divergence of scale-invariant ES. In Chapter 2, the Law of Large Numbers (LLN) for orthogonal random variables has been used to show the log-linear convergence of the scale-invariant $(1 + 1)$-ES when minimizing spherical functions. In Chapter 3, the Borel-Cantelli Lemma was used to show the almost sure convergence (or divergence) of the scale-invariant $(1 + 1)$-ES when minimizing noisy spherical functions. Then, in the same chapter, we used the LLN for Markov chains to rigorously derive the expressions of convergence (or divergence) rates of the algorithm. However, in order to obtain the log-linear behavior of the algorithm, one has to show that the convergence (or divergence) rates are not equal to 0. Though it is difficult to have a theoretical estimation of this convergence rate, our study shows that the expressions of the convergence (or divergence) rates derived can be estimated using Monte Carlo simulations. Therefore, one can show numerically that the convergence (or divergence) rates are not equal to 0. For the scale-invariant $(1, \lambda)$-ES minimizing noisy spherical functions, the LLN for orthogonal random variables is used again in Chapter 4 to show the log-linear of behavior of the algorithm. In the same chapter, numerical simulations have investigated the convergence (or divergence) rate that was theoretically derived to distiguish convergence and divergence cases. Moreover, it is theoretically proven (Chapter 4) that the convergence rate is

asymptotically (in the search space dimension) linear as a function of the inverse of the search space dimension. Note that for rank-based algorithms [137] or any Hit-and-Run direct search method [75], we know that the convergence rate is asymptotically linear as a function of the inverse of the search space dimension.

The convergence results obtained in Chapter 2 (Theorem 2.10) and in Chapter 4 (Theorem 4.8) were obtained using the LLN for orthogonal random variables. Note that the same results can be obtained using LLN for independent random variables.

Optimization of noisy objective functions When objective functions are noisy, ES had been shown to be more robust than other optimization methods in previous empirical studies [9, 106]. As pointed in [24], the difficulty when handling noisy objective functions arises for high noise levels. If the noise level is high, relatively to the ideal objective function value, the selection process can be deceived and therefore the performance of the algorithm is altered. This may lead to a non convergence of the method. Therefore, we investigated a multiplicative noise model for which the random noise is the ratio between the noisy objective function value and the ideal one. We investigated both the scale-invariant plus and comma strategies:

1. For the $(1+1)$-ES (Chapter 3), the only relevant fact is whether the noisy function can take negative fitness values or not. If a negative fitness value can happen, the scale-invariant $(1+1)$-ES will diverge, because of the elitist selection. This result may appear in contradiction with the result that has been previously derived in [8], stating that the algorithm is expected to converge, because of its positive expected progress rate. The point is that, in the numerical simulations investigated in that paper, negative fitness values were never sampled because they had a very small probability to occur. This was due to the use of normalizations of the noise strength with respect to the search space dimension. This also shows that numerical simulations have to be considered with care, and that both theoretical and numerical approaches have to be investigated in a complementary approach.

2. For the $(1, \lambda)$-ES (Chapter 4), the conclusions are different. The $(1, \lambda)$-ES can converge even in the case where negative fitness values can happen, provided that the variance of the noise (the noise strength) is sufficiently small. On the other hand, if the noise strength is sufficiently high, divergence occurs. In the specific case of Gaussian noise, the distinction between convergence and divergence cases was theoretically (respectively numerically) shown for infinite (respectively finite) search space dimension. For infinite dimension, similar results had been obtained using the limit of the normalized progress rate [25], which is equal to the opposite of the limit of the normalized convergence rate derived in our study. Moreover, for 'large' noise strength values where divergence occurs, convergence can nevertheless be obtained by increasing the number of offspring λ, and/or reevaluating each offspring several times and setting its fitness value to the average of these reevaluations. These solutions had been previously proposed in [25], and are also discussed in Chapter 4.

