INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAX115

Thèse de doctorat

Linear convergence of evolution strategies with covariance matrix adaptation

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 16 décembre 2024, par

ARMAND GISSLER

Composition du Jury :

Gersende Fort Directrice de recherche, CNRS, Institut de mathématiques de Toulouse, France	Présidente du jury
Johannes Lengler Professeur, ETH Zurich, Suisse	Rapporteur
Sean Meyn Professeur, University of Florida, États-Unis	Rapporteur
Randal Douc Professeur, Télécom SudParis, France	Examinateur
Alain Durmus Professeur, École polytechnique, France	Examinateur
Tobias Glasmachers Professeur, Ruhr-Universität Bochum, Allemagne	Examinateur
Anne Auger Directrice de recherche, Inria, École polytechnique, France	Directrice de thèse
Nikolaus Hansen Directeur de recherche, Inria, École polytechnique, France	Invité

Linear convergence of evolution strategies with covariance matrix adaptation

Armand Gissler

16 décembre 2024

Thèse de doctorat

Préparée à l'École polytechnique
Sous la direction de Anne Auger et Nikolaus Hansen

Spécialité: Mathématiques appliquées

Jury:
Gersende Fort, Présidente du jury
Johannes Lengler, Rapporteur
Sean Meyn, Rapporteur
Randal Douc, Examinateur
Alain Durmus, Examinateur
Tobias Glasmachers, Examinateur

Abstract

Standing as the state-of-the-art algorithm among the evolution strategies, CMA-ES is a derivative-free optimization algorithm with many applications. However, the mathematical proof of its convergence remains an open problem for more than 20 years. The main goal of this thesis is therefore to bring theoretical guarantees of convergence of CMA-ES. More precisely, we prove that CMA-ES approaches the minimum of an ellipsoidal function at a geometric rate. Furthermore, we confirm the conjecture that the covariance matrix in CMA-ES approximates the inverse Hessian of a convex-quadratic function.

Our proof relies on the analysis of stochastic processes and is decomposed in several steps. Indeed, we define a stochastic process by normalizing the state variables of CMA-ES. This is inspired from previous works analyzing stepsize adaptive ES: the normalization of the mean variable (translated by the optimum) by the stepsize yields to the definition of a Markov chain, assuming that the objective function is scaling-invariant. Under additional assumptions the chain is geometrically ergodic and by limit theorems we find that the algorithm converges. For CMA-ES, we have to include the covariance matrix to obtain a Markov chain. First we define a normalization function R on the space of positive definite matrices, and by normalizing the mean by the stepsize and the function R applied to the covariance matrix, we aim to find a stationary process. With a normalized covariance matrix and possibly normalized evolution paths, this process forms a time-homogeneous Markov chain when the objective function is scaling-invariant. Proving that this normalized Markov chain converges to a stationary probability distribution is the key to our proof of convergence of CMA-ES and will occupy Chapters 2 and 4.

First, in Chapter 1, we give a methodology to establish the irreducibility, aperiodicity and topological properties of time-homogeneous Markov chains valued in manifolds and possibly with nonsmooth updates. These tools are the generalization of a previously developed analysis of nonlinear state-space models, which however only includes Euclidean state spaces and continuously differentiable update functions. By using results from the nonsmooth analysis and the theory of measures on topological manifolds, we were able to extend this work.

These preliminary results unlock a convergence proof of CMA-ES based on the stability analysis of underlying Markov chains, since the normalization of the covariance matrix transforms the state space of the algorithm into a manifold, and standard stepsize adaptations include nonsmooth updates. Chapter 2 explains how to use the fore-mentioned methodology and prove that the normalized Markov chain is an irreducible and aperiodic T-chain.

We then prove its ergodicity by means of a Foster-Lyapunov method. In Chapter 4 we derive a potential function for which a state-dependent drift condition holds outside of a compact. Since the chain is a T-chain, compact sets are small, and since it is irreducible and aperiodic, a geometric drift

condition outside of a small set proves the geometric ergodicity of the chain. Yet the complexity of the chain (several state variables including a normalized covariance matrix) imposes us to restrict our proof to ellipsoidal objective functions.

The final step of our convergence proof is shown in Chapter 5. We use an ergodic theorem and a Law of Large Numbers to deduce the linear convergence of CMA-ES. Moreover, we use the affine-invariance of the algorithm to find that the covariance matrix in CMA-ES learns second-order information and that the convergence rate is independent of which ellipsoidal objective function is minimized.

Keywords: CMA-ES, Linear convergence, Markov chains, Irreducibility, Ergodicity, Covariance matrix

Résumé

En tant que méthode à l'état de l'art parmi les stratégies d'évolution, CMA-ES un algorithme d'optimisation sans dérivées avec de nombreuses applications, mais dont la convergence est restée un problème ouvert depuis plus de 20 ans. Le but de cette thèse est d'apporter des garanties théoriques de convergence de CMA-ES. Ainsi, nous prouvons que CMA-ES approche le minimum de fonctions ellipsoïdes avec une erreur géométrique, et nous vérifions la conjecture de la matrice de covariance dans CMA-ES qui estime l'inverse de la Hessienne d'une fonction convexe-quadratique.

Notre démonstration s'appuie sur l'analyse de processus stochastiques et est établie en plusieurs étapes. En effet, nous définissons un processus par la normalisation des variables de CMA-ES. Cette approche a réussi à analyser des ES avec adaptation du pas : en normalisant la variable de moyenne (translaté par l'optimum) par le pas, cela forme une chaîne de Markov lorsque la fonction objective est invariante par changement d'échelle. Sous des hypothèses supplémentaires, cette chaîne est géométriquement ergodique, ce qui permet de prouver la convergence de l'algorithme. Pour CMA-ES, nous devons inclure la matrice de covariance. Nous introduisons d'abord une fonction de normalisation R sur l'espace des matrices définies positives, et la normalisation de la moyenne par le pas et la fonction R appliquée à la matrice de covariance, nous espérons obtenir un processus stationnaire. Avec une matrice de covariance normalisée et des chemins d'évolution normalisés, ce processus est une chaîne de Markov pour des fonctions objectives invariantes par changement d'échelle. La preuve de sa convergence vers une probabilité stationnaire est la clé de notre démonstration et occupera les chapitres 2 et 4.

Tout d'abord, nous donnons dans le chapitre 1 une méthode pour établir l'irréductibilité, l'apériodicité et des propriétés topologiques de chaînes de Markov homogènes, à valeurs dans des variétés et avec des mises à jour non lisses. Ceci est la généralisation d'une analyse de modèles d'espace d'état non-linéaires qui n'incluaient que des espaces d'états euclidiens et des fonctions de mise à jour continuement différentiables. En s'appuyant sur des résultats de l'analyse non-lisse et de la théorie de la mesure sur des variétés topologique, nous avons pu étendre ce travail.

Ces résultats préliminaires permettent une preuve de convergence de CMA-ES qui repose sur l'analyse de stabilité de chaînes de Markov sous-jacentes, puisque la normalisation de la matrice de covariance définit un espace d'état qui est une variété, et car les adaptations de pas standards peuvent être non-lisses. Le chapitre 2 explique comment utiliser la méthode mentionnée ci-dessus pour prouver que la chaîne normalisée est une T-chaîne irréductible et apériodique.

Nous démontrons ensuite son ergodicité en utilisant une méthode de Foster-Lyapunov. Dans le chapitre 4, nous trouvons une fonction de potentiel qui satisfait une condition de dérive en-dehors d'un compact. Puisque la chaîne est une T-chaîne, les compacts sont petits et puisqu'elle est irréductible et apériodique, une condition de dérive géométrique en-dehors d'un ensemble petit démontre l'ergodicité

géométrique. Cependant, la complexité de la chaîne (avec plusieurs variables dont une matrice de covariance normalisée) nous impose de restreindre notre preuve à des fonctions objectifs ellipsoïdales.

L'étape finale de notre démonstration est donnée dans le chapitre 5. Nous utilisons un théorème ergodique et une loi des grands nombres pour déduire la convergence linéaire de CMA-ES. De plus, nous utilisons l'invariance par transformations affines de l'algorithme pour établir que la matrice de covariance de CMA-ES apprend l'inverse de la Hessienne de fonctions convexe-quadratiques, et que le taux de convergence est indépendant de quelle fonction ellipsoïdale est minimisée.

Mots clés: CMA-ES, Convergence linéaire, Chaînes de Markov, Irréductibilité, Ergodicité, Matrice de covariance

Remerciements

Ces trois dernières années, j'ai été occupé par mes travaux de thèse qui sont présentés dans ce manuscrit. Durant cette aventure, j'ai été accompagné par de nombreuses personnes que je tenais à remercier ici.

Je remercie avant tout mes encadrants Anne et Niko. J'ai découvert au cours de mon stage de M2 le monde fascinant des *Evolution strategies* et j'ai vite été captivé par ces méthodes. Vous avez su nourrir mon curiosité mathématique avec ce sujet cependant très concret. J'ai du mal à imaginer une question qui serait plus appropriée pour moi.

Je remercie les rapporteurs Johannes et Sean pour le temps passé à la relecture attentive de ma thèse. Merci aussi à Alain, Gersende, Randal et Tobias, pour avoir accepté de faire partie du jury. En particulier, je souhaite remercier Alain pour sa précieuse aide pour préparer la suite.

Je suis aussi heureux d'avoir été membre de l'équipe RandOpt, dont les réunions hebdomadaires ont rythmé mes années de thèse. J'ai eu le plaisir de rencontrer des co-équipiers, actuels ou anciens : Alexandre, Cheikh, Dimo, Eugénie, Lorenzo, Mohamed, Nikita, Oskar, Paul, Shan, Tristan. Merci par ailleurs aux organisateurs et aux participants des différents séminaires, conférences, colloques auxquels j'ai pu assisté et participé, notamment des deux sessions *Theory of Randomized Optimization Heuristics* à Dagstuhl et du Colloque des Jeunes Probabilistes et Statisticiens à l'île d'Oléron.

J'ai cherché à faire une thèse d'abord grâce à mon goût pour les maths. Je n'ai aucun doute sur le rôle joué par les profs qui ont croisé mon chemin. En particulier, merci à Monsieur Lucas, c'est pendant l'année de MP que j'ai décidé finalement de continuer les maths. Big up aussi à la promo MPSI-MP dont je retiens d'excellents souvenirs. Bien entendu, je n'oublie pas mes amis cachanais : Antoine, Jessie, Luc, Raymond, Simon, qui ont continué de me supporter pendant mes années de thèse.

Avant ma thèse, j'ai eu un bel aperçu du monde de la recherche en maths, grâce à l'accueil par Lukasz à Édimbourg et Tim à Montréal. J'ai pu y découvrir des questions nouvelles et variées en mathématiques. Je veux aussi remercier Francis pour la confiance qui m'accorde pour ma prochaine aventure.

Le CMAP a été un environnement exceptionnel pendant ces trois dernières années. Il m'est impossible ne pas remercier Nasséra et les gestionnaires pour le travail qui est accompli pour faire du labo un lieu de travail agréable et fonctionnel. Merci aussi aux assistantes Inria, Marie, Julienne, Amandine, pour leur aide administrative. Je remercie également l'ensemble des chercheurs du CMAP qui participent à une émulation scientifique bonne de par les différents séminaires régulièrement proposées ou tout simplement les discussions mathématiques (et aussi non-mathématiques) au quotidien.

Je remercie également Lucas et Igor pour l'organisation du monitorat au département de maths

appliquées. Je crois avoir beaucoup appris en enseignant tout au long de ma thèse. J'en profite pour saluer Julien et Inna qui donnaient le cours LAB102 ainsi que les autres chargés de TD Fabien, Grégoire, Maxime et Jules, et les élèves du Bachelor qui étaient présents durant mes sessions.

J'adresse de chaleureux remerciements aux doctorants du CMAP : mes camarades du bureau 2008 – Adrien, Adriano, Ana, Emmanuel, Grégoire mais aussi à Guillaume et Pierre qui m'ont accueilli dans ce bureau au tout début de ma thèse ; aux co-organisateurs du séminaire de doctorants, notamment Grégoire, Leila et Mano qui y étaient très investis pendant deux ans, mais aussi Antoine, Matthias, Meggie, Nathan ; et à tous les doctorants avec qui j'ai pu échanger en salle café ou au détour d'un départ au Magnan, dont Alexandre, Arthur, Baptiste, Clément, Christoph, Constantin, Corentin, Jean, Loïc[2], Louis[3], Madeleine, Mahmoud, Manon, Marta, Mouad, Quentin, Théo, Thomas, Yoann, Yushan, Ward. J'ai des pensées particulières pour mes amis doctorants, Adam, Antoine, Benjamin, Guillaume, Jessie, Louis, Luce, Raphaël, Richard. Merci Oskar pour tous les moments passés dans ou en-dehors du CMAP, notamment pendant les quelques *courtes* pauses que j'ai beaucoup appréciées. Enfin je voudrais remercier Orso pour toute son amitié. Je crois avoir réussi la quête annexe que tu m'as confié et j'espère qu'on pourra profiter de la récompense.

Pour terminer sur une note plus personnelle, je suis reconnaissant de la proximité que j'ai avec mes frères et soeur, Gautier, Danaé et Béranger, ainsi qu'avec ma nièce Adeline et bientôt aussi une nouvelle personne dans la famille. Malgré le temps passé, ma mère reste toujours dans mes pensées.

Contents

Introduction **1**

1 Mathematical formulation of CMA-ES 2
1.1 Sampling and ranking of candidate solutions *2*
1.2 Update of the mean *3*
1.3 Update of the stepsize *3*
1.4 Update of the covariance matrix *4*
1.5 Summary of CMA-ES *4*
2 Related optimization methods 5
2.1 Derivative-free optimization algorithms *5*
2.2 Newton's methods *7*
3 Linear Convergence 8
4 Introduction to Markov chains 10
4.1 Kernels and Markov chains *10*
4.2 Irreducibility *11*
4.3 Periodicity, aperiodicity *11*
4.4 Petite sets, small sets, T-chains *12*
4.5 Recurrence, Harris recurrence *12*
4.6 Invariant measures, positivity *13*
4.7 Ergodicity *13*
4.8 Stability of nonlinear Markov chains via the analysis of a deterministic control model *14*
5 Convergence and theoretical guarantees of evolution strategies 15
5.1 Proving linear behavior of ES by analyzing a normalized Markov chain *17*
6 Methodology and overview 18
6.1 Chapter 1: On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds [52] *18*
6.2 Chapter 2: Irreducibility of nonsmooth state-space models with an application to CMA-ES [53] *19*
6.3 Chapter 3: Asymptotic estimation of a perturbed symmetric eigenproblem [51] . . *20*
6.4 Chapter 4: Geometric ergodicity of Markov chains underlying CMA-ES *21*
6.5 Chapter 5: Linear convergence of CMA-ES on ellipsoidal problems and learning of second-order information *22*

6.6 Appendix A: Evaluation of the impact of various modifications to CMA-ES that facilitate its theoretical analysis [49] *22*

CHAPTER 1
On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds **23**

1 Introduction 24
2 Main results 25
2.1 The model and assumptions *25*
2.2 Main results *32*
3 Applications 33
3.1 CMA-ES *33*
3.2 The step-size adaptive ES with nonsmooth update *34*
4 Proofs 35
4.1 Preliminary results *35*
4.2 Proofs of the main results: verifiable conditions for irreducibility and aperiodicity *40*
4.3 Proofs for the application to CMA-ES *46*
A Background on manifolds 48
B Clarke's generalized derivative of locally Lipschitz functions on manifolds 49
C Additional proofs 53

CHAPTER 2
Irreducibility of nonsmooth state-space models with an application to CMA-ES **57**