Elitist strategies and comparing ES in noisy environments The results of Chapter 3 show that, if negative objective functions values have a strictly postive probability to happen, then the scale-invariant $(1+1)$-ES cannot converge becasue of the elitist selection. Therefore, the non convergence holds also even if the number of the offspring is increased, i.e., even when using a $(1+\lambda)$-ES with $\lambda > 1$. It is worth noticing that the non robustness of the elitist selection have been already noticed in previous studies [119] (where the objective functions is not noisy), where it had been shown that the $(1+1)$-ES using the 1/5-success rule can get stuck in a local optimum. To overcome the non convergence of the $(1+1)$-ES (when minimizing noisy objective functions) shown in Chapter 3, a possible solution is to reevaluate the parent at each selection step. Therefore, the objective functions values of the solutions generated by the algorithm are no more decreasing. Another solution is to use the $(1,\lambda)$-ES which has been analyzed in Chapter 4 using the LLN for orthogonal random variables. The study of the scale-invariant $(1+\lambda)$-ES with reevaluation of the parent has not been investigated here but it can be done, similarly to the $(1,\lambda)$-ES, using the LLN for orthogonal random variables. Moreover, the $(1+\lambda)$-ES with reevaluation is similar, for infinite dimension, to a $(1,\lambda+1)$-ES as suggested by the limits of the normalized progress rates derived in [25]. Note that our study does not include the comparison of the performances of plus and comma ES in noisy environments. However, our study gives a guideline for practitionners about which strategy to use when some qualitative or quantitative informations on the noise distribution are available. If the noise is such that negative objective function values can happen one should not use plus strategies with no reevaluation of the parent. In this case, comma strategies (and probably plus strategies with reevaluation, relying on results in [25]) can be used with the possible solutions of reevaluating offspring or increasing their number if the noise level is 'high'.

In a previous study that compared the performances of ES in the presence of a Gaussian noise [7], it had been shown that, for small values of the noise strength, the plus strategies (with or without reevaluation) perform better than the comma strategies, and that the opposite happens for large normalized noise strength values. However, according to our study and from a theoretical view point, plus strategies with no reevaluation should not be used in the case of Gaussian noise as they lead to a non convergence of the algorithm. Therefore one has to investigate, in case of (theoretical) convergence, the comparison of the convergence rates of the $(1+\lambda)$-ES with reevaluation of the parent, and of the $(1,(\lambda+1))$-ES. Note that in practice, when the noise is Gaussian with a sufficiently small noise strength, the study of Chapter 3 shows that convergence can be seen in numerical simulations as the event leading to the non-convergence of the algorithm requires a huge number of iterations which is not the case of almost all numerical simulations. In these cases, and if one knows that (ideal) objective functions have to be positive, the $(1+1)$-ES can be used as a fast strategy (as suggested by the study in [7]) until a negative fitness value is sampled or another stopping criteria is met.
Finally, ES with recombination has to be theoretically and numerically investigated and compared with the other strategies. Another point that should be investigated, in noisy environments, is the behavior of ES using actual adaptation techniques (e.g. SA-ES and, of course, CMA-ES).

On the use of infinite dimension approximations and link with the progress rate theory The limit of the (normalized) convergence rate (or normalized progress rate) of an ES has in general, a simpler expression than that relative to a fixed dimension. This makes the distinction of convergence and divergence cases easier and the results obtained when the search space dimension goes to infinity can be considered to be reliable for sufficiently large dimensions.

In Chapter 4, we also extend a result from [17] to the noisy case: when optimizing spherical functions, the normalized progress rate, which is related to the convergence in mean of an ES, and (the opposite of) the normalized convergence rate, which gives the almost sure convergence, have the same limit when d goes to infinity.
On the other hand, for finite dimensions of the search space, Figure 4.5 and Figure 4.6 in Chapter 4 show that for some cases, divergence can hold in the case of infinite dimension while convergence holds for some finite dimensions. This confirms the observation that has already been done in the case of sphere function in [27]: The authors show that infinite dimension results do not cover all convergence cases for finite dimensions.

Moreover, our study shows rigorously (Chapter 4) the reliablity of an approximation for large dimensions that has been previously done in [8] when optimizing noisy objective functions. This approximation assumes that, for high dimensions of the search space, the parent and its offspring have the same noise level.However, the finite dimension plots of the convergence rates that are shown in Chapter 4, and especially in Figure 4.2, demonstrate that for the same noise variance and the same step-size mutation, the original model and the approximating one can have completely different behaviors (convergence for the former and divergence for the latter). Therefore, such approximations has to be taken with care.

Application

In Chapter 5, CMA-ES was applied to solve a real-world problem encountered in chemical engineering. This study confirms previous empirical comparison dealing whith the efficiency and the robustness of deterministic and randomized search methods. In this specific case study, CMA-ES is demonstrated to be more robust than a gradient based approach: CMA-ES found the same solution than the gradient method, but independently of the starting point, whereas gradient search is very sensitive to its initialization. In fact, the solutions proposed by CMA-ES were also slightly more accurate. But the most striking result is that CMA-ES succeeded to handle the full optimization problem whereas the gradient-based approach failed unless some parameters were fixed by the user to some experimentally determined values.

Bibliography

[1] C code of the standard pso 2006. http://www.particleswarm.info/Standard_PSO_2006.c.

[2] Comparison of evolutionary algorithms on a benchmarck function set. http://www.bionik.tu-berlin.de/user/niko/cec2005.html.

[3] E. Aarts and J. Korst. *Simulated Annealing and Boltzmann Machines: A stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley & Sons, Inc., New York, NY, USA, 1989.

[4] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in Order Statistics*. Wiley, 1992.

[5] D. V. Arnold. *Noisy Optimization with Evolution Strategies*. GENA. Kluwer Academic Publishers, 2002.

[6] D. V. Arnold and H.-G. Beyer. Efficiency and mutation strength adaptation of the $(\mu/\mu_{\mathrm{I}}, \lambda)$-ES in a noisy environment. In M. S. et al, editor, *Proceedings of Parallel Problem Solving from Nature - PPSN VI*, volume 1917 of *LNCS*, pages 39–48. Springer, 2000.

[7] D. V. Arnold and H.-G. Beyer. Investigation of the (μ, λ)-ES in the presence of noise. In *Proceedings of 2001 IEEE Congress on Evolutionary Computation*, pages 332–339. IEEE Press, 2001.

[8] D. V. Arnold and H.-G. Beyer. Local performance of the (1+1)-ES in a noisy environment. *IEEE Transactions on Evolutionary Computation*, 6(1):30–41, 2002.

[9] D. V. Arnold and H.-G. Beyer. A comparison of Evolution Strategies with other direct search methods in the presence of noise. *Computational Optimization and Applications*, 24:135–159, 2003.

[10] D. V. Arnold and H.-G. Beyer. A general noise model and its effects on evolution strategy performance. *IEEE Transactions on Evolutionary Computation*, 10(4):380–391, 2006.

[11] C. Audet and J. E. Dennis. Analysis of genaralized pattern searches. *SIAM Journal on Optimization*, 13(3):889–903, 2003.

[12] A. Auger. *Contributions théoriques et Numériques à l'optimisation continue par algorithmes évolutionnaires*. PhD thesis, Université Paris 6, 2004.

[13] A. Auger. Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible Markov chains. *Theoretical Computer Science*, 334:35–69, 2005.

[14] A. Auger, C. L. Bris, and M. Schoenauer. Rigorous analysis of some simple adaptive es. Technical Report RR-4914, INRIA, 2003.

[15] A. Auger and N. Hansen. Performance evaluation of an advanced local search evolutionary algorithm. In *Proc. IEEE Congress On Evolutionary Computation*, pages 1777–1784, 2005.

[16] A. Auger and N. Hansen. A restart CMA evolution strategy with increasing population size. In *Proc. IEEE Congress On Evolutionary Computation*, pages 1769–1776, 2005.

[17] A. Auger and N. Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In A. Press, editor, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, pages 445–452, 2006.

[18] A. Auger and N. Hansen. Tutorial (PPSN'2008) - Evolution Strategies and Related Estimation of Distribution Algorithms, September 2008. http://www.bionik.tu-berlin.de/user/niko/ppsn2008tutorial.pdf.

[19] A. Auger, M. Jebalia, and O. Teytaud. XSE: quasi-random mutations for evolution strategies. In *Proceedings of Evolutionary Algorithms, 12 pages*, 2005.

[20] A. Auger and O. Teytaud. Continuous lunches are free plus the design of optimal optimization algorithms. *Algorithmica*, 2009. (accepted).

[21] S. Baluja and R. Caruana. Removing the genetics from the standard genetic algorithms. In A. Prieditis and S. Russel, editors, *ICML95*, pages 38–46. Morgan Kaufmann, 1995.

[22] A. Ben Haj-Yedder, A. Auger, C. M. Dion, E. Cancès, A. Keller, C. Le Bris, and O. Atabek. Numerical optimization of laser fields to control molecular orientation. *Physical Review A*, 66(6):063401, Dec 2002.

[23] H.-G. Beyer. Toward a theory of evolution strategies: Some asymptotical results from the (1,+ λ)-theory. *Evol. Comput.*, 1(2):165–188, 1993.