1 Introduction 58
2 Definition of Markov chains arising from a normalization of CMA-ES 60
2.1 Presentation of CMA-ES *61*
2.2 Assumptions *63*
2.3 Proving the stability of a normalized Markov chain leads to linear convergence .. *64*
3 Main Results I: Irreducibility, aperiodicity, and T-chain property of normalized Markov chains underlying the CMA-ES algorithm 69
4 Main results II: Extension of the analysis of nonlinear state-space models 71
4.1 Deterministic control model and sufficient conditions for irreducibility and aperiodicity *71*
4.2 Irreducibility and aperiodicity of a projected Markov chain *73*
4.3 Homeomorphic transformation of an irreducible aperiodic T-chain *76*
5 Proof of Theorem 2.1 77
5.1 Definition of normalized chains underlying CMA-ES following (2.25) *and satisfying* **H**1-**H**2 ..... *77*
5.2 Finding steadily attracting states *82*
5.3 Controllability condition *86*
5.4 Proof of Theorem 2.1 ..... *90*

6 Conclusion and perspectives . . . 91
A Proofs in Section 5.2 . . . 92
A.1 Proof of Lemma 2.8 . . . *92*
B Proofs in Section 5.3 . . . 94
B.1 Proof of Lemma 2.12 . . . *94*
B.2 Proof of Lemma 2.13 . . . *99*
B.3 Proof of Proposition 2.12 . . . *104*

CHAPTER 3
Asymptotic estimations of a perturbed symmetric eigenproblem **107**

1 Introduction . . . 108
2 Bounds on the eigenvalues of (P_m) . . . 109
3 Estimating the eigenvectors of (P_m) . . . 110
3.1 Rank-one perturbation . . . *110*
3.2 Sum of m rank-one matrices pertubation . . . *112*
3.3 Thightness . . . *113*

CHAPTER 4
Geometric ergodicity of Markov chains underlying CMA-ES **115**

1 Introduction . . . 116
1.1 Notations . . . *118*
2 Geometric ergodicity of a normalized chain underlying CMA-ES . . . 119
2.1 The CMA-ES algorithm and first assumptions . . . *119*
2.2 Objective function assumptions . . . *122*
2.3 Definition of a normalized chain underlying the CMA-ES algorithm . . . *123*
2.4 Main results . . . *125*
3 Proof of the main results . . . 129
3.1 Methodology to prove the geometric ergodicity of the normalized Markov chain . . . *129*
3.2 Proof of Theorem 4.3 . . . *132*
3.3 Preliminary definitions and results for the proof of Proposition 4.4 . . . *133*
3.4 Bounding the expected largest eigenvalue of the (normalized) covariance matrix . . . *147*
3.5 Bounding the expected (normalized) mean . . . *156*
3.6 Proof of Proposition 4.4 . . . *166*
4 Discussion . . . 178
A Technical results . . . 179
B Proof of Proposition 4.1 . . . 179
C Proofs in Section 3.3 . . . 181
C.1 Proof of Theorem 4.6 . . . *181*
C.2 Proof of Corollary 4.2 . . . *181*
C.3 Proof of Lemma 4.2 . . . *182*
C.4 Proof of Proposition 4.6 . . . *183*
C.5 Proof of Proposition 4.7 . . . *184*
C.6 Proof of Proposition 4.8 . . . *185*
C.7 Proof of Proposition 4.9 . . . *186*

C.8 Proof of Lemma 4.3 ..... *188*

CHAPTER 5
Linear convergence of CMA-ES on ellipsoidal problems and learning of second-order information **191**

1 Assumptions for the theoretical analysis of CMA-ES 192
1.1 Assumptions on the objective function *192*
1.2 Assumption on the sampling distribution *192*
1.3 Assumption on the weights *192*
1.4 Assumptions on the stepsize change *192*
2 Invariance to rotation and to affine transformation 193
3 Linear convergence of CMA-ES on ellipsoidal objective functions 194
3.1 Integrability with respect to the invariant probability measure *195*
3.2 Linear behavior *200*
3.3 Equal convergence rates for different ellipsoidal functions *202*
3.4 Positivity of the convergence rate *202*
4 Limit distribution of the covariance matrix 207
5 Learning of the inverse Hessian of convex-quadratic functions 209

Conclusion and discussion **213**

APPENDIX A
Evaluation of the impact of various modifications to CMA-ES that facilitate its theoretical analysis **217**

1 Introduction 218
2 Algorithm presentation 218
3 Implementation and experimental procedure 220
4 CPU Timings 220
5 Results 220
6 Conclusion / Discussion 221

APPENDIX B
Introduction en français **229**

1 Formulation mathématique de CMA-ES 230
1.1 Échantillonnage et classement des solutions candidates *231*
1.2 Mise à jour de la moyenne *231*
1.3 Mise à jour du pas *231*
1.4 Mise à jour de la matrice de covariance *232*
1.5 Résumé de CMA-ES *232*
2 D'autres méthodes d'optimisation 233
2.1 Algorithmes sans dérivées *233*
2.2 Méthodes de Newton *236*

3 Convergence linéaire 237
4 Introduction aux chaînes de Markov 238
4.1 Noyaux et chaînes de Markov 239
4.2 Irréductibilité 239
4.3 Périodicité, apériodicité 240
4.4 Ensembles petits et small, T-chaînes 240
4.5 Récurrence, récurrence Harris 241
4.6 Mesures invaiantes 241
4.7 Ergodicité 241
4.8 Stabilité de chaînes de Markov non-linéaires par l'analyse d'un modèle de contrôle déterministe 242
5 Convergence et garanties théoriques pour les ES 244
5.1 Prouver un comportement linéaire des ES en analysant une chaîne de Markov normalisée 245
6 Méthodologie et aperçu 246
6.1 Chapitre 1: Sur l'irréductibilité et la convergence d'une classe de modèles d'espaces d'états non-lisses sur des variétés [52] 247
6.2 Chapitre 2: Irréductibilité de modèles à espaces d'états non-lisses avec application à CMA-ES[53] 247
6.3 Chapitre 3: Estimation asymptotique d'un problème de vecteurs propres symétrique perturbé [51] 248
6.4 Chapitre 4: Ergodicité géométrique de chaîne de Markov sous-jacentes à CMA-ES 249
6.5 Chapitre 5: Convergence linéaire de CMA-ES sur des problèmes ellipsoïdaux et apprentissage d'informations de second ordre 250
6.6 Annexe A: Évaluation de l'impact de modifications variées de CMA-ES qui facilitent son analyse théorique [49] 251

Bibliography **252**

Index of Notations **265**

Introduction

1 Mathematical formulation of CMA-ES . . . 2
1.1 Sampling and ranking of candidate solutions . . . 2
1.2 Update of the mean . . . 3
1.3 Update of the stepsize . . . 3
1.4 Update of the covariance matrix . . . 4
1.5 Summary of CMA-ES . . . 4
2 Related optimization methods . . . 5
2.1 Derivative-free optimization algorithms . . . 5
2.2 Newton's methods . . . 7
3 Linear Convergence . . . 8
4 Introduction to Markov chains . . . 10
4.1 Kernels and Markov chains . . . 10
4.2 Irreducibility . . . 11
4.3 Periodicity, aperiodicity . . . 11
4.4 Petite sets, small sets, T-chains . . . 12
4.5 Recurrence, Harris recurrence . . . 12
4.6 Invariant measures, positivity . . . 13
4.7 Ergodicity . . . 13
4.8 Stability of nonlinear Markov chains via the analysis of a deterministic control model 14
5 Convergence and theoretical guarantees of evolution strategies . . . 15
5.1 Proving linear behavior of ES by analyzing a normalized Markov chain . . . 17
6 Methodology and overview . . . 18
6.1 Chapter 1: On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds [52] . . . 18
6.2 Chapter 2: Irreducibility of nonsmooth state-space models with an application to CMA-ES [53] . . . 19
6.3 Chapter 3: Asymptotic estimation of a perturbed symmetric eigenproblem [51] . . . 20
6.4 Chapter 4: Geometric ergodicity of Markov chains underlying CMA-ES . . . 21
6.5 Chapter 5: Linear convergence of CMA-ES on ellipsoidal problems and learning of second-order information . . . 22
6.6 Appendix A: Evaluation of the impact of various modifications to CMA-ES that facilitate its theoretical analysis [49] . . . 22

Mathematical optimization usually seeks to find a solution to a minimization problem:

$$\text{find } x^* \in \operatorname{Arg\,min} f \tag{P}$$

where $f\colon \Omega \to \mathbb{R}$ is a real-valued function defined on a domain Ω. Specifically, in this thesis we take $\Omega = \mathbb{R}^d$ where d is a positive integer. In this context, finding an exact solution to (P) is often challenging, and instead we aim to approximate the global minimum x^* of f with algorithms that produce a sequence of points that converges to x^*.

Among derivative-free[1] optimization algorithms, evolution strategies (ES) form a class of algorithms which have known their first developments in the '60s and the '70s [131, 124]. Inspired by evolutionary biology, these methods are based on the principle of survival of the fittest. The evolution strategy with covariance matrix adaptation (CMA-ES) stands as the state-of-the-art algorithm in this class, widely used such as in biology [129, 22] and medicine [121], energy [93, 125], machine learning [60, 80, 123] and hyperparameter optimization [103, 2, 118], computer vision [31, 122, 140], including applications to autonomous vehicles [104, 1], web pages [95] or video game levels [144, 137]. It approximates the solution x^* by a multivariate normal distribution, and adapts a covariance matrix to favor sampling directions based on the rankings of the f-values of the previous candidate solutions.

While the initial propositions of CMA-ES averaged the previous covariance matrix with a *rank-one update* [74, 75, 76], successive improvements introduced a *rank-mu update* with equal weights [73], then with positive recombination weights [72] and later an *active update* based on negative weights on the worse candidate solutions [91]. Empirical evidence [76, 73] suggests that the mean provided by CMA-ES converges to the solution of many optimization problems with high probability, including nonconvex, nonseparable, ill-conditioned and multimodal problems. Additionally, CMA-ES seems to learn second-order information, a property satisfied by quasi-Newton methods, remarkable for CMA-ES since it only relies on comparisons of function evaluations.

However, mathematical proofs supporting these observations are lacking. While several variants of ES have been successfully demonstrated to linearly converge to the global optimum on specific classes of objective functions [23, 39, 17, 4, 6, 54, 141], the linear convergence to the global optimum and the approximation of the inverse Hessian by CMA-ES remain open problems. This thesis aims to address these two questions.

1 Mathematical formulation of CMA-ES

The principle of CMA-ES is to approximate the minimum x^* of the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ by a multivariate normal distribution $\mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$ where $t \in \mathbb{N}$ denotes the current iteration, the vector $m_t \in \mathbb{R}^d$ the mean, the scalar $\sigma_t > 0$ the stepsize, and $\mathbf{C}_t \in \mathcal{S}_{++}^d$ the covariance matrix at iteration t. Besides, we use two paths $p_t^\sigma, p_t^c \in \mathbb{R}^d$ for the update of the stepsize and the covariance matrix, respectively. At iteration $t \in \mathbb{N}$, given $(m_t, \sigma_t, \mathbf{C}_t, p_t^\sigma, p_t^c) \in \mathbb{R}^d \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \times \mathbb{R}^d \times \mathbb{R}^d$, the update of the parameters of CMA-ES is described below. This thesis does not cover the active update of CMA-ES [91] which relies on negative weights to adapt the covariance matrix.

1.1 Sampling and ranking of candidate solutions

Firstly, a population of λ offspring $x_{t+1}^1, \dots, x_{t+1}^\lambda$ is sampled from the distribution $\mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$. In this section and throughout the entire manuscript, the integer $\lambda \geqslant 2$ will represent the population size used by CMA-ES. More precisely, we consider random variables

$$U_{t+1}^1, \dots, U_{t+1}^\lambda \sim \mathcal{N}(0, \mathbf{I}_d) \text{ i.i.d.}, \tag{1}$$

[1] do not rely on or assume the existence of derivatives

and define candidate solutions

$$x_{t+1}^i = m_t + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^i \quad \text{for } i = 1, \dots, \lambda. \tag{2}$$

For a positive definite matrix $\mathbf{C} \in \mathcal{S}_{++}^d$, its square root $\sqrt{\mathbf{C}}$ is defined as the unique matrix of $\mathcal{S}_{++}^d$ such that $\sqrt{\mathbf{C}}^2 = \mathbf{C}$. The next step is to rank the candidate solutions with respect to their f-values. Only here CMA-ES relies on the evaluation of the objective function f for this iteration. Formally, we define a permutation $s_{t+1} \in \mathfrak{S}_\lambda$ satisfying

$$f\left(x_{t+1}^{s_{t+1}(1)}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{s_{t+1}(\lambda)}\right). \tag{3}$$

The choice of the permutation is not unique when the f-values of two different solutions are equal. As a tie-break, we impose throughout the thesis that when $i < j$ and $f(x_{t+1}^i) = f(x_{t+1}^j)$, we have $s_{t+1}^{-1}(i) < s_{t+1}^{-1}(j)$.

1.2 *Update of the mean*

In order to update the mean, one can define a weighted average of the μ best offspring. The integer $\mu \in \{1, \dots, \lambda\}$ is called the parent number, and often chosen as $\mu \approx \lambda/2$. Besides, we fix weights $\mathbf{w}_m = (w_1^m, \dots, w_\mu^m)$ satisfying $w_1^m \geqslant \cdots \geqslant w_\mu^m > 0$ and $\sum_{i=1}^\mu w_i^m = 1$. Then, the mean obeys

$$m_{t+1} = \sum_{i=1}^{\mu} w_i^m x_{t+1}^{s_{t+1}(i)}. \tag{4}$$

However, throughout the manuscript, we consider a more general update for the mean [77]. We choose a learning rate $c_m > 0$, and update the mean by

$$m_{t+1} = (1 - c_m) m_t + c_m \sum_{i=1}^{\mu} w_i^m x_{t+1}^{s_{t+1}(i)} = m_t + c_m \sigma_t \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}. \tag{5}$$

We recover the update equation (4) by setting $c_m = 1$. This is in practice the default value for c_m.

1.3 *Update of the stepsize*

The stepsize update relies on the cumulation path p_t^σ. It aims to weight the previous average best directions and compare its length to its expectation under neutral selection, i.e., when we do not rank the offspring and define a permutation by (3). The update of the path is as follows:

$$p_{t+1}^\sigma = (1 - c_\sigma) p_t^\sigma + \sqrt{c_\sigma (2 - c_\sigma) \mu_{\text{eff}}} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}. \tag{6}$$

The positive number $c_\sigma \in (0, 1]$ is the decay rate for the path. When $c_\sigma = 1$, only the last iteration is used to update the stepsize, and we say that we have no cumulation on the stepsize. Besides, the quantity $\mu_{\text{eff}} > 0$ is defined as $\mu_{\text{eff}} = \|\mathbf{w}_m\|_2^2$. The stepsize is updated then using an abstract stepsize change function $\Gamma \colon \mathbb{R}^d \to \mathbb{R}_{++}$, i.e.,

$$\sigma_{t+1} = \sigma_t \times \Gamma(p_{t+1}^\sigma). \tag{7}$$

One usual choice for the stepsize change corresponds to the cumulative stepsize adaptation (CSA) [63]:

$$\Gamma_{\text{CSA}}^1 \colon p \in \mathbb{R}^d \mapsto \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|p\|}{\mathbb{E}\|\mathcal{N}(0, \mathbf{I}_d)\|} - 1\right)\right). \tag{8}$$

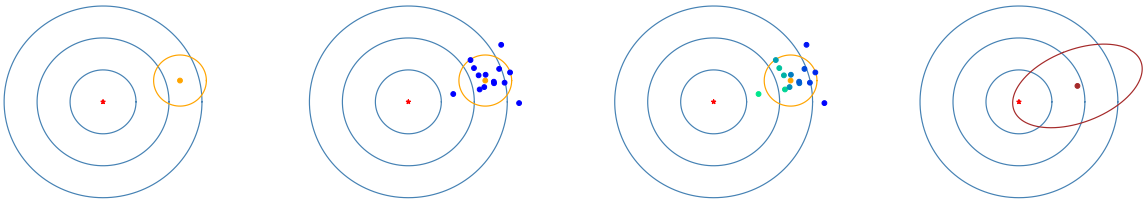


Figure 1: An iteration of CMA-ES: the function to minimize is represented by some of its level sets in blue and its global minimum in red. From left to right: 0. the initial mean m_t is represented in orange, the orange circle shows the area where the candidates solutions (given by the distribution $\mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$) are likely to be located; 1. the samples candidate solutions $m_t + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^i$ are represented in blue; 2. after being ranked (based on their f-values), the candidate solutions are colored accordingly; 3. the updated mean m_{t+1} (brown) is obtained by taking a weighted average of the best candidate solutions.