[24] H.-G. Beyer. Evolutionary algorithms in noisy environments: Theoretical issues and guidelines for practice. *Computer Methods in Applied Mechanics and Engineering*, 186(2-4):239–267, 2000.

[25] H.-G. Beyer. *The Theory of Evolution Strategies*. Natural Computing Series. Springer-Verlag, 2001.

[26] H.-G. Beyer and K. Deb. On the desired behaviors of self-adaptive evolutionary algorithms. In M. S. et al, editor, *PPSN2000*, pages 59–68. Springer Verlag LNCS 1917, 2000.

[27] A. Bienvenüe and O. François. Global convergence for evolution strategies in spherical problems: some simple proofs and difficulties. *Theor. Comput. Sci.*, 306(1-3):269–289, 2003.

[28] D. Boeringer and D. Werner. Efficiency-constrained particle swarm optimization of a modified bernstein polynomial for conformal array excitation amplitude synthesis. *IEEE transactions on antennas and propagation*, 53:2662–2673, 2005.

[29] J. Bonnans, J. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical Optimization – Theoretical and Practical Aspects*. Universitext. Springer Verlag, Berlin, 2006.

[30] P. A. N. Bosman and D. Thierens. Expanding from discrete to continuous estimation of distribution algorithms: The idea. In *PPSN VI: Proceedings of the 6th International Conference on Parallel Problem Solving from Nature*, LNCS, pages 767–776. Springer Verlag, 2000.

[31] S. H. Brooks. A discussion of random methods for seeking maxima. *Operations Research*, 6:244–251, 1958.

[32] A. Chatterjee, K. Pulasinghe, K. Watanabe, and K. Izumi. A particle-swarm-optimized fuzzy-neural network for voice-controlled robot systems. *Industrial Electronics, IEEE Transactions on*, 52(6):1478–1489, 2005.

[33] G. Chonghui and T. Huanwen. Global convergence properties of evolution strategies. *Chinese Journal of Numerical Mathematics and Applications*, pages 78–84, 2001.

[34] M. Clerc and J. Kennedy. The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *Evolutionary Computation, IEEE Transactions on*, 6(1):58–73, February 2002.

[35] X. Cui, T. Potok, and P. Palathingal. Document clustering using particle swarm optimization. In *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE*, pages 185–191. IEE, 2005.

[36] J.-C. Culioli. *Introduction à l'optimisation*. Ellipses, 1994.

[37] V. D. and L. S. G. A monte carlo simulated annealing approach to optimization over continuous variables. *Journal of computational physics*, 56:259–271, 1984.

[38] K. DeJong. *Evolutionary Computation. A unified Approach*. MIT Press, 2006.

[39] J. Dennis and V. Torczon. Derivative-free pattern search methods for multidisciplinary design problems. In *Proccedings of the 5th AIAA/ USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, pages 922–932, 1994.

[40] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations (Classics in Applied Mathematics, 16)*. SIAM, 1996.

[41] A. Eiben and J. Smith. *Introduction to Evolutionary Computing*. Springer Verlag, 2003.

[42] A. E. Eiben, Z. Michalewicz, M. Schoenauer, and J. E. Smith. Parameter control in evolutionary algorithms. In F. Lobo, C. Lima, and Z. Michalewicz, editors, *Parameter Setting in Evolutionary Algorithms*, chapter 2, pages 19–46. Springer, 2007.

[43] A. Fadda and M. Schoenauer. Evolutionary chromatographic law identification by recurrent neural nets. In D. Fogel and W. Atmar, editors, *Proc. 3^{rd} Annual Conference on Evolutionary Programming*, pages 219–235. MIT PRESS, 1994.

[44] J. Fitzpatrick and J. Grefenstette. Genetic algorithms in noisy environments. *Machine Learning*, 3:101–120, 1988.

[45] L. J. Fogel, A. J. Owens, and M. J. Walsh. *Artificial Intelligence through Simulated Evolution*. New York: John Wiley, 1966.

[46] R. Gämperle, S. D. Müller, and P. Koumoutsakos. A parameter study for differential evolution. In *WSEAS Int. Conf. on Advances in Intelligent Systems, Fuzzy Systems, Evolutionary Computation*, pages 293–298. WSEAS Press, 2002.

[47] P. Gilmore and C. Kelley. An implicit filtering algorithm for optimization of functions with many local minima. *SIAM Journal on optimization*, 5:269–285, 1995.

[48] E. Godlewski and P.-A. Raviart. *Hyperbolic systems of conservation laws*, volume 3/4 of *Mathematiques et applications*. Ed. Ellipses, SMAI, 1991.

[49] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley, 1989.

[50] G. Guiochon, A. Feilinger, S. Golshan Shirazi, and A. Katti. *Fundamentals of preparative and nonlinear chromatography*. Academic Press, Boston, second edition, 2006.

[51] H. Hamda and M. Schoenauer. Adaptive techniques for evolutionary topological optimum design. In *In I. Parmee, editor, Evolutionary Design and Manufacture*, pages 123–136. Springer, 2000.

[52] N. Hansen. The CMA Evolution Strategy. http://www.bionik.tu-berlin.de/user/niko/cmaesintro.html.

[53] N. Hansen. References to CMA-ES applications. http://www.bionik.tu-berlin.de/user/niko/cmaapplications.pdf.

[54] N. Hansen. Invariance, self-adaptation and correlated mutations and evolution strategies. In M. S. et al, editor, *Proceedings of Parallel Problem Solving from Nature (PPSN VI)*, volume 1917 of *Lecture Notes in Computer Science*, pages 355–364. Springer, 2000.

[55] N. Hansen. The CMA evolution strategy: a comparing review. In J. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, editors, *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, pages 75–102. Springer, 2006.

[56] N. Hansen and S. Kern. Evaluating the CMA evolution strategy on multimodal test functions. In X. Yao et al., editors, *Parallel Problem Solving from Nature PPSN VIII*, volume 3242 of *LNCS*, pages 282–291. Springer, 2004.

[57] N. Hansen and S. Kern. Evaluating the CMA evolution strategy on multimodal test functions. In X. Yao et al., editors, *Parallel Problem Solving from Nature - PPSN VIII, LNCS 3242*, pages 282–291. Springer, 2004.

[58] N. Hansen and S. Kern. Evaluting the cma evolution strategy on multimodal test functions. In *PPSN'04*. Springer Verlag, 2004.

[59] N. Hansen, S. D. Müller, and P. Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation. *Evolutionary Computation*, 11(1):1–18, 2003.

[60] N. Hansen and A. Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *ICEC96*, pages 312–317. IEEE Press, 1996.

[61] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[62] N. Hansen, R. Ros, N. Mauny, M. Schoenauer, and A. Auger. PSO Facing Non-Separable and Ill-Conditioned Problems. Technical Report RR-6447, INRIA, 2008.

[63] M. Herdy. Reproductive isolation as strategy parameter in hierarichally organized evolution strategies. In *PPSN*, pages 209–, 1992.

[64] J. H. Holland. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, 1975.

[65] R. Hooke and T. Jeeves. 'direct search' solution of numerical and statistical problems. *Journal of the ACM*, 8:212–229, 1961.

[66] J. F. I. Quiñones and G. Guiochon. High concentration band profiles and system peaks for a ternary solute system. *Anal. Chem*, pages 1495–1502, 2000.

[67] C. Igel, N. Hansen, and S. Roth. Covariance matrix adaptation for multi-objective optimization. *Evolutionary Computation*, 15(1):1–28, 2007.

[68] L. Ingber. Very fast simulated re-annealing. Lester Ingber Papers 89vf, Lester Ingber, 1989.

[69] L. Ingber. Adaptive simulated annealing (asa): Lessons learned. *Control and Cybernetics*, 25:33–54, 1996.

[70] J. Jägersküpper. Probabilistic runtime analysis of $(1 +/, \lambda)$,es using isotropic mutations. In *GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 461–468, New York, NY, USA, 2006. ACM.

[71] J. Jägersküpper. Algorithmic analysis of a basic evolutionary algorithm for continuous optimization. *Theoretical Computer Science*, 379(3):329–347, 2007.

[72] J. Jägersküpper and C. Witt. Rigorous runtime analysis of a $(\mu+1)$es for the sphere function. In *GECCO '05: Proceedings of the 2005 conference on Genetic and evolutionary computation*, pages 849–856, New York, NY, USA, 2005. ACM.

[73] F. James and M. Postel. Numerical gradient methods for flux identification in a system of conservation laws. *Journal of Engineering Mathematics*, 60:293–317, 2007.

[74] F. James, M. Sepúlveda, I. Q. nones, F. Charton, and G. Guiochon. Determination of binary competitive equilibrium isotherms from the individual chromatographic band profiles. *Chem. Eng. Sci.*, 54(11):1677–1696, 1999.