The quantity $d_\sigma > 0$ (often called damping parameter) is in practice often set proportional to $\sqrt{\mu_{\text{eff}}}$. We will be interested in this thesis in the following smooth alternative [68]:

$$\Gamma_{\text{CSA}}^2 \colon p \in \mathbb{R}^d \mapsto \exp\left(\frac{c_\sigma}{2d_\sigma}\left(\frac{\|p\|^2}{d} - 1\right)\right). \tag{9}$$

Our analysis will consider general stepsize change functions, that encompass (8) and (9).

1.4 *Update of the covariance matrix*

Like the stepsize, the update of the covariance matrix $\mathbf{C}_t$ requires a path p_t^c, such that

$$p_{t+1}^c = (1 - c_c) p_t^c + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}} \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}. \tag{10}$$

The covariance matrix is then updated so that it favors sampling in the direction p_{t+1}^c. Besides, it also exploits the μ best directions of the current iteration. Overall, we have

$$\mathbf{C}_{t+1} = (1 - c_1 - c_\mu)\mathbf{C}_t + c_1 \left[p_{t+1}^c\right]\left[p_{t+1}^c\right]^\top + c_\mu \sqrt{\mathbf{C}_t} \times \sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top \times \sqrt{\mathbf{C}_t}. \tag{11}$$

The learning rates $c_1, c_\mu \in [0, 1]$ should satisfy $c_1 + c_\mu \leqslant 1$. The update of the covariance matrix uses the weights $\mathbf{w}_c = (w_1^c, \dots, w_\mu^c)$ with the same assumptions as $\mathbf{w}_m$. The matrix $[p_{t+1}^c][p_{t+1}^c]^\top$ is usually designated as the rank-one update, and $\sqrt{\mathbf{C}_t} \times \sum_{i=1}^{\mu} w_i^c [U_{t+1}^{s_{t+1}(i)}][U_{t+1}^{s_{t+1}(i)}]^\top \times \sqrt{\mathbf{C}_t}$ the rank-mu update, being almost surely a matrix of rank $\min\{\mu, d\}$.

1.5 *Summary of CMA-ES*

The CMA-ES algorithm can be summarized by the next steps and is illustrated in Figure 1. We initialize $m_0 \in \mathbb{R}^d$, $\sigma_0 > 0$, $\mathbf{C}_0 \in \mathcal{S}_{++}^d$, $p_0^\sigma \in \mathbb{R}^d$, $p_0^c \in \mathbb{R}^d$. For $t \in \mathbb{N}$, we repeat the following lines until we reach a stopping criterion.

1. **Sample** $U_{t+1}^1, \dots, U_{t+1}^\lambda \sim \mathcal{N}(0, \mathbf{I}_d)$ i.i.d., independently of $(m_t, \sigma_t, \mathbf{C}_t, p_t^\sigma, p_t^c)$.

2. **Rank** the candidate solutions, by defining a permutation $s_{t+1} \in \mathfrak{S}_\lambda$ that satisfies:

$$f\left(m_t + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^{s_{t+1}(1)}\right) \leqslant \cdots \leqslant f\left(m_t + \sqrt{\mathbf{C}_t} \sigma_t U_{t+1}^{s_{t+1}(\lambda)}\right) \quad .$$

3. **Update** the parameters:

$$\begin{aligned}
m_{t+1} &= m_t + c_m \sigma_t \sqrt{\mathbf{C}_t} \textstyle\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \\
p_{t+1}^\sigma &= (1-c_\sigma) p_t^\sigma + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}} \textstyle\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \\
\sigma_{t+1} &= \sigma_t \times \Gamma(p_{t+1}^\sigma) \\
p_{t+1}^c &= (1-c_c) p_t^c + \sqrt{c_c(2-c_c)\mu_{\text{eff}}} \sqrt{\mathbf{C}_t} \textstyle\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \\
\mathbf{C}_{t+1} &= (1-c_1-c_\mu)\mathbf{C}_t + c_1 [p_{t+1}^c][p_{t+1}^c]^\top + c_\mu \sqrt{\mathbf{C}_t} \textstyle\sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right] \left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\mathbf{C}_t} \quad .
\end{aligned}$$

2 Related optimization methods

In this section, we present optimization algorithms that relate to CMA-ES or to our convergence analysis.

2.1 Derivative-free optimization algorithms

Evolution strategies are optimization methods that do not exploit the derivatives (or subderivatives) of the objective function, or even assume their existence. Yet there are other derivative-free methods that can be used in practice. We focus here first on variants of ES, with or without covariance matrix, on the simulated annealing algorithm which is derived from the Metropolis-Hastings algorithm, and on the Nelder-Mead algorithm.

2.1.1 Evolution strategies

We briefly present several ES algorithms that are not analyzed in this thesis. It should be noted that ES can be divided in two subclasses: the plus and the comma strategies [68]. The plus strategy corresponds to *elitist* algorithms, for which the function values of the mean m_t can only decrease [124, Chapter C]. On the opposite, comma strategies use an average of candidate solutions to update the mean which is not compared to the previous mean [132, Section 5.2]. The denomination of an ES depend on which of this class it belongs to. For instance, the CMA-ES algorithm presented in Section 1 is called $(\mu/\mu_w, \lambda)$-CMA-ES, where μ denotes the number of candidate solutions used to compute the updated mean, μ_w indicates that recombination weights are used, and λ denotes the total population size at each iteration.

We first present a (1+1)-ES with stepsize adaptation minimizing an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$. We set $\gamma \in (0,1)$ to be a fixed stepsize change factor, we initialize $m_0 \in \mathbb{R}^d$ and $\sigma_0 > 0$, and we proceed as follows for $t \in \mathbb{N}$.

1. **Sample** $x_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 \mathbf{I}_d)$.
2. **If** $f(x_{t+1}) \leqslant f(m_t)$: $m_{t+1} = x_{t+1}$ and $\sigma_{t+1} = \gamma^{4/5}\sigma_t$.
3. **Else**, $m_{t+1} = m_t$ and $\sigma_{t+1} = \gamma^{-1/5}\sigma_t$.

There exist also variants of (1+1)-CMA-ES [83]. We set $\gamma \in (0,1)$ to be a fixed stepsize change factor, we initialize $m_0 \in \mathbb{R}^d$, $p_0 \in \mathbb{R}^d$, $\sigma_0 > 0$ and $\mathbf{C}_0 \in \mathcal{S}_{++}^d$, and we proceed as follows for $t \in \mathbb{N}$.

1. **Sample** $U_{t+1} \sim \mathcal{N}(0, \mathbf{I}_d)$ i.i.d., independently of $(m_t, p_t, \sigma_t, \mathbf{C}_t)$ and compute the candidate solution $x_{t+1} = m_t + \sigma_t \mathbf{C}_t^{1/2} U_{t+1}$.

2. **If** $f(x_{t+1}) \leqslant f(m_t)$:

$$\begin{aligned} m_{t+1} &= x_{t+1} \\ \sigma_{t+1} &= \gamma^{4/5} \sigma_t \\ p_{t+1} &= (1 - c_c) p_t + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}}\, \sqrt{\mathbf{C}_t} U_{t+1} \\ \mathbf{C}_{t+1} &= (1 - c_1)\mathbf{C}_t + c_1 p_{t+1} p_{t+1}^\top \quad . \end{aligned}$$

3. **Else**:

$$\begin{aligned} m_{t+1} &= m_{t+1} \\ \sigma_{t+1} &= \gamma^{-1/5} \sigma_t \\ p_{t+1} &= p_t \\ \mathbf{C}_{t+1} &= \mathbf{C}_t \quad . \end{aligned}$$

Last, we introduce a $(\mu/\mu_w, \lambda)$-ES without covariance matrix adaptation [76]. We initialize $m_0 \in \mathbb{R}^d$ and $\sigma_0 > 0$, and for $t \in \mathbb{N}$ we repeat the following operations.

1. **Sample** $U_{t+1}^1, \ldots, U_{t+1}^\lambda \sim \mathcal{N}(0, \mathbf{I}_d)$ i.i.d., independently of $(m_t, \sigma_t, C_t, p_t^\sigma, p_t^c)$.

2. **Rank** the candidate solutions, by defining a permutation $s_{t+1} \in \mathfrak{S}_\lambda$ that satisfies:

$$f\left(m_t + \sigma_t U_{t+1}^{s_{t+1}(1)}\right) \leqslant \cdots \leqslant f\left(m_t + \sigma_t U_{t+1}^{s_{t+1}(\lambda)}\right) \quad .$$

3. **Update** the parameters:

$$\begin{aligned} m_{t+1} &= m_t + \sigma_t \textstyle\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \\ \sigma_{t+1} &= \sigma_t \times \Gamma\left(\sqrt{\mu_{\text{eff}}} \textstyle\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}\right) \quad . \end{aligned}$$

This algorithm can be obtained with a $(\mu/\mu_w, \lambda)$-CMA-ES by setting $c_c = c_\sigma = c_m = 1$ and $c_1 = c_\mu = 0$.

2.1.2 *The Metropolis-Hastings algorithm and simulated annealing*

Markov chain Monte-Carlo (MCMC) methods [25] and more specifically the Metropolis-Hastings algorithm [106, 78] aim to produce samples from a target probability distribution π on $\mathbb{R}^d$ whose density $p(\cdot)$ is known up to a multiplicative constant. The principle is to define a Markov chain $\{z_t\}_{t\in\mathbb{N}}$ which admits π as a unique invariant probability measure. This relates to our work in two ways: as presented below, the Metropolis-Hastings algorithm may be adapted into a simulated annealing method in order to solve optimization problems, and the theoretical analysis of MCMC methods heavily relies on convergence results for Markov chains—so does the convergence proof of CMA-ES presented in this thesis.

The Metropolis-Hastings algorithm is summarized in the following lines. Given $z_0 \in \mathbb{R}^d$, we repeat for $t = 0, 1, \ldots$:

1. **Sample** $y_{t+1} \sim \nu_{z_t}(\cdot)$ and $U_{t+1} \sim \text{Uniform}([0, 1])$ independently of z_t;

2. **Update** $z_{t+1} = z_t + (y_{t+1} - z_t)\, \mathbb{1}\left\{U_{t+1} \leqslant \frac{p(y_{t+1})}{p(z_t)}\right\}$.

In Step 1., we use a proposal probability distribution $\nu_{z_t}(\cdot)$ with a density positive on $\mathbb{R}^d$, satisfying $\nu_z(y) = \nu_y(z)$, e.g., $\nu_z = \mathcal{N}(z, \mathbf{I}_d)$. The simulated annealing [96] is a method which adapts the Metropolis-Hastings algorithm to solve optimization problems. Despite being initially designed to solve discrete problems (like the traveling salesman problem [28]), the following variant of the simulated annealing aims to minimize an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ [24, 37]. Given $m_0 \in \mathbb{R}^d$, we repeat for $t = 0, 1, \ldots$:

1. **Sample** $x_{t+1} \sim \mathcal{N}(m_t, \mathbf{I}_d)$ and $U_{t+1} \sim \mathrm{Uniform}([0, 1])$ independently of m_t;
2. **Set** $m_{t+1} = m_t + (x_{t+1} - m_t)\, \mathbb{1}\left\{U_{t+1} \leqslant \exp\left(-\frac{f(x_{t+1}) - f(m_t)}{T_t}\right)\right\}$.

The positive constants T_t for $t \in \mathbb{N}$ are called temperature parameters, and are chosen such that T_t tends to 0 when t to $+\infty$ in order to ensure convergence. However, contrary to ES, the simulated annealing does not converge linearly—with a geometric rate. Besides, it relies not only on the comparison of function values, as would ES algorithms, since it uses function values to compute the acceptance probability of Step 2.

2.1.3 *Nelder-Mead algorithm*

We mention here the Nelder-Mead algorithm [115]. Despite being purely deterministic, it relates to CMA-ES in several ways. It is a derivative-free algorithm and is based on the comparison of function values. Moreover, like CMA-ES for certain hyperparameter settings, it is affine-invariant [99, Lemma 3.2]. Moreover, it has been proven to converge linearly for a certain class of strictly convex functions in dimensions 1 and 2 [99]. The principle of the Nelder-Mead algorithm is to update a simplex $\theta = (x_1, \ldots, x_n)$ for which at each iteration we replace the point of θ with the highest f-value by a transformed point (either reflected, expanded or contracted with respect to a centroid of the simplex), or when it fails shrink all the points except the one with lowest f-value.

2.2 *Newton's methods*

In this section, we draw a parallel between CMA-ES and Newton's methods for optimization. Originally, the exact Newton method iteratively approximates an objective function f by a quadratic function using second-order information. However, it requires to solve a linear system which implies a high cost when the dimension of the problem becomes large. To deal with this, approximations of the Newton's method, often called quasi-Newton algorithms, have emerged. The principle is then to approximate the inverse of the Hessian matrix of f. Like CMA-ES (at least for some hyperparameter settings), quasi-Newton methods are affine-invariant. Besides, among the results of this thesis, we show that the covariance matrix in CMA-ES learns the inverse Hessian of a convex-quadratic function.

We first provide a general formulation of exact and inexact Newton's methods. When minimizing a continuously differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$, we initialize $x_0 \in \mathbb{R}^d$ and $\mathbf{B}_0 \in \mathcal{S}^d_{++}$. We repeat for $t \in \mathbb{N}$ the following lines.

1. **Update** $x_{t+1} = x_t - \sigma_t \mathbf{B}_t \nabla f(x_k)$.
2. **Update** $\mathbf{B}_{t+1} = F_B(x_{t+1}, x_t, \mathbf{B}_t; f)$.

We do not specify here how the stepsize $\sigma_t > 0$ is chosen.[2] The update function F_B is then specific to a given Newton's method, with two examples given below. A sufficient condition for this algorithm

[2] Usually for convergence analysis, a line-search satisfying Wolfe's condition is required [116, Eq. (3.7)].

to be affine-invariant is that

$$F_B\left(\mathbf{A}^{-1}(x_{t+1}-a), \mathbf{A}^{-1}(x_t-a), \mathbf{A}^{-1}\mathbf{B}_t\mathbf{A}^{-\top}; f(\mathbf{A}\cdot+a)\right) = \mathbf{A}^{-1}F_B(x_{t+1}, x_t, \mathbf{B}_t; f)\mathbf{A}^{-\top} \quad (12)$$

for every $\mathbf{A} \in \mathsf{GL}_d(\mathbb{R})$ and $a \in \mathbb{R}^d$. However, this does not imply that we learn second-order information. For instance if $F_B(x_{t+1}, x_t, \mathbf{B}_t; f) = \mathbf{B}_t$ then (12) holds but the matrix $\mathbf{B}_t$ remains constant and thus cannot approximate the inverse Hessian of f.

2.2.1 *Exact Newton's method*

The exact Newton's method is obtained with the previous algorithm when f is twice continuously differentiable and $F_B(x_{t+1}, x_t, \mathbf{B}_t; f) = \nabla^2 f(x_{t+1})^{-1}$. When the Hessian $\nabla^2 f$ is Lipschitz continuous, then the Newton's method converges to a stationary point at a quadratic rate [116, Theorem 3.5].

2.2.2 *BFGS*

One popular quasi-Newton algorithm is named after Broyden, Fletcher, Goldfarb and Shanno (BFGS) [26, 44, 56, 134]. The BFGS algorithm uses the update function defined via

$$F_B(x_{t+1}, x_t, \mathbf{B}_t; f) = \left(\mathbf{I}_d - \frac{\Delta x_t \Delta d_t^\top}{\Delta d_t^\top \Delta x_t}\right)\mathbf{B}_t\left(\mathbf{I}_d - \frac{\Delta d_t \Delta x_t^\top}{\Delta d_t^\top \Delta x_t}\right) + \frac{\Delta x_t \Delta x_t^\top}{\Delta d_t^\top \Delta x_t} \quad (13)$$

where $\Delta x_t = x_{t+1} - x_t$ and $\Delta d_t = \nabla f(x_{t+1}) - \nabla f(x_t)$. On top of being affine-invariant, the BFGS algorithms learns the inverse Hessian of a convex-quadratic function in less than d steps [116, Theorem 6.4] and converge with a superlinear rate to a minimizer of a convex twice continuously differentiable function with Lispchitz Hessian [116, Theorem 6.6].

3 Linear Convergence

It has been observed [76, 73] that CMA-ES converges linearly (or geometrically) with high probability to the global optimum x^* for a wide range of problems. In Figures 2 and 3, we have results of numerical experiments (produced with the Python package `pycma` [66]) with that observation. We also see (right part of Figure 3) that for multimodal objective functions, CMA-ES might converge to local optima only. Mathematically, the linear convergence of CMA-ES can be seen as the almost sure limit

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|} = \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|m_t - x^*\|}{\|m_{t-1} - x^*\|}\right] = -\mathrm{CR} \quad (14)$$

where the convergence rate CR is a positive constant. In this thesis, we prove the latter limit under mild assumptions (most notably we assume that the objective function has ellipsoidal level sets). Besides, we prove that CR is positive when using the stepsize change function (9) and when the hyper-parameters of CMA-ES are chosen correctly.

Furthermore, it seems that CMA-ES approximates second-order information of the function f to tackle ill-conditioning. Indeed, empirical convergence rates of CMA-ES while minimizing a spherical function or a ill-conditioned ellipsoidal function seem to be equal, see Figure 2. Beyond the equality of convergence rates for ellipsoidal objective function, we establish the limit for some positive constant $\rho > 0$:

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\mathbf{C}_t}{\det(\mathbf{C}_t)^{1/d}}\right] = \rho\mathbf{H}^{-1} \quad (15)$$

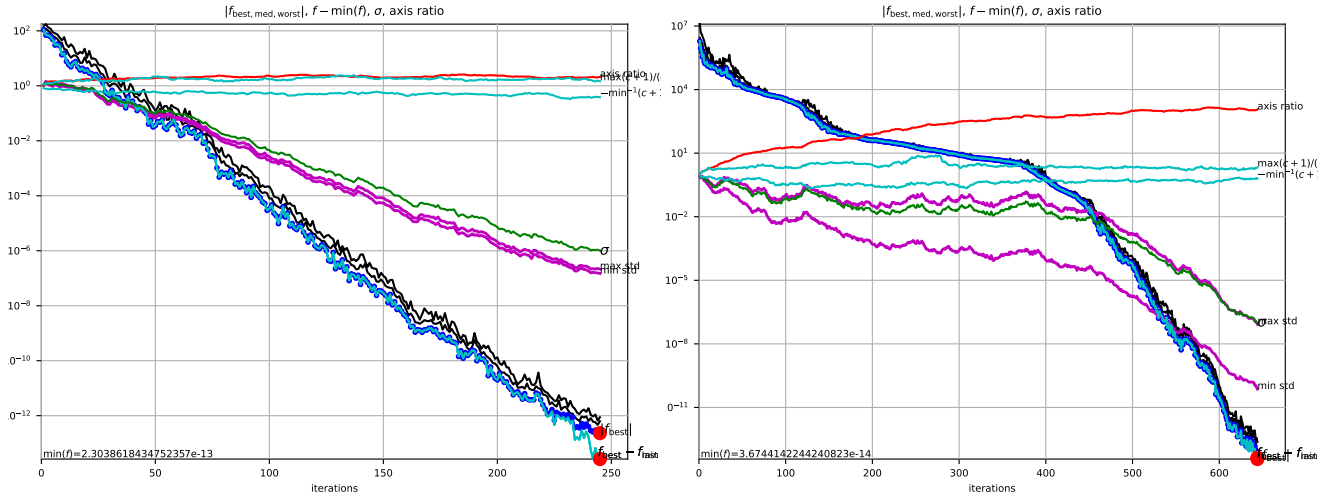


Figure 2: Empirical results of CMA-ES while minimizing the sphere function (left) or a ill-conditioned ($\mathrm{Cond}(\mathbf{H}) = 10^6$) ellipsoidal function (right) in dimension 10. The blue and black lines show the f-values of candidate solutions that are evaluated during the optimization process. Since the y-axis is displayed in a logarithmic scale, a linear decrease is interpreted as linear convergence.

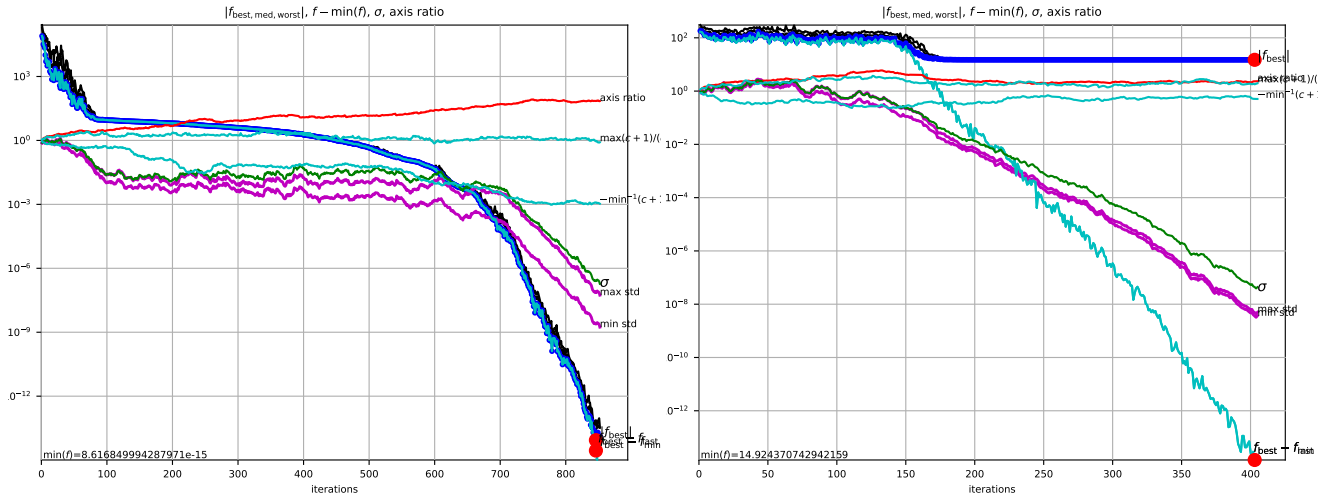


Figure 3: Empirical results of CMA-ES while minimizing the Rosenbrock function (left) or the Rastrigin function (right) in dimension 10. The blue and black lines show the f-values of candidate solutions that are evaluated during the optimization process. Since the y-axis is displayed in a logarithmic scale, a linear decrease is interpreted as linear convergence.

when the objective function f is ellipsoidal, i.e., an increasing transformation of the convex-quadratic $x \mapsto x^\top \mathbf{H} x$, where $\mathbf{H}$ is a positive definite matrix. As an analogy to quasi-convex functions that are increasing transformations of convex functions, the matrix $\mathbf{H}$ is the Hessian matrix of an increasing transformation of f, and thus will be referred to as the *quasi-Hessian matrix* of f.

This last observation is mostly a consequence of the affine-invariance of CMA-ES. This property, illustrated in Figure 4, has been established for alternative variants of CMA-ES [70, 15], and is true for the CMA-ES algorithm studied in this thesis when the decay rate c_σ equals one (we say that we do not use cumulation for the stepsize update).

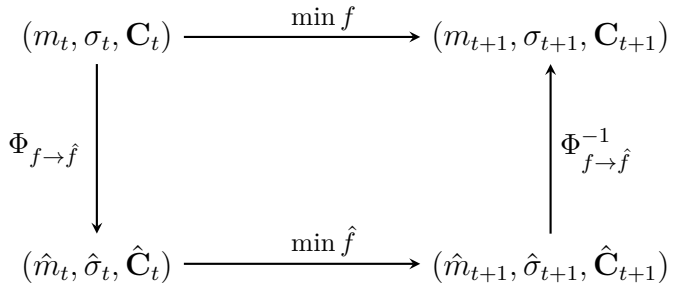


Figure 4: Commutative diagram for CMA-ES. Note however that the affine-invariance is true only for some hyperparameter settings of CMA-ES (in particular when $c_\sigma = 1$).

The results that CMA-ES converges linearly and learns second-order information are exposed in Chapter 5.

4 Introduction to Markov chains

This section is devoted to pose the mathematical framework for the definitions of Markov chains and their transition kernels. Beyond the theoretical analysis of evolution strategies, the study of Markov chains and more specifically of their convergence towards a stationary distribution is an active fields of research and has many applications. For instance, Markov chain Monte Carlo (MCMC) methods [25] aim to sample from a target probability distribution $\pi(\cdot)$, known up to a multiplicative constant. The convergence of a Markov chain at a geometric rate towards its stationary probability distribution π can be established by using a drift criteria such as Theorem 1 below. This has been applied to the Metropolis-Hastings algorithm [128] or adaptive MCMC [127].

We refer to a more specialized textbook on the topic of Markov chains [110] for the interested readers.

Let $(\Omega, \mathcal{F})$ be a measurable space. The set Ω is called *set of all possible outcomes*, and $\mathcal{F}$ is a σ-field on Ω, called *event space*. An *event* is an element of $\mathcal{F}$. We equip $(\Omega, \mathcal{F})$ with a probability measure $\mathbb{P}$. The probability of an event $\mathsf{W} \in \mathcal{F}$ is thus $\mathbb{P}[\mathsf{W}]$. A random variable valued in a measurable space $(\mathsf{X}, \mathcal{X})$ is a measurable map $X\colon \Omega \to \mathsf{X}$. If ν is a measure on $\mathcal{X}$, we say that X is distributed with respect to ν when

$$\mathbb{P}[X \in \mathsf{A}] \coloneqq \mathbb{P}\left[\{\omega \in \Omega \mid X(\omega) \in \mathsf{A}\}\right] = \nu(\mathsf{A}) \quad \text{for } \mathsf{A} \in \mathcal{X}. \tag{16}$$

4.1 Kernels and Markov chains

We suppose that the state space X is a topological space, which is locally compact, separable and metrizable (e.g., an Euclidean space). We denote $\mathcal{B}(\mathsf{X})$ the Borelian σ-fields of X. Note however that we can define kernels and Markov chains on more general spaces. A kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ is a function $P\colon \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}_+$ such that

(a) for every $x \in \mathsf{X}$, $\mathsf{A} \in \mathcal{B}(\mathsf{X}) \mapsto P(x, \mathsf{A})$ is a measure,

(b) for every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$, $x \in \mathsf{X} \mapsto P(x, \mathsf{A})$ is a measurable map.

Moreover, when for any $x \in \mathsf{X}$, $P(x, \mathsf{X}) = 1$, we say that P is a transition kernel [110, Chapter 3]. We say that P is a substochastic kernel when $P(x, \mathsf{X}) \leqslant 1$ for $x \in \mathsf{X}$.

In the sequel, P is a transition kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. In order to define then a (time-homogeneous) Markov chain $\{\theta_k\}_{k\in\mathbb{N}}$ on X with transition kernel P, we equip the set of all outcomes $(\Omega, \mathcal{F})$ with a probability measure $\mathbb{P}[\cdot \mid \theta_0 \sim \nu]$, where ν is a probability measure on $\mathcal{B}(\mathsf{X})$. Then, we define the sequence $\{\theta_k\}_{k\in\mathbb{N}}$, such that the random variable θ_0 is distributed with respect to ν, and for every $k \in \mathbb{N}^*$ the random variable θ_k satisfies

$$\mathbb{P}[\theta_k \in \mathsf{A} \mid \theta_0 \sim \nu] = \nu P^k(\mathsf{A}) \coloneqq \int_{\mathsf{X}^{k+1}} P(x^k, \mathsf{A}) P(x^{k-1}, \mathrm{d}x^k) \dots P(x^0, \mathrm{d}x^1)\nu(\mathrm{d}x^0) \tag{17}$$

for every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$. For every $x \in \mathsf{X}$, we define the k-steps transition kernel $P^k(x, \cdot) = \delta_x P^k(\cdot)$, where δ_x is the Dirac measure at x.

We define in the next paragraphs essential notions to characterize the convergence of Markov chains towards stationary measures.

4.2 *Irreducibility*

Irreducibility is central for the convergence of Markov chains [110, Chapter 4]. On finite state spaces, a Markov chain is irreducible if every state is reachable in a finite number of steps with positive probability for every starting state. For more general state spaces, we define irreducibility with respect to a measure φ. For a (nonnegative) measure φ on $\mathcal{B}(\mathsf{X})$, the support of φ is the set $\operatorname{supp}\varphi$ composed of subsets $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ such that $\varphi(A) > 0$.

Definition 1. The transition kernel P on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ is said to be *irreducible* if there exists a nontrivial measure φ on $\mathcal{B}(\mathsf{X})$ such that for every $\mathsf{A} \in \operatorname{supp}\varphi$ and every $x \in \mathsf{X}$, we have

$$\sum_{k\geqslant 1} P^k(x, \mathsf{A}) > 0.$$

We also introduce the notion of maximal irreducibility measure.

Definition 2. Let P be an irreducible transition kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. A *maximal irreducibility measure* of P is an irreducibility measure ψ of P such that for every irreducibility measure φ of P, we have

$$\operatorname{supp}\varphi \subset \operatorname{supp}\psi.$$

4.3 *Periodicity, aperiodicity*

For an irreducible transition kernel P on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ with maximal irreducibility measure ψ, a *d-cycle* is a collection of disjoint subsets $\{\mathsf{A}_1, \dots, \mathsf{A}_d\}$ of $\mathcal{B}(\mathsf{X})$ such that

(1) for $k = 1, \dots, d-1$ and $x_k \in \mathsf{A}_k$, $P(x_k, \mathsf{A}_{k+1}) = 1$ and for $x_d \in \mathsf{A}_d$, $P(x_d, \mathsf{A}_1) = 1$;

(2) the support of ψ is included in $\mathsf{A}_1 \cup \dots \cup \mathsf{A}_d$.

Definition 3 ([110, Chapter 5])**.** The *period* of an irreducible transition kernel P is the largest integer d such that a d-cycle exists. When the period of P equals 1, we say that P is *aperiodic*.

4.4 *Petite sets, small sets, T-chains*

Petite and small sets [110, Chapter 5] play an important role in the convergence analysis of Markov chains. Below in Theorem 1, we give a *drift condition* for the geometric convergence of Markov chains which relies on small sets.

Definition 4. Let P be a transition kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ and let $\mathsf{C} \in \mathcal{B}(\mathsf{X})$.

(i) If there exists a probability measure b on $\mathbb{N}^*$ and a nontrivial measure ν_b on $\mathcal{B}(\mathsf{X})$ such that

$$\sum_{k \geqslant 1} b(k) P^k(x, \mathsf{A}) \geqslant \nu_b(\mathsf{A}) \quad \text{for every } x \in \mathsf{C} \text{ and } \mathsf{A} \in \mathcal{B}(\mathsf{X}),$$

then C is said to be *petite*.

(ii) If there exists $a \in \mathbb{N}^*$ and a nontrivial measure ν_a on $\mathcal{B}(\mathsf{X})$ such that

$$P^a(x, \mathsf{A}) \geqslant \nu_a(\mathsf{A}) \quad \text{for every } x \in \mathsf{C} \text{ and } \mathsf{A} \in \mathcal{B}(\mathsf{X}),$$

then C is said to be *small*.

We define now the notion of T-chain [110, Chapter 6], which is a topological property of a transition kernel. This facilitates the identification of petite sets: for an irreducible T-chain, every compact set is petite [110, Theorem 6.2.5].