[75] J. Jägersküpper. Lower bounds for hit-and-run direct search. In Yao, Xin et al., editor, *Stochastic Algorithms: Foundations and Applications - SAGA 2007, LNCS 4665*, pages 118–129. Springer Berlin, Heidelberg, 2007.

[76] M. Jebalia and A. Auger. On multiplicative noise models for stochastic search. In G. Rudolph, T. Jansen, S. Lucas, C. Polini, and N. Beume, editors, *Proceedings of Parallel Problem Solving from Nature (PPSN X)*, volume 5199 of *Lecture Notes in Computer Science*, pages 52–61. Springer Verlag, 2008.

[77] M. Jebalia, A. Auger, and P. Liardet. Log-linear convergence and optimal bounds for the $(1 + 1)$-ES. In N. Monmarché and al., editors, *Proceedings of Evolution Artificielle (EA'07)*, volume 4926 of *LNCS*, pages 207–218. Springer, 2008.

[78] M. Jebalia, A. Auger, M. Schoenauer, F. James, and M. Postel. Identification of the isotherm function in chromatography using CMA-ES. In *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*, pages 4289–4296. IEEE, September 2007.

[79] Y. Jin and J. Branke. Evolutionary Optimization in Uncertain Environments-A Survey. *IEEE Transactions on Evolutionary Computation*, 9(3):303–317, June 2005.

[80] C. Kelley. *Iterative Methods for Optimization*. SIAM Frontiers in Applied Mathematics. SIAM, Philadelphia, USA, 1999.

[81] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, volume 4, pages 1942–1948, 1995.

[82] S. Kern, S. Müller, N. Hansen, D. Büche, J. Ocenasek, and P. Koumoutsakos. Learning Probability Distributions in Continuous Evolutionary Algorithms - A Comparative Review. *Natural Computing*, 3(1):77–112, 2004.

[83] J. Kiefer and J. Wolfowitz. Stochastic estimation of a regression function. *Annals of Mathematical statistics*, 23:462–466, 1952.

[84] S. Kirkpatrick, C. D. G. Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671 – 680, May 1983.

[85] T. G. Kolda, R. M. Lewis, and V. Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM REVIEW*, 45(3):385–482, 2003.

[86] H. J. Kushner and D. S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, volume 26 of *Applied Mathematical Sciences*. Springer-Verlag, 1978.

[87] H. J. Kushner and G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer Verlag, New York, 1997.

[88] J. Lampinen and I. Zelinka. On stagnation of the differential evolution algorithm. In *Proceedings of MENDEL 2000, 6th International Mendel Conference on Soft Computing*, pages 76–83, 2000.

[89] I. Langmuir. The adsorption of gases on plane surfaces of glass, mica and platinum. *Jour. Am. Chem. Soc.*, 40(9):1361–1403, 1918.

[90] P. Larranaga, J. A. Lozano, and E. Bengoetxea. Estimation of distribution algorithms based on multivariate normal and gaussian networks. Technical Report EHU-KZAA-IK-1-01, University of the Basque Country, 2001.

[91] P. Larrañaga and J. A. Lozano. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. GENA. Kluwer Academic Publishers, 2001.

[92] F. G. Lobo, C. F. Lima, and Z. Michalewicz, editors. *Parameter Setting in Evolutionary Algorithms*. Studies in Computational Intelligence. Springer Verlag, 2007.

[93] M. Loève. *Probability Theory*. Van Nostrand, 1955.

[94] Y. Ma, C. Jiang, Z. Hou, and C. Wang. The formulation of the optimal strategies for the electricity producers based on the particle swarm optimization algorithm. *Power Systems, IEEE Transactions on*, 21(4):1663 – 1671, November 2006.

[95] Y. Marinakis, M. Marinaki, and G. Dounias. Particle swarm optimization for pap-smear diagnosis. *Expert Syst. Appl.*, 35(4):1645–1656, 2008.

[96] K. McKinnon. Convergence of the nelder-mead simplex method to a nonstationary point. *SIAM Journal on Optimization*, 9(1):148–158, 1998.

[97] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, New York, 1993.

[98] E. Mezura-Montes, J. Velázquez-Reyes, and C. A. C. Coello. A comparative study of differential evolution variants for global optimization. In *GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 485–492, New York, NY, USA, 2006. ACM.