Definition 5. Let P be an irreducible transition kernel and T a substochastic kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$.

(i) If $x \in \mathsf{X} \mapsto T(x, \mathsf{A})$ is lower semicontinuous for every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ and if there exists a measure b on $\mathbb{N}$ such that

$$\sum_{k \in \mathbb{N}} b(k) P^k(x, \mathsf{A}) \geqslant T(x, \mathsf{A}) \quad \text{for } x \in \mathsf{X} \text{ and } \mathsf{A} \in \mathcal{B}(\mathsf{X}),$$

then T is called a *continuous component* of P.

(ii) If P possesses a continuous component T such that $T(x, \mathsf{X}) > 0$ for every $x \in \mathsf{X}$, then P is the transition kernel of a *T-chain*.

4.5 *Recurrence, Harris recurrence*

One central step of the convergence proof in this thesis is to prove that a normalized Markov chain underlying CMA-ES is recurrent. The definition of recurrence relates to the occupation time of a Markov chain. For $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ and $x \in \mathsf{X}$, we denote $N_{\mathsf{A}}(x)$ the expected occupation time of the Markov chain $\{\theta_k\}_{k \in \mathbb{N}}$ conditionally to $\theta_0 = x$, that is,

$$N_{\mathsf{A}}(x) = \sum_{k \geqslant 1} P^k(x, \mathsf{A}) \ . \tag{18}$$

Definition 6 ([110, Chapter 8])**.** Let P be a transition kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$.

(i) A subset $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ is *recurrent* when for every $x \in \mathsf{A}$, $N_{\mathsf{A}}(x) = +\infty$.

(ii) Suppose that P is irreducible and denote ψ its maximal irreducibility measure. Then P is *recurrent* if every $\mathsf{A} \in \operatorname{supp} \psi$ is recurrent.

Likewise, we can define transient subsets when they are visited finitely often (in expectation), and transient kernels as irreducible chains that are not recurrent. However, only recurrent chains will be of interest in this thesis. A stronger notion is Harris recurrence. Define the random variable η_{A} as the occupation time in A by

$$\eta_{\mathsf{A}} = \sum_{k \geqslant 1} \mathbb{1}_{\theta_k \in \mathsf{A}} \quad . \tag{19}$$

Definition 7 ([110, Chapter 9])**.** Let P be a transition kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$.

(i) A subset $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ is *Harris recurrent* when for all $x \in \mathsf{A}$, we have $\eta_{\mathsf{A}} = +\infty$ almost surely.

(ii) Suppose that P is irreducible and denote ψ its maximal irreducibility measure. Then P is *Harris recurrent* if every $\mathsf{A} \in \operatorname{supp} \psi$ is Harris recurrent.

4.6 *Invariant measures, positivity*

Our interest is to the convergence of Markov chains to stationary measures, which have the particularity to be invariant for the transition kernel of the Markov chain.

Definition 8 ([110, Chapter 10])**.** Let P be a transition kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. A measure π on $\mathcal{B}(\mathsf{X})$ is *invariant* when $\pi P = \pi$. When there exists an invariant probability measure on $\mathcal{B}(\mathsf{X})$, then P is said to be *positive*.

4.7 *Ergodicity*

Ergodicity [110, Chapter 13] characterizes the convergence of a Markov chain to a stationary measure. We are interested in geometric ergodicity [110, Chapter 15] in this thesis. Let $V : \mathsf{X} \to [1, +\infty]$ be a measurable function that we call *potential function*. Given ν a (signed) measure on $\mathcal{X}$, we set

$$\|\nu\|_V = \sup_{|g| \leqslant V} \int g \, \mathrm{d}\nu. \tag{20}$$

Definition 9. Let P be a transition kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$, and let $V : \mathsf{X} \to [1, +\infty]$ be measurable. We say that P is *V-geometrically ergodic* when it is positive Harris recurrent with invariant probability measure π, where $V \in \mathsf{L}^1(\pi)$, and when there exists a constant $r > 0$ such that for every $x \in \mathsf{X}$

$$\sum_{k \geqslant 1} r^k \left\| P^k(x, \cdot) - \pi(\cdot) \right\|_V < +\infty \quad .$$

When P is 1-geometrically ergodic, we simply say that it is *geometrically ergodic*.

Our proof of convergence is based on the following criterion for geometric ergodicity. This is known as a geometric drift condition, or Foster-Lyapunov condition, for ergodicity. However, note that we use a *state-dependent* condition, as the standard drift condition would assume $n(x) = 1$ in the theorem below.

Theorem 1 ([110, Theorem 19.1.3]). Suppose that P is a irreducible, aperiodic transition kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. Let $n\colon \mathsf{X} \to \mathbb{N}^*$ and $V\colon \mathsf{X} \to [1, +\infty]$ be two measurable functions. If there exist a small set $\mathsf{C} \in \mathcal{B}(\mathsf{X})$ such that V is bounded on C, and positive constants $\rho < 1$ and $b < +\infty$ such that

$$\int P^{n(x)}(x, \mathrm{d}y) V(y) \leqslant \rho^{n(x)} \times (V(x) + b\mathbb{1}_{x \in \mathsf{C}}) \quad \text{for } x \in \mathsf{X}, \tag{21}$$

then, P is geometrically ergodic. Moreover, we have

$$\sum_{k \geqslant 1} r^k \left\| P^k(x, \cdot) - \pi(\cdot) \right\|_1 \leqslant RV(x) \quad \text{for } x \in \mathsf{X}, \tag{22}$$

for some constants $R < \infty$ and $r > 1$, where π is the unique invariant probability measure of P.

4.8 *Stability of nonlinear Markov chains via the analysis of a deterministic control model*

In this section, we recall results [109, 29] that provide sufficient condition for a Markov chain to be an irreducible aperiodic T-chain.

Let X and W be locally compact, separable, metrizable spaces equipped with their respective Borelian σ-fields $\mathcal{B}(\mathsf{X})$ and $\mathcal{B}(\mathsf{W})$, and let $(\mathsf{U}, \mathcal{U})$ be a measured space. Consider $\{\phi_k\}_{k \in \mathbb{N}}$ a Markov chain with state space X such that

$$\phi_{k+1} = F\left(\phi_k, \alpha(\phi_k, U_{k+1})\right) \tag{23}$$

where $F\colon \mathsf{X} \times \mathsf{W} \to \mathsf{X}$ and $\alpha\colon \mathsf{X} \times \mathsf{U} \to \mathsf{W}$ be measurable functions, and $\{U_{k+1}\}_{k \in \mathbb{N}}$ is an i.i.d. process valued in U independent of ϕ_0.

We define then the *extended transition map* S_x^k for $x \in \mathsf{X}$ and $k \in \mathbb{N}$ by

$$\begin{aligned} &S_x^0 = x \\ &S_x^{k+1}(w_{1:k+1}) = F\left(S_x^k(w_{1:k}), w_{k+1}\right) \quad \text{for } w_{1:k+1} = (w_1, \ldots, w_{k+1}) \in \mathsf{W}^{k+1}. \end{aligned} \tag{24}$$

Likewise, we define the *extended probability density* p_x^k for $x \in \mathsf{X}$ and $k \geqslant 1$ by

$$\begin{aligned} &p_x^1(w_1) = p_x(w_1) \quad \text{for } w_1 \in \mathsf{W} \\ &p_x^{k+1}(w_{1:k+1}) = p_x^k(w_{1:k}) p_{S_x^k(w_{1:k})}(w_{k+1}) \quad \text{for } w_{1:k+1} = (w_1, \ldots, w_{k+1}) \in \mathsf{W}^{k+1}, \end{aligned} \tag{25}$$

where $p_x(\cdot)$ denotes a probability density function w.r.t. a σ-finite measure ζ_{W} on W for the random variable $\alpha(x, U_1)$. In this section, we only consider the cases where ζ_{W} is the Lebesgue measure on the open subset W of $\mathbb{R}^p$. Besides, we suppose that $(x, w) \mapsto p_x(w)$ is lower semi-continuous (l.s.c.) on $\mathsf{X} \times \mathsf{W}$. We define the control sets by

$$\mathcal{O}_x^k = \{w_{1:k} \in \mathsf{W}^k \mid p_x^k(w_{1:k}) > 0\} \ . \tag{26}$$

Then, the set of attainable states by $\{\phi_k\}_{k \in \mathbb{N}}$ starting at $x \in \mathsf{X}$ equals

$$A_+(x) = \bigcup_{k \in \mathbb{N}} A_+^k(x) := \bigcup_{k \in \mathbb{N}} \left\{S_x^k(w_{1:k}) \mid w_{1:k} \in \mathcal{O}_x^k\right\} \ . \tag{27}$$

When a state $x^* \in \mathsf{X}$ satisfies

$$x^* \in \bigcap_{x \in \mathsf{X}} \overline{A_+(x)} \ , \tag{28}$$

we say that x^* is a *globally attracting state*. When F is differentiable, we define for $x \in \mathsf{X}$ the following *controllability condition*:

there exist $k \geqslant 1$ and $w_{1:k} \in \mathcal{O}_x^k$ such that the derivative $\mathcal{D}S_x^k(w_{1:k})$ is of maximal rank. (29)

We recall a first criterion for the chain $\{\phi_k\}_{k\in\mathbb{N}}$ to be an irreducible T-chain, with stronger assumptions on F and α.

Theorem 2 ([110, Theorems 7.1.1 and 7.2.6, Propositions 7.1.2 and 7.2.5]). Suppose that $\mathsf{X} = \mathbb{R}^n$ and $\mathsf{U} = \mathsf{W} = \mathbb{R}^p$. If moreover F is infinitely many times continuously differentiable, and if $\alpha(x, u) = u$ for $(x, u) \in \mathsf{X} \times \mathsf{U}$, then we have:

(i) if every $x \in \mathsf{X}$ satisfies (29), then $\{\phi_k\}_{k\in\mathbb{N}}$ is a T-chain;

(ii) if there exists a globally attracting state x^* which satisfies (29), then $\{\phi_k\}_{k\in\mathbb{N}}$ is an irreducible T-chain.

To characterize the aperiodicity, we introduce a stronger notion than globally attracting states. A state $x^* \in \mathsf{X}$ is steadily attracting when for every $x \in \mathsf{X}$ and for every neighborhood U of x^*, there exists $T \in \mathbb{N}$ such that $A_+^k(x)$ intersects U for every $k \geqslant T$.

The following criterion generalizes the conditions to apply Theorem 2, and provide a supplementary condition to obtain aperiodicity.

Theorem 3 ([29, Corollary 4.1, Theorems 4.2 and 4.4]). Suppose that X, U and W are open subsets of $\mathbb{R}^n$, $\mathbb{R}^m$, $\mathbb{R}^p$, respectively. If F is $\mathcal{C}^1$, then we have:

(i) if every $x \in \mathsf{X}$ satisfies (29), then $\{\phi_k\}_{k\in\mathbb{N}}$ is a T-chain;

(ii) if there exists a globally attracting state x^* which satisfies (29), then $\{\phi_k\}_{k\in\mathbb{N}}$ is an irreducible T-chain;

(iii) if there exists a steadily attracting state x^* which satisfies (29), then $\{\phi_k\}_{k\in\mathbb{N}}$ is an irreducible aperiodic T-chain.

In Chapter 1, we generalize the conditions of Theorem 3 to when the sets X and W are manifolds, and the function F is locally Lipschitz. In Chapter 2, we apply these results to Markov chain underlying CMA-ES and prove that it is an irreducible aperiodic T-chain.

5 Convergence and theoretical guarantees of evolution strategies

In this section, we provide an overview of previous works that give theoretical insights of different variants of ES, in particular for convergence. Early works [23] have proven that ES with stepsize proportional to the distance to the optimum, and slightly later ES with adaptive stepsize [14], converge linearly to the optimum of a spherical function $f = g(\|\cdot\|_2)$, with $g\colon \mathbb{R} \to \mathbb{R}$ a monotonous function. This approach is very similar to the one used in this thesis: by defining a normalized process $z_t = (m_t - x^*)/\sigma_t$, we obtain a geometrically ergodic Markov chain. This yields to the linear behavior of the log progress of the mean to the optimum.

Later, linear convergence of the (1+1)-ES with stepsize adaptation is obtained first for spherical [87] and ellipsoidal problems [88, 89, 90] and for noisy objective functions [92]. More recently, drift

analysis allow to estimate the convergence rate and expected hitting time on spherical problems [3] and on smooth strongly convex problems [4]. This methodology differs from ours since it requires to obtain a drift condition for every initialization of the state space (whereas we can start outside an arbitrary large compact). It was inspired from approaches used in the analysis of evolutionary algorithms for combinatorial problems [42, 40, 145, 101].

Moreover, different theoretical results have been established to stochastic algorithms and in particular the ES class coming from the information-geometric optimization (IGO) point of view [119, 8]. It has been shown that a simplified version of CMA-ES, without the stepsize adaptation or the rank-one update, can be seen as a natural gradient update on the space of probability measures. Moreover, there have been local convergence analysis of ordinary differential equations (ODE) following from the IGO setting of stepsize adaptive ES for increasing transformation of twice continuously differentiable objective functions [5]. The analysis of an underlying ODE associated to stepsize adaptive ES has recently allowed to deduce the linear convergence for spherical (i.e., increasing transformation of the Euclidean norm) objective functions [6].

The first convergence proof of algorithms that include ES with similar mechanisms to CMA-ES for the covariance matrix adaptation has exploited a different stepsize update [39], ensuring the limit $\sigma_t \to 0$ for objective functions that are bounded by below, and a sufficient decrease condition on the f-value of the mean before the update. Under additional assumptions, it has been proven that the mean then converges to a stationary point. However these modifications do not reflect important features of CMA-ES, as they do not allow the f-value of the mean to increase and thus limit the exploration of the state space, and force the stepsize to quickly decrease, affecting the performances of the algorithm on multi-modal problems, for which exploration is essential.

More recently, the convergence analysis of ES with a different covariance matrix adaptation, namely the Hessian estimation ES (HE-ES), has shown the linear convergence of the mean to the optimum and the learning of the inverse Hessian [54] for convex-quadratic problems. Furthermore, it was shown that the convergence rate of HE-ES did not depend on the conditioning of the problem. The algorithm HE-ES, although it was "designed to be conceptually closed to CMA-ES" imposes the offspring to be sampled on orthogonal mirrored directions, which facilitates the theoretical analysis of the covariance matrix update. Moreover, it approximates the Hessian by a finite differences scheme, relying on values of the objective function, in contrast to CMA-ES, which uses comparisons of function values only. This approach is also based on a Markov chain analysis, but yet assumed the algorithm to be elitist, i.e., it did not allow the f-value of the mean to increase.

Recent works have used a Markov chain approach to analyze the global convergence of ES with stepsize adaptation on scaling-invariant functions [17, 141]. Yet, the definition of a scaling-invariant function allows in these works to define a normalized process and prove that it is a geometrically ergodic Markov chain. This have been the main inspiration for the work presented in this thesis, as the key ideas of the proofs follow the same scheme. However, there are several improvements of the method used at that time that our proof relies on. Firstly, a previous convergence proof for stepsize adaptive ES [141] was based on conditions for irreducibility, as well as for aperiodicity and topological properties, that required to have continuously differentiable updates on open subsets of Euclidean spaces [29]. Chapter 1 generalizes these conditions to include locally Lipschitz updates on manifolds, that allow to analyze the nonsmooth CSA update in (8), and facilitate the analysis of a chain which involves a normalized covariance matrix, valued in a manifold of normalized positive definite matrices. Besides, our results are based on a state-dependent drift condition (instead of a standard drift condition) to obtain the geometric ergodicity of a normalized Markov chain, easing the analysis when we use cumulation on the stepsize.

5.1 *Proving linear behavior of ES by analyzing a normalized Markov chain*

We focus here on approaches based on the analysis of an underlying normalized Markov chain. This is the approach adopted in this thesis in order to prove linear convergence of CMA-ES.

While this was applied to several variants of ES, we state the results on a stepsize adaptive $(\mu/\mu_w, \lambda)$-ES (without covariance matrix adaptation) minimizing an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ introduced in Section 2.1.1.

The convergence of this algorithm [141] was proven for different stepsize changes $\Gamma\colon \mathbb{R}^d \to \mathbb{R}$ satisfying various assumptions. We state these results in the particular case $\Gamma = \Gamma^2_{\text{CSA}}$, see (9). The convergence proof is mainly based on the assumption that the objective function f is *scaling-invariant* [142], i.e., when there exists a point $x^* \in \mathbb{R}^d$ such that

$$f(x^* + x) \leqslant f(x^* + y) \Rightarrow f(x^* + \rho x) \leqslant f(x^* + \rho y) \quad \text{for } x, y \in \mathbb{R}^d \text{ and } \rho > 0. \tag{30}$$

In this case, if we define the following process by

$$z_t = \frac{m_t - x^*}{\sigma_t} \quad \text{for } t \in \mathbb{N}, \tag{31}$$

then $\{z_t\}_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain [141, Proposition 2]. Based on Theorem 3, this chain is an irreducible aperiodic T-chain, and using a geometric drift condition like Theorem 1, we obtain that this chain is ergodic.

Theorem 4 ([141, Theorem 6])**.** If f is smooth and scaling-invariant with respect to x^* and $\Gamma = \Gamma^2_{\text{CSA}}$, then $\{z_t\}_{t\in\mathbb{N}}$ is a geometrically ergodic Markov chain.

Geometric ergodicity allows to conclude to the linear behavior of the algorithm. Indeed,

$$\begin{aligned}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} &= \frac{1}{T}\sum_{t=0}^{T-1}\log\|m_{t+1} - x^*\| - \log\|m_t - x^*\| \\ &= \frac{1}{T}\sum_{t=0}^{T-1}\log\|z_{t+1}\| - \log\|z_t\| + \log\Gamma^2_{\text{CSA}}\left(\sqrt{\mu_{\text{eff}}}\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}\right).\end{aligned}$$

Then, by proving some integrability properties [141, Propositions 10 and 11], we find that the above average log-progress converges almost surely to a real limit.

Theorem 5 ([141, Theorem 7])**.** If f is smooth and scaling-invariant with respect to x^* and $\Gamma = \Gamma^2_{\text{CSA}}$, then there exists $\text{CR} \in \mathbb{R}$ such that for every initial condition:

$$\lim_{T\to\infty}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty}\frac{1}{T}\log\frac{\sigma_T}{\sigma_0} = -\text{CR}$$

and

$$\lim_{t\to\infty}\mathbb{E}\left[\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}\right] = \lim_{t\to\infty}\mathbb{E}\left[\log\frac{\sigma_{t+1}}{\sigma_t}\right] = -\text{CR}\ .$$

The case $\text{CR} > 0$ corresponds to linear convergence.

However a proof that CR is positive is missing. We prove in Chapter 5 that when minimizing ellipsoid scaling-invariant functions, under conditions on the hyperparameters, CMA-ES with stepsize change Γ^2_{CSA} converges linearly. The same arguments would apply for stepsize adaptive ES.

6 Methodology and overview

We give in this section the overview of the thesis. Each of the subsection below shows the contributions of the chapters of the thesis.

The main goal of the thesis is to provide a rigorous proof of linear convergence of the mean m_t produced by CMA-ES to the optimum x^* for a certain class of objective functions $f\colon \mathbb{R}^d \to \mathbb{R}$. Moreover, we prove that the covariance matrix $\mathbf{C}_t$ learns second-order information.

Throughout the manuscript, we consider several assumptions that are required for different parts of the proof. Hence, in Chapter 2, we suppose that the objective function f is scaling-invariant and a monotonously increasing transformation of a continuous function, with Lebesgue-negligible level sets, whereas in Chapter 4, we restrict to ellipsoidal objective functions, that is, when there exists a strictly increasing function $g\colon \mathbb{R} \to \mathbb{R}$ such that $f(x) = g(x^\top \mathbf{H} x)$ for some matrix $\mathbf{H} \in \mathcal{S}^d_{++}$. The name *ellipsoidal* is due to the ellipsoidal shape of the level sets of f.

Besides, in order to pave the way to a more general analysis, in Chapter 2 we do not assume that the random variables U^i_{t+1} in (1) follow a standard normal distribution, but instead a generic probability distribution on $\mathcal{B}(\mathbb{R}^d)$ with minimal assumptions. In Chapters 4 and 5 our assumptions will include only normal distributions though.

6.1 *Chapter 1: On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds [52]*

Throughout the thesis, and particularly in Chapters 2, 4 and 5 we analyze a normalized Markov chain underlying CMA-ES to deduce finally linear convergence. Before that, we introduce in Chapter 1 the methodology that is used in Chapter 2. Indeed, while previous works [109, 29] provided conditions to analyze the irreducibility of Markov chains underlying various ES algorithms, they have limitations that prevent their application to CMA-ES. We generalize these conditions in Chapter 1, and we give here a brief overview of the main result of this chapter.

Consider a Markov chain $\Theta = \{\theta_t\}_{t\in\mathbb{N}}$ such that

$$\theta_{t+1} = F(\theta_t, \alpha(\theta_t, U_{t+1})) \tag{32}$$

where $F\colon \mathsf{X} \times \mathsf{W} \to \mathsf{X}$ is a locally Lipschitz function between the manifolds $\mathsf{X} \times \mathsf{W}$ and W, α is a measurable function between the measured spaces $\mathsf{X} \times \mathsf{U}$ and W satisfying mild assumptions, and $\{U_{t+1}\}_{t\in\mathbb{N}}$ is a i.i.d. process independent of θ_0.

We give in Chapter 1 various definitions associated to the control model (32) and some of them are recalled above in Section 4.8. The main contribution of Chapter 1 is summarized in the next theorem.

Theorem 6 (Theorem 1.2 in Chapter 1)**.** Suppose that there exists a globally attracting state θ^* such that the controllability condition (29) holds. Then, Θ is an irreducible T-chain. If moreover θ^* is steadily attracting, then Θ is aperiodic.

It remains several limitations to this result.

First, we encounter in Chapter 2 the issue that the normalized Markov chain underlying CMA-ES is defined on a nonsmooth manifold when the normalization is equal, e.g., to the minimum eigenvalue. By defining a homeomorphism between two Markov chains, with one defined on a smooth manifolds and satisfying the conditions required to apply Theorem 6, we are able to prove our results. It would be however interesting to have more general conditions which includes nonsmooth manifolds.

Second, the assumption that the random variable $\alpha(\theta, U_1)$ is absolutely continuous with respect to a σ-finite measure locally equivalent to Lebesgue's measure cannot apply to ES with a plus strategy (see Section 2.1.1 where we introduce the difference between plus and comma strategies). An interesting extension would be to include cases where $\alpha(\theta, U_1)$ is a mixture of distributions with density and Dirac distributions.

6.2 *Chapter 2: Irreducibility of nonsmooth state-space models with an application to CMA-ES [53]*

Chapter 2 constitutes the starting point of our convergence proof for CMA-ES. In this chapter, we introduce a normalized process underlying CMA-ES, defined by

$$\begin{aligned} z_t &= \frac{m_t - x^*}{\sigma_t \sqrt{R(\mathbf{C}_t)}} \\ \mathbf{\Sigma}_t &= \frac{1}{R(\mathbf{C}_t)} \mathbf{C}_t \\ p_t &= p_t^\sigma \\ q_t &= \frac{p_t^c}{\sqrt{R(\mathbf{C}_{t-1})}} \\ r_t &= \frac{R(\mathbf{C}_t)}{R(\mathbf{C}_{t-1})} \end{aligned} \tag{33}$$

where $R \colon \mathcal{S}_{++}^d \to \mathbb{R}_{++}$ is a normalization function for the covariance matrix. Proving that this process is a geometrically ergodic Markov chain is the central step of this proof of linear convergence. In this chapter, we prove that it defines a Markov chain when the objective function is scaling-invariant.

Theorem 7 (Proposition 2.2 in Chapter 2)**.** If f is scaling-invariant with respect to x^* and R is positively homogeneous, then $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain.

Moreover, we use Theorem 6 and the methodology developed in Chapter 1 to deduce that it is an irreducible aperiodic T-chain. However, Theorem 6 does not allow to include all the hyperparameter settings of CMA-ES that we are interested in without repeating a similar proof. For instance, if the parameter for the stepsize update c_σ equals 1, the Markov chain that we prove to be irreducible is not the normalized chain (33), but the chain $\{(z_t, \mathbf{\Sigma}_t, q_t, r_t)\}_{t\in\mathbb{N}}$. To deal with these different Markov chains, we introduce the notions of *redundant* and *projected* Markov chains. More precisely, a Markov chain $\{(\theta_t, \chi_t)\}_{t\in\mathbb{N}}$ valued in a state space $\mathsf{X} \times \mathsf{Y}$ is said to be redundant, when the process $\{\theta_t\}_{t\in\mathbb{N}}$ valued in X is itself a Markov chain. In this case, the Markov chain $\{\theta_t\}_{t\in\mathbb{N}}$ is called projected. We provide in Chapter 2 a generalization of Theorem 6 for redundant state space models. We consider the Markov chain

$$(\theta_{t+1}, \chi_{t+1}) = \tilde{F}((\theta_t, \chi_t), \tilde{\alpha}((\theta_t, \chi_t), U_{t+1})) \tag{34}$$

where $\tilde{F}$ and $\tilde{\alpha}$ obey similar assumptions than for F and α in the previous section. We introduce then the following controllability condition at a point $(\theta, \chi) \in \mathsf{X}$:

$$\exists w_{1:k} \in \overline{\tilde{\mathcal{O}}^k_{(\theta,\chi)}}, \forall (h_\theta, h_\chi) \in \mathsf{T}_{\tilde{S}^k_{(\theta,\chi)}(w_{1:k})}(\mathsf{X} \times \mathsf{Y}), h_\theta \in \operatorname{range} \mathcal{D}(\Pi_{\mathsf{X}} \circ \tilde{S}^k_{(\theta,\chi)})(w_{1:k}) \tag{35}$$

where Π_{X} is the projection of $\mathsf{X} \times \mathsf{Y}$ on X.

Theorem 8 (Theorem 2.3 in Chapter 2)**.** Suppose that there exists a steadily attracting state (θ^*, χ^*) such that the the controllability condition (35) holds. Then, $\{\theta_t\}_{t\in\mathbb{N}}$ is an irreducible aperiodic T-chain.

This allows us to prove the irreducibility of the normalized Markov chain.

Theorem 9 (Theorem 2.1 in Chapter 2)**.** Under supplementary assumptions on the objective function f, the normalization function R, the stepsize change Γ, the sampling distribution ν_U^d and the hyperparameters of CMA-ES, the process

(i) $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ if $c_c \neq 1$ and $c_\sigma \neq 1$;

(ii) $\{z_t, q_t, \mathbf{\Sigma}_t, r_t\}_{t\in\mathbb{N}}$ if $c_c \neq 1$ and $c_\sigma = 1$;

(iii) $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ if $c_c = 1$ and $c_\sigma \neq 1$;

(iv) $\{(z_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ if $c_c = 1$ and $c_\sigma = 1$;

is an irreducible aperiodic T-chain, and compact sets are small.

6.3 *Chapter 3: Asymptotic estimation of a perturbed symmetric eigenproblem [51]*

The central step of the present convergence proof for CMA-ES is presented in Chapter 4. We show there that the normalized process introduced earlier satisfies a drift condition and thus is geometrically ergodic. However, this step was very challenging, as it requires to obtain precise upper bounds on the estimated parameters of CMA-ES. Particularly, we have to determine a tight bound for the expected largest eigenvalue of the covariance matrix in cases it should decrease, when the condition number of the covariance matrix is large. Yet, the normalization of the covariance matrix imposes to control as well the increase of the smallest eigenvalue, and therefore understanding how the axis of the covariance matrix change over iterations is determinant.

This is the subject of Chapter 3. We analyze there a perturbation of a definite positive matrix similar to the rank-mu update of CMA-ES, and we provide bounds on the projection of the eigenvectors of the updated system on the initial one, when the condition number of the matrix is high. Formally, the problem writes as

$$\mathbf{B} = \mathbf{A} + \sqrt{\mathbf{A}} \sum_{i=1}^{\mu} v_i v_i^\top \sqrt{\mathbf{A}} \tag{36}$$

where $\mathbf{A} \in \mathcal{S}_{++}^d$ represent the initial matrix and the vectors $v_i \in \mathbb{R}^d$, $i = 1, \dots, \mu$, are used to compute the so-called rank-mu update. In the following, we denote λ_i the i-th largest (counted with multiplicity) eigenvalue function for a positive definite matrices, and e_i a (normalized) associated eigenvector function.

Theorem 10 (Theorem 3.1 in Chapter 3)**.** We have

$$|\langle e_i(\mathbf{A}), e_j(\mathbf{B})\rangle| \leqslant C \times \sqrt{\frac{\min\left(\lambda_i(\mathbf{A}), \lambda_j(\mathbf{A})\right)}{\max\left(\lambda_i(\mathbf{A}), \lambda_j(\mathbf{A})\right)}}$$

where $C > 0$ is a positive number that only depends (polynomially) on μ, d and $\|v_i\|$, $i = 1, \dots, \mu$.

While this asymptotic rate is sufficient for our analysis, the estimation of the constant C that we find in Chapter 3 seems not be tight and could probably be improved.

6.4 *Chapter 4: Geometric ergodicity of Markov chains underlying CMA-ES*

After the irreducibility and aperiodicity, the next step in the convergence analysis of the normalized Markov chain (33) is the proof of geometrical ergodicity. To this aim, we rely in Chapter 4 on a state-dependent drift condition (see Theorem 1 above). However, proving that it holds for such a complex process is a very complicated task and requires additional assumptions. At this point of the proof, we consider only ellipsoidal objective functions, equivalently objective functions that are increasing transformations of convex-quadratic functions. Moreover, we restrict our analysis to specific normalization functions $R(\cdot)$, more precisely to transformations of the smallest eigenvalue.

Theorem 11 (Proposition 4.4 in Chapter 4)**.** Consider the normalized Markov chain $\{\theta_t\} = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ via (33) with an ellipsoidal objective function f which is an increasing transformation of $x \mapsto (x - x^*)^\top \mathbf{H}(x - x^*)$ for some matrix $\mathbf{H} \in \mathcal{S}^d_{++}$, and the normalization function $R \mapsto \mathbf{C} \in \mathcal{S}^d_{++} \mapsto \lambda_{\min}(\mathbf{H}^{1/2}\mathbf{C}\mathbf{H}^{1/2})$. Consider the potential function V defined by

$$V(z, p, q, \mathbf{\Sigma}, r) = \|\mathbf{H}^{1/2}z\|^2 + \beta\lambda_{\max}(\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}) + \gamma_p\|p\| + \gamma_q\|\mathbf{H}^{1/2}q\|^2 + \gamma_r r \qquad (37)$$

for some well-chosen constants $\beta, \gamma_p, \gamma_q, \gamma_r > 0$. Then, if the hyperparameters and the stepsize change function $\Gamma(\cdot)$ of CMA-ES are well-chosen, we have

$$\mathbb{E}\left[V(\theta_{n(\theta)}) \mid \theta_0 = \theta\right] \leqslant \rho \times V(\theta) \qquad (38)$$

for every θ outside a compact K, and where $\rho \in (0, 1)$ and $n(\theta) = 1$ or 2.

As in Chapter 2, we want to include different hyperparameter settings and thus we extend the state-dependent drift criterion for ergodicity to projected Markov chains. Indeed, we prove that if a redundant Markov chain satisfies a state-dependent drift condition outside a small set and that it possesses an irreducible aperiodic projected chain, then the latter is ergodic. We use this to prove directly that several algorithm variants (in particular when c_c or c_σ is equal to 1) are ergodic by analyzing only the chain (33).