[99] Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer Verlag, New-York, 1992-1996. 1st-3rd edition.

[100] B. Miller and D. Goldberg. Genetic algorithms, selection schemes, and the varying effects of noise. *Evolutionary computation*, 4(2):113–131, 1997.

[101] M. Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, 1996.

[102] M.Locatelli. *Handbook of Global Optimization II*, chapter Simulated annealing algorithms for continuous global optimization, pages 179–230. Kluwer Academic Publishers, 2002.

[103] S. Mottelet. Optimisation non-linéaire, 2003. http://www.dma.utc.fr/polytex/cours.pdf.

[104] S. D. Müller. *Bio-inspired optimization algorithms for engineering applications*. PhD thesis, Technische Wissenschaften ETH Zürich, 2002. Diss., ETH, Nr. 14719, 2002.

[105] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, pages 308–313, 1965.

[106] V. Nissen and J. Propach. On the robustness of population-based versus point-based optimization in the presence of noise. *IEEE Trans. Evolutionary Computation*, 2(3):107–119, 1998.

[107] F. J. (PI). CHROMALGEMA, Numerical Resolution of the Inverse Problem for Chromatography using Evolutionary Algorithms and Adaptive Multiresolution. http://sourceforge.net/projects/chromalgema.

[108] M. Powell. Direct search algorithms for optimization calculations. *Acta Numerica*, 7:287–336, 1998.

[109] M. J. Powell. The newuoa software for unconstrained optimization without derivatives. Technical Report DAMTP 2004/NA05, CMS, University of Cambridge, Cambridge CB3 0WA, UK, November 2004.

[110] M. J. Powell. Developments of newuoa for unconstrained minimization without derivatives. Technical Report DAMTP 2007/NA05, CMS, University of Cambridge, Cambridge CB3 0WA, UK, June 2007.

[111] K. Price, R. Storn, and J. Lampinen. *Differential Evolution - A Practical Approach to Global Optimization*. Springer, 2005.

[112] M. Rattray and J. Shapiro. Noisy fitness evaluations in genetic algorithms and the dynamics of learning. In R. Belew and M. Vose, editors, *Foundations of Genetic Algorithms 4*, pages 117–139, 1997.

[113] I. Rechenberg. *Evolutionsstrategie*. Friedrich Frommann Verlag (Günther Holzboog KG), Stuttgart, 1973.

[114] I. Rechenberg. *Evolutionstrategie: Optimierung Technisher Systeme nach Prinzipien des Biologischen Evolution*. Fromman-Hozlboog Verlag, Stuttgart, 1973.

[115] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical statistics*, 29:400–407, 1951.

[116] R. Ros and N. Hansen. A simple modification in cma-es achieving linear time and space complexity. In G. Rudolph, T. Jansen, S. Lucas, C. Polini, and N. Beume, editors, *Proceedings of Parallel Problem Solving from Nature (PPSN X)*, volume 5199 of *Lecture Notes in Computer Science*, pages 296–305. Springer, 2008.

[117] G. Rudolph. *Convergence Properties of Evolutionary algorithms*. Verlag Dr. Kovac, Hamburg, 1997.

[118] G. Rudolph. Convergence rates of evolutionary algorithms for a class of convex objective functions. *Control and Cybernetics*, 26(3):375–390, 1997.

[119] G. U. Rudolph. Self-adaptive mutations may lead to premature convergence. *IEEE Transactions on Evolutionary Computation*, 5:410–414, 2001.

[120] M. Schoenauer and M. Sebag. Using Domain Knowledge in Evolutionary System Identification. In K. G. et al., editor, *Evolutionary Algorithms in Engineering and Computer Science*. John Wiley, 2002.

[121] M. Schumer and K. Steiglitz. Adaptive step size random search. *Automatic Control, IEEE Transactions on*, 13:270–276, 1968.

[122] H.-P. Schwefel. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*. Birkhaeuser, 1977.

[123] H.-P. Schwefel. *Numerical Optimization of Computer Models*. John Wiley & Sons, Inc., New York, NY, USA, 1981.

[124] H.-P. Schwefel. Collective phenomena in evolutionary systems. In P. Checkland and I. Kiss, editors, *Problems of Constancy and Change – The Complementarity of Systems Approaches to Complexity, Proc. 31st Annual Meeting*, volume 2, pages 1025–1033, Budapest, 1987. Int'l Soc. for General System Research.