Yet, we did not obtain the geometric ergodicity of the normalized chain when two evolution paths are used to update the stepsize and the covariance matrix.

Theorem 12 (Theorem 4.3 in Chapter 4)**.** Assume that the objective function f is ellipsoidal and is an increasing transformation of $x \mapsto (x - x^*)^\top \mathbf{H}(x - x^*)$ for some matrix $\mathbf{H} \in \mathcal{S}^d_{++}$. Consider the normalization function $R \mapsto \mathbf{C} \in \mathcal{S}^d_{++} \mapsto \lambda_{\min}(\mathbf{H}^{1/2}\mathbf{C}\mathbf{H}^{1/2})$ which defines the normalized Markov chain $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ via (33). Then, if the hyperparameters of CMA-ES are well-chosen:

(i) when $c_c \neq 1$ and $c_\sigma = 1$, the projected Markov chain $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ is geometrically ergodic;

(ii) when $c_c = 1$ and $c_\sigma \neq 1$, the projected Markov chain $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ is geometrically ergodic;

(iii) when $c_c = c_\sigma = 1$, the projected Markov chain $\{(z_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ is geometrically ergodic.

6.5 *Chapter 5: Linear convergence of CMA-ES on ellipsoidal problems and learning of second-order information*

Chapter 5 ends our convergence proof of CMA-ES. Since in Chapter 4 we have obtained the geometric ergodicity of the Markov chain (33), we obtain the linear behavior of the mean towards the optimum of the objective function when the latter is ellipsoidal. Moreover, when we have no cumulation on the stepsize (i.e., $c_\sigma = 1$), we can use the affine-invariance of CMA-ES and find that the convergence rate is the same when minimizing any ellipsoidal objective function and that the covariance matrix approximates the inverse Hessian of a convex-quadratic function. Besides, we prove that for a specific stepsize update, the convergence rate is positive, which proves that CMA-ES converges linearly.

Theorem 13 (Theorems 5.3 and 5.5 in Chapter 5)**.** Consider an ellipsoidal objective function f which is an increasing transformation of the convex-quadratic function $x \mapsto (x-x^*)^\top \mathbf{H}(x-x^*)$ for some matrix $\mathbf{H} \in \mathcal{S}^d_{++}$ and some point $x^* \in \mathbb{R}^d$. If the hyperparameters of CMA-ES are chosen correctly and the stepsize is updated via (9), then the mean m_t converges to x^* at a geometric rate, that is, there exists $C > 0$ and $\rho \in (0, 1)$ such that for every $t \in \mathbb{N}$:

$$\|m_t - x^*\| \leqslant C\rho^t \quad . \tag{39}$$

Moreover, the covariance matrix approximates $\mathbf{H}^{-1}$ in the sense that there exists $\alpha > 0$ such that:

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\mathbf{C}_t}{\det(\mathbf{C}_t)^{1/d}}\right] = \alpha\mathbf{H}^{-1} \quad . \tag{40}$$

6.6 *Appendix A: Evaluation of the impact of various modifications to CMA-ES that facilitate its theoretical analysis [49]*

We provide in Appendix A several experimental results that evaluate the impact of modifications of CMA-ES that have been or could have been used in the proof of convergence presented in this thesis.

We can find there the impact of cumulation on the stepsize and on the covariance matrix on the performances of the algorithm. This is particularly interesting in the light of our results, since we were able to prove the learning of the inverse Hessian, as well as that the convergence rate is the same for all ellipsoidal functions, only when we do not have cumulation on the stepsize.

We also analyzed the performances of the alternative stepsize update for which we were able to prove that the convergence rate is positive. We did not find sufficiently important differences that would suggest that this update is not relevant in practice. An other stepsize update that was included in this work is based on the cumulation path p^c_t, and modifies the CMA-ES algorithm to make it affine-invariant. However, we did not analyzed this variant in our theoretical works.

Chapter 1

On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds

Comments on Chapter 1: The content of this chapter is presented in the paper "On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds" (Armand Gissler, Alain Durmus, Anne Auger) submitted in 2024 to the journal *Annals of Applied Probability* [52]. The main result of this chapter is stated in Theorem 1.2. It generalizes previous works [29] that were useful to analyze the stability of Markov chains underlying evolution strategy (ES) algorithms. For instance, they were applied to a step-size adaptive ES [141]. However, they had two limitations. First, one of the functions that defines the update of the analyzed Markov chain was supposed to be continuously differentiable, excluding to the analysis for instance ES with a nonsmooth step-size update. Second, the state space of the Markov chain was required to be an open subset of an Euclidean space. In this chapter and in Chapter 2, we analyze Markov chains underlying CMA-ES which violate these two assumptions.

Abstract

In this paper, we analyze a large class of general nonlinear state-space models on a state space X, defined by the recursion $\phi_{k+1} = F(\phi_k, \alpha(\phi_k, U_{k+1}))$, $k \in \mathbb{N}$, where F, α are some functions and $\{U_{k+1}\}_{k\in\mathbb{N}}$ is a sequence of i.i.d. random variables. More precisely, we extend conditions under which this class of Markov chains is irreducible, aperiodic and satisfies important continuity properties, relaxing two key assumptions from prior works. First, the state space X is supposed to be a smooth manifold instead of an open subset of a Euclidean space. Second, we only suppose that F is locally Lipschitz continuous.

We demonstrate the significance of our results through their application to Markov chains underlying optimization algorithms. These schemes belong to the class of evolution strategies with covariance matrix adaptation and step-size adaptation.

Keywords: Markov chains, irreducibility, aperiodicity, T-chain, deterministic control model, CMA-ES.

1 Introduction 24
2 Main results 25
2.1 The model and assumptions *25*
2.2 Main results *32*
3 Applications 33
3.1 CMA-ES *33*
3.2 The step-size adaptive ES with nonsmooth update *34*
4 Proofs 35
4.1 Preliminary results *35*
4.2 Proofs of the main results: verifiable conditions for irreducibility and aperiodicity .. *40*
4.3 Proofs for the application to CMA-ES *46*
A Background on manifolds 48
B Clarke's generalized derivative of locally Lipschitz functions on manifolds 49
C Additional proofs 53

1 Introduction

Consider a nonlinear state-space model defined by the recursion:

$$\phi_{k+1} = G(\phi_k, \xi_{k+1}), \tag{1.1}$$

where the sequence $\{\xi_{k+1}\}_{k\in\mathbb{N}}$ consists of independent and identically distributed (i.i.d.) random variables, $G : \mathsf{X} \times \mathsf{W} \to \mathsf{X}$ is a continuous function, and X, W are two measurable spaces. Nonlinear state-space models (1.1) form a class of Markov chains that have been first popularized in stochastic control theory [111, 107, 108, 109]. This has spurred extensive analysis and has a well-established historical context. In particular, for nonlinear autoregressive models, i.e., where G can be written as $G(x, u) = \tilde{G}(x) + u$, ergodicity has been widely investigated [21, 10, 146, 112, 55]. Moreover, connections have been established between the stability of (1.1) and the one of some Ordinary Differential Equation (ODE) [81]. The idea of analyzing (1.1) from the perspective of control theory, where u is regarded as a control parameter, was initially proposed in [136] within the context of diffusion processes. This approach was subsequently employed with success in [82] and [98]. It has been then applied in [109, 110] when G is infinity differentiable and X, W are open sets of Euclidean spaces, to establish the irreducibility, aperiodicity and topological properties of the Markov kernel associated to (1.1).

Besides stochastic control models, (1.1) also encompasses many algorithms in optimization and Markov chain Monte Carlo algorithms. The variable ϕ_k corresponds to the state of an algorithm at iteration k and ξ_{k+1} represents the random components used to update this state. However, for certain classes of algorithms, especially those arising from zeroth-order optimization, the function G may not be continuous. Nevertheless, it can be written using the Markov chain model introduced in [29] as

$$\phi_{k+1} = F(\phi_k, \alpha(\phi_k, U_{k+1})), \tag{1.2}$$

for some continuous function $F : \mathsf{X} \times \mathsf{W} \to \mathsf{X}$ and a potentially discontinuous function $\alpha : \mathsf{X} \times \mathsf{U} \to \mathsf{W}$ for some measurable space U and $\{U_{k+1}\}_{k\in\mathbb{N}}$ a sequence of i.i.d. random variables valued in U. Extensions of the connection between the φ-irreducibility and the stability of the associated deterministic control model [110, Chapter 7] to Markov chains following (1.3) have been established in [29].

A particularly relevant algorithm of the form (1.2) in Evolution Strategies (ES) is ES with Covariance Matrix Adaptation (CMA-ES) [76, 73] often regarded as the state-of-the-art algorithm for numerical derivative-free optimization of difficult problems with tremendous applications in many domains (e.g., in biology [22, 129], medicine [121], machine learning [2, 60])[1]. Yet, while we have ample empirical evidences of its linear convergence on wide classes on functions, a convergence proof together with a convergence rate is still an open question. In order to extend linear convergence results from step-size adaptive ES [17, 141] to CMA-ES, a first step is to show the irreducibility and topological properties of the kernel associated to a normalized Markov chain underlying the algorithm. However, previous works [109, 29] cannot be applied since (i) the state space X of this chain is a smooth manifold whereas previous analysis supposed that they were open subsets of a Euclidean space, (ii) the function F is supposed to be continuously differentiable in existing results while certain step-size updates used in CMA-ES are only locally Lipschitz. The purpose of the present paper is to resolve these two limitations and pave the way to a complete convergence analysis of CMA-ES.

In this context, the objective of this paper is to extend these results [29] to the case where the sets X and W are smooth manifolds and where F is locally Lipschitz instead of continuously differentiable.

The paper is organized as follows. In Section 2.1, we provide a precise definition of the class of nonlinear state-space models under investigation. In Sections 2.1.1 and 2.1.2, we illustrate and motivate our analysis through two ES algorithms–one employing covariance matrix adaptation and the other utilizing step-size adaptation. Next, in Section 2.1.3, we outline the assumptions necessary for establishing our main results and deriving the irreducibility and aperiodicity of our model. Our main results are presented in Section 2.2 and are subsequently applied in Section 3 to the two algorithms introduced in Sections 2.1.1 and 2.1.2. Finally, proofs are gathered in Section 4. Note that some of the proofs and useful definitions are given in the appendix.

2 Main results

2.1 The model and assumptions

Let X, W be two (smooth, connected) manifolds (see Definition 1.2) of dimensions n and p respectively, endowed with their Borel σ-fields denoted by $\mathcal{B}(\mathsf{X})$ and $\mathcal{B}(\mathsf{W})$ respectively. We let $\mathrm{dist}_{\mathsf{X}}$ and $\mathrm{dist}_{\mathsf{W}}$ be two distance functions on X and W which induce the topology of X and W respectively. As a consequence of [100, Proposition 13.2, Theorem 13.29], such distance functions always exist.

We consider in this paper Markov chains taking values in X and associated with the general recursion

$$\phi_{k+1} = F(\phi_k, \alpha(\phi_k, U_{k+1})) \tag{1.3}$$

where $F : \mathsf{X} \times \mathsf{W} \to \mathsf{X}$ and $\alpha : \mathsf{X} \times \mathsf{U} \to \mathsf{W}$ are measurable functions, and $\{U_{k+1}\}_{k\in\mathbb{N}}$ is a sequence of i.i.d. random variables valued in a measurable space $(\mathsf{U}, \mathcal{U})$, chosen independently of the initial state ϕ_0. Throughout the paper, we denote by P the Markov kernel associated to (1.3). As emphasized in the introduction, this class of models is a natural extension of nonlinear state-space models defined on manifolds. To illustrate the interest of such processes, we provide the following examples.

[1] As of September 2023, the two main Python implementations of the CMA-ES algorithm cma and cmaes have more than 5 millions and 45 millions downloads respectively.

2.1.1 An instructive example: CMA-ES

We introduce here a simplified version of the numerical optimization algorithm called evolution strategy with covariance matrix adaptation (CMA-ES) [76, 73], which, for an objective function $f : \mathbb{R}^d \to \mathbb{R}$, aims to solve:

$$\text{find } x^* \in \operatorname*{Arg\,min}_{x\in\mathbb{R}^d} f(x)\,. \tag{P}$$

To this end, it approximates the optimum x^* of the objective function f by a multivariate normal distribution $\mathcal{N}(m_k, \mathbf{C}_k)$ for a mean $m_k \in \mathbb{R}^d$ and a covariance matrix $\mathbf{C}_k \in \mathcal{S}^d_{++}$ that are updated iteratively. More precisely, for each $k \in \mathbb{N}$, given $m_k \in \mathbb{R}^d$ and $\mathbf{C}_k \in \mathcal{S}^d_{++}$, the algorithm can be described as follows. First, a population of $\lambda \geqslant 2$ offspring is sampled using

$$U^1_{k+1}, \dots, U^\lambda_{k+1} \sim \mathcal{N}(0, I_d) \quad \text{i.i.d. and independently of } (m_k, \mathbf{C}_k), \tag{1.4}$$

so that, conditionally to $(m_k, \mathbf{C}_k)$, the offspring satisfy $m_k + \sqrt{\mathbf{C}_k}U^i_{k+1} \sim \mathcal{N}(m_k, \mathbf{C}_k)$, for $i = 1, \dots, \lambda$. Next, we rank the offspring so that we define a permutation $s_{k+1} \in \mathfrak{S}_\lambda$ of $\{1, \dots, \lambda\}$ satisfying

$$f\left(m_k + \sqrt{\mathbf{C}_k}U^{s_{k+1}(1)}_{k+1}\right) \leqslant \cdots \leqslant f\left(m_k + \sqrt{\mathbf{C}_k}U^{s_{k+1}(\lambda)}_{k+1}\right). \tag{1.5}$$

Then, given the $\mu \in \{1, \dots, \lambda\}$ best offspring, the mean is moved towards the best solutions with the following update

$$m_{k+1} = m_k + \sqrt{\mathbf{C}_k}\sum_{i=1}^{\mu} w_i U^{s_{k+1}(i)}_{k+1} \tag{1.6}$$

and the covariance matrix update reads

$$\mathbf{C}_{k+1} = (1-c)\mathbf{C}_k + c\sqrt{\mathbf{C}_k}\left(\sum_{i=1}^{\mu} w_i \left(U^{s_{k+1}(i)}_{k+1}\right)\left(U^{s_{k+1}(i)}_{k+1}\right)^\top\right)\sqrt{\mathbf{C}_k}. \tag{1.7}$$

It increases the likelihood to sample in the directions where good solutions were found. In the above equations, the weights $w_1 \geqslant \cdots \geqslant w_\mu > 0$ satisfy $\sum_{i=1}^{\mu} w_i = 1$, and we call $c \in (0, 1)$ the learning rate for the covariance matrix. In ES, the function values are not used explicitly to update the state variables. It influences the update only through the ranking of candidate solutions via the permutation s_{k+1}. Consequently, the algorithms are invariant with respect to strictly increasing transformations of the objective function (that preserve the ranking). In this context, a natural class of functions to analyze the convergence of ES are scaling-invariant functions [17, 142]. A function f is said to be scaling-invariant w.r.t. x^* if, for every $x, y \in \mathbb{R}^d$ and $\rho > 0$, we have

$$f(x + x^*) \leqslant f(y + x^*) \Leftrightarrow f(\rho x + x^*) \leqslant f(\rho y + x^*). \tag{1.8}$$