[125] M. Sebag and A. Ducoulombier. Extending population-based incremental learning to continuous search spaces. In *PPSN V: Proceedings of the 5th International Conference on Parallel Problem Solving from Nature*, pages 418–427, London, UK, 1998. Springer-Verlag.

[126] Y. Shi and R. Eberhart. Modified particle swarm optimizer. In *The 1998 IEEE International Conference on Evolutionary Computation, ICEC'98*, pages 69–73, 1998.

[127] Y. Shi, R. Eberhart, E. Center, and I. Carmel. Empirical study of particle swarm optimization. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, volume 3, 1999.

[128] J. Spall, S. Hill, and D. Stark. Theoretical comparisons of evolutionary computation and other optimization approaches. In *Proccedings of the 1999 IEEE Congress on Evolutionary Computation*, pages 1398–1405, 1998.

[129] W. Spendley, G. Hext, and F. Himsworth. Sequential application of simplex designs in optimisation and evolutionary operation. *Technometrics*, 4:441–461, 1962.

[130] R. Storn and K. Price. Differential Evolution Homepage. http://www.icsi.berkeley.edu/%7Estorn/code.html.

[131] R. Storn and K. Price. Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces. Technical Report TR-95-012, International Computer Science Institute, March 1995.

[132] R. Storn and K. Price. Minimizing the real functions of the icec'96 contest by differential evolution. In *The 1996 IEEE International Conference on Evolutionary Computation, ICEC'96*, pages 842–844, May 1996.

[133] R. Storn and K. Price. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, December 1997.

[134] P. Suganthan, N. Hansen, J. Liang, K. Deb, Y. P. Chen, A. Auger, and S. Tiwari. Problem definitions and evaluation criteria for the CEC 2005 special session on real-parameter optimization. Technical report, Nanyang Technological University, Singapore and KanGAL Report Number 2005005 (Kanpur Genetic Algorithms Laboratory, IIT Kanpur), May 2005.

[135] P. D. Surry and N. J. Radcliffe. Real representations. In *Foundations of Genetic Algorithms 4*. Morgan Kaufmann, 1997.

[136] O. Teytaud and A. Auger. On the adaptation of the noise level for stochastic optimization. In *IEEE Congress on Evolutionary Computation - CEC 2007*. IEEE, 2007.

[137] O. Teytaud and S. Gelly. General lower bounds for evolutionary algorithms. In 10^{th} *International Conference on Parallel Problem Solving from Nature (PPSN 2006)*, 2006.

[138] V. Torczon. *Multi-directional-search: A direct search algorithm for parallel machines*. PhD thesis, Department of Mathematical Sciences, RICE University, Houston, 1989.

[139] V. Torczon. On the convergence of pattern search algorithms. *SIAM Journal on optimization*, 7(1):1–25, 1997.

[140] P. Valentin and G. Guiochon. Propagation of finite concentration in gas chromatography. *Separation Science*, 10:245–305, 1976.

[141] P. Valentin, F. James, and M. Sepúlveda. Statistical thermodynamics models for a multicomponent two-phases equilibrium isotherm. *Math. Models and Methods in Applied Science*, 1:1–29, 1997.

[142] N. Vanhaecke, C. Lisdat, B. TrsquoJampens, D. Comparat, A. Crubellier, and P. Pillet. Accurate asymptotic ground state potential curves of cs_2 from two-colour photoassociation. *The European Physical Journal D - Atomic, Molecular, Optical and Plasma Physics*, 28(3):351–360, 2004.

[143] M. Vose. *The Simple Genetic Algorithm: foundations and theory*. MIT Press, 1999.

[144] D. Winfield. Function minimization by interpolation in a data table. *J. Inst. Maths Applics*, 12:339–347, 1973.

[145] D. Wolpert and W. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1:67–82, 1997.

[146] M. Wright. Direct search methods: Once scorned, now respectable. *Numerical Analysis*, pages 191–208, 1995.

[147] X. Yao and Y. Liu. Fast evolution strategies. *Control and Cybernetics*, 26:467–496, 1997.

[148] Z. B. Zabinsky and R. L. Smith. Pure adaptive search in global optimization. *Mathematical Programming*, 53(3):323–338, 1992.

[149] A. Zhigljavsky and A. Zilinskas. *Stochastic global optimization*, volume 1 of *Springer Optimization and its applications*. Springer, 2008.

[150] A. A. Zhigljavsky. *Theory of Global Random search*. Kluwer Academic Publishers, 1991.