Convergence of step-size adaptive ES on scaling-invariant functions with smooth level sets was established—for specific assumptions on the step-size update—in previous works [141]. Assuming that the objective function f satisfies (1.8), we define then the following quantities

$$z_k = \frac{m_k - x^*}{\sqrt{R(\mathbf{C}_k)}} \quad ; \quad \mathbf{\Sigma}_k = \frac{\mathbf{C}_k}{R(\mathbf{C}_k)} \tag{1.9}$$

where $R = \det(\cdot)^{1/d} \colon \mathcal{S}^d_{++} \to \mathbb{R}_+$. We assume w.l.o.g. that $x^* = 0$. Then the sequence $\{(z_k, \mathbf{\Sigma}_k)\}_{k\in\mathbb{N}}$ defines a time-homogeneous Markov chain which obeys to the model (1.3), see Proposition 1.22,

with $\mathsf{X} = \mathbb{R}^d \times R^{-1}(\{1\})$, $\mathsf{U} = \mathbb{R}^{d\times\lambda}$, $\mathsf{W} = \mathbb{R}^{d\times\mu}$, and

$$\begin{aligned} F\colon\ & \mathsf{X}\times\mathsf{W}\to\mathsf{X} \\ & ((z,\boldsymbol{\Sigma}),(v_1,\dots,v_\mu)) \mapsto \left(\frac{z+\sqrt{\boldsymbol{\Sigma}}\sum_{i=1}^{\mu} w_i v_i}{R^{1/2}\left(\mathbf{K}(\boldsymbol{\Sigma},v_1,\dots,v_\mu)\right)}, \frac{\mathbf{K}(\boldsymbol{\Sigma},v_1,\dots,v_\mu)}{R\left(\mathbf{K}(\boldsymbol{\Sigma},v_1,\dots,v_\mu)\right)}\right) \end{aligned} \tag{1.10}$$

where

$$\mathbf{K}(\boldsymbol{\Sigma},v_1,\dots,v_\mu) = (1-c)\boldsymbol{\Sigma} + c\sqrt{\boldsymbol{\Sigma}}\left(\sum_{i=1}^{\mu} w_i v_i v_i^\top\right)\sqrt{\boldsymbol{\Sigma}},$$

and with

$$\begin{aligned} \alpha\colon\ & \mathsf{X}\times\mathsf{U}\to\mathsf{W} \\ & ((z,\boldsymbol{\Sigma}),(u_1,\dots,u_\lambda)) \mapsto \left(u_{s(1;z,\boldsymbol{\Sigma},u_{1:\lambda})},\dots,u_{s(\mu;z,\boldsymbol{\Sigma},u_{1:\lambda})}\right) \end{aligned} \tag{1.11}$$

where given $u_{1:\lambda} = (u_1,\dots,u_\lambda) \in (\mathbb{R}^d)^\lambda$, $z\in\mathbb{R}^d$ and $\boldsymbol{\Sigma}\in\mathcal{S}_{++}^d$, we denote by $s(\cdot;z,\boldsymbol{\Sigma},u_{1:\lambda})$ a permutation that sorts the $f(z+\sqrt{\boldsymbol{\Sigma}}u_i)$, $i=1,\dots,\lambda$. To ensure uniqueness of this permutation, we impose a tie-break, e.g., if $i<j$ are such that $f(z+\sqrt{\boldsymbol{\Sigma}}u_i) = f(z+\sqrt{\boldsymbol{\Sigma}}u_j)$, then $s(\cdot;z,\boldsymbol{\Sigma},u_{1:\lambda})^{-1}(i) < s(\cdot;z,\boldsymbol{\Sigma},u_{1:\lambda})^{-1}(j)$. Note that X is not an open subset of a Euclidean space (but it is a smooth manifold by the preimage theorem, see e.g. [59, Chapter 1, Section 4]), hence the results in [29] do not apply. We show in Section 3 that our results apply and we prove that $\{(z_k,\boldsymbol{\Sigma}_k)\}_{k\in\mathbb{N}}$ defines a φ-irreducible aperiodic T-chain and that all compact subsets of X are small.

If we establish moreover that the chain $\{(z_k,\boldsymbol{\Sigma}_k)\}_{k\in\mathbb{N}}$ is positive recurrent, then we obtain that CMA-ES behaves linearly, as stated below.

Theorem 1.1. Consider a scaling-invariant function with respect to x^* and the Markov chain $\{(z_k,\boldsymbol{\Sigma}_k)\}_{k\in\mathbb{N}}$ defined in (1.9) ensuing from CMA-ES minimizing f. Suppose that $\{(z_k,\boldsymbol{\Sigma}_k)\}_{k\in\mathbb{N}}$ is a φ-irreducible aperiodic positive recurrent chain with invariant probability measure π. If the function $(z,\boldsymbol{\Sigma})\mapsto\log\|z\|$ is π-integrable on $\mathbb{R}^d\times\mathcal{S}_{++}^d$, then almost surely we have

$$\lim_{k\to\infty}\frac{1}{k}\log\frac{\|m_k-x^*\|}{\|m_0-x^*\|} = -\mathrm{CR}\in\mathbb{R}. \tag{1.12}$$

When moreover $\mathrm{CR}>0$, we say that CMA-ES converges linearly to x^*.

Proof. Assume that $x^*=0$. Since CMA-ES is invariant by translation [15], (1.12) would generalize to any value of x^*. Since $\{(z_k,\boldsymbol{\Sigma}_k)\}_{k\in\mathbb{N}}$ is supposed to be φ-irreducible, aperiodic and positive recurrent, by [110, Theorem 17.0.1], we know that for all π-integrable function g, we have that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1} g(z_k,\boldsymbol{\Sigma}_k) = \int g(z,\boldsymbol{\Sigma})\mathrm{d}\pi(z,\boldsymbol{\Sigma}). \tag{1.13}$$

However, we have

$$\frac{1}{T}\log\frac{\|m_T\|}{\|m_0\|} = \frac{1}{T}\sum_{k=0}^{T-1}\left(\log\|m_{k+1}\| - \log\|m_k\|\right)$$

$$= \frac{1}{T}\sum_{k=0}^{T-1}\left(\log\|z_{k+1}\| - \log\|z_k\|\right) + \tag{1.14}$$

$$\frac{1}{2dT}\sum_{k=0}^{T-1}\log\det\underbrace{\left((1-c)\boldsymbol{\Sigma}_k + \sqrt{\boldsymbol{\Sigma}_k}\left(\sum_{i=1}^{\mu} w_i \left(U_{k+1}^{s_{k+1}(i)}\right)\left(U_{k+1}^{s_{k+1}(i)}\right)^\top\right)\sqrt{\boldsymbol{\Sigma}_k}\right)}_{=:\tilde{\boldsymbol{\Sigma}}_{k+1}}. \tag{1.15}$$

But, by assumption, $(z,\boldsymbol{\Sigma}) \mapsto \log\|z\|$ is π-integrable. Moreover, $\det(\boldsymbol{\Sigma}_k) = 1$, hence

$$\det\left(\tilde{\boldsymbol{\Sigma}}_{k+1}\right) = \det\left((1-c)I_d + c\sum_{i=1}^{\mu} w_i \left(U_{k+1}^{s_{k+1}(i)}\right)\left(U_{k+1}^{s_{k+1}(i)}\right)^\top\right).$$

Moreover,

$$1-c \leqslant \det\left((1-c)I_d + c\sum_{i=1}^{\mu} w_i \left(U_{k+1}^{s_{k+1}(i)}\right)\left(U_{k+1}^{s_{k+1}(i)}\right)^\top\right)^{1/d} \leqslant 1-c+c\max_{i=1,\dots,\mu}\|U_{k+1}^i\|^2$$

which defines an integrable quantity, since the vectors U_{k+1}^i, $k\in\mathbb{N}$, $i=1,\dots,\lambda$, are standard Gaussian vectors of $\mathbb{R}^d$. Applying (1.13) to (1.14) and (1.15), we find the stated result with

$$\mathrm{CR} = -\frac{1}{2d}\mathbb{E}_{(z,\boldsymbol{\Sigma})\sim\pi}\left[\det\left((1-c)I_d + c\sum_{i=1}^{\mu} w_i \left(U_1^{s(i;z,\boldsymbol{\Sigma},U_1^{1:\lambda})}\right)\left(U_1^{s(i;z,\boldsymbol{\Sigma},U_1^{1:\lambda})}\right)^\top\right)\right]. \tag{1.16}$$

□

The previous theorem illustrates how the φ-irreducibility and aperiodicity of $\{(z_k,\boldsymbol{\Sigma}_k)\}_{k\in\mathbb{N}}$ are instrumental to obtain linear convergence of CMA-ES.[2]

2.1.2 *A nonsmooth example: a step-size adaptive ES*

We present here an other simplification of CMA-ES where instead of adapting a full covariance matrix, a scaling factor called step-size is adapted such that the covariance matrix reads $\sigma_k^2 I_d$. In this step-size adaptive algorithm, the optimum $x^* \in \mathbb{R}^d$ of the problem (P) is approximated by a multivariate normal distribution $\mathcal{N}(m_k,\sigma_k^2 I_d)$, where the mean $m_k\in\mathbb{R}^d$ and the step-size $\sigma_k>0$ are updated as follows. For $k\in\mathbb{N}$, given a mean $m_k\in\mathbb{R}^d$ and a step-size $\sigma_k>0$, we sample $U_{k+1}^1,\dots,U_{k+1}^\lambda$, rank them by defining the permutation $s_{k+1}\in\mathfrak{S}_\lambda$ and update the mean m_{k+1} according to (1.4), (1.5),

[2]The variant of CMA-ES presented here differs significantly from the default CMA-ES (used in applications) where both step-size adaptation and covariance matrix adaptation are used. In addition, the covariance matrix update presents an additional mechanism (rank-one update). The combination of all the mechanisms is important to obtain fast convergence in many situations. This variant with however a learning rate on the mean update has been analyzed in previous theoretical works [8], and it has been proven to be a discretized version of a natural gradient update on the manifold of probability distributions [7].

(1.6), respectively, where we replace C_k by $\sigma_k^2 I_d$. The step-size update obeys

$$\sigma_{k+1} = \sigma_k \times \exp\left(\frac{1}{d_\sigma}\left(\frac{\sqrt{\mu_{\text{eff}}}\left\|\sum_{i=1}^{\mu} w_i U_{k+1}^{s_{k+1}(i)}\right\|}{\mathbb{E}\|\mathcal{N}(0, I_d)\|} - 1\right)\right) \tag{1.17}$$

where we define $\mu_{\text{eff}} = \sum_{i=1}^{\mu} w_i^2$ and fix $d_\sigma > 0$ (usually $d_\sigma \approx 1$). Moreover, as in Section 2.1.1, we assume f to be scaling-invariant, see (1.8). W.l.o.g. we suppose that f is scaling-invariant w.r.t. $x^* = 0$. Then, by defining

$$z_k = \frac{m_k - x^*}{\sigma_k}, \tag{1.18}$$

we get that the sequence $\{z_k\}_{k\in\mathbb{N}}$ is a time-homogeneous Markov chain which obeys to the model (1.3) (see [141, Proposition 4]) with $\mathsf{X} = \mathbb{R}^d$, $\mathsf{U} = \mathbb{R}^{d\times\lambda}$, $\mathsf{W} = \mathbb{R}^{d\times\mu}$,

$$\begin{aligned} F\colon\ & \mathsf{X}\times\mathsf{W} \to \mathsf{X} \\ & (z, (v_1, \dots, v_\mu)) \mapsto (z + \textstyle\sum_{i=1}^{\mu} w_i v_i) \times \exp\left(-\frac{1}{d_\sigma}\left(\frac{\sqrt{\mu_{\text{eff}}}\|\sum w_i v_i\|}{\mathbb{E}\|\mathcal{N}(0, I_d)\|} - 1\right)\right) \end{aligned} \tag{1.19}$$

and

$$\begin{aligned} \alpha\colon\ & \mathsf{X}\times\mathsf{U} \to \mathsf{W} \\ & (z, (u_1, \dots, u_\lambda)) \mapsto \left(u_{s(1;z,I_d,u_{1:\lambda})}, \dots, u_{s(\lambda;z,I_d,u_{1:\lambda})}\right) \end{aligned} \tag{1.20}$$

where we define the permutation $s(\cdot; z, I_d, u_{1:\lambda})$ as in Section 2.1.1. Here, F is not continuously differentiable, and we cannot use the results of [29] to analyze this chain. However the stability of an alternative strategy where (1.17) is replaced by a smooth update of the step-size has already been analyzed [141].

2.1.3 Assumptions

We consider the following assumptions on the functions F and α to establish ergodicity of the Markov kernel defined via (1.3):

H1. *For any $x \in \mathsf{X}$, the distribution μ_x of the random variable $\alpha(x, U_1)$ admits a density, denoted by p_x, with respect to a σ-finite measure ζ_{W}, such that:*

(i) *The function $(x, w) \mapsto p_x(w)$ is lower semicontinuous (l.s.c.), i.e., for any $(\bar{x}, \bar{w}) \in \mathsf{X}\times\mathsf{W}$, we have $\liminf p_x(w) \geqslant p_{\bar{x}}(\bar{w})$ when $(x, w) \to (\bar{x}, \bar{w})$.*

(ii) *For $\mathsf{A} \subset \mathcal{B}(\mathsf{W})$, $\zeta_{\mathsf{W}}(\mathsf{A}) = 0$ if and only if A is negligeable, i.e., $\text{Leb}(\varphi(\mathsf{A}\cap U)) = 0$ for any chart (φ, U) of W, where Leb stands for the Lebesgue measure.*

The condition **H**1 is a generalization of [29, A4], where $\mathsf{W} \subset \mathbb{R}^p$ was instead an open subset of an Euclidean space and ζ_{W} a Lebesgue measure. Besides, in that context, **H**1(i) has already been considered in [29]. If W is equipped with a smooth Riemannian metric which makes W a Riemannian manifold, a σ-finite measure satisfying **H**1(ii) would be the Lebesgue-Riemann volume measure [9, Chapter XII and Proposition XII.1.6].

We assume moreover the following on the map $F\colon \mathsf{X}\times\mathsf{W} \to \mathsf{X}$.

H2. *The map $F\colon \mathsf{X}\times\mathsf{W} \to \mathsf{X}$ is locally Lipschitz, see Definition 1.4, on $\mathsf{X}\times\mathsf{W}$ with respect to the distance $\text{dist}_{\mathsf{X}} \oplus \text{dist}_{\mathsf{W}}$, defined by $\text{dist}_{\mathsf{X}} \oplus \text{dist}_{\mathsf{W}}((x, w), (x', w')) = \text{dist}_{\mathsf{X}}(x, x') + \text{dist}_{\mathsf{W}}(w, w')$ for every $((x, w), (x', w')) \in (\mathsf{X}\times\mathsf{W})^2$.*

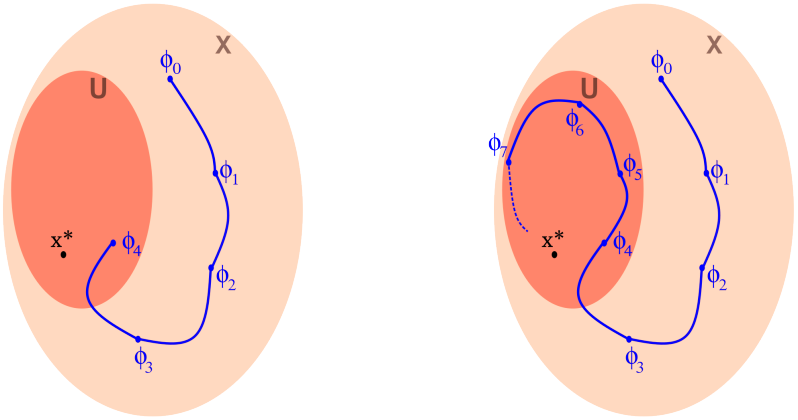


Figure 1.1: Left: Illustration of a globally attracting state x^*, for any neighborhood U of x^* and any starting state ϕ_0, there exists a k-steps path from ϕ_0 to U.
Right: Illustration of a steadily attracting state x^*, for any neighborhood U of x^* and any starting state ϕ_0, there exist $T > 0$ and k-steps paths from ϕ_0 to U for every $k \geqslant T$.

This assumption encompasses the continuous differentiability assumption made in [29], in the case where X and W are open subsets of some Euclidean spaces. Indeed, any continuously differentiable function is in particular locally Lipschitz. For example, the update (1.19), introduced in Section 2.1.2, satisfies the assumption **H**2.

For our last assumption regarding the functions F and α, we need to introduce further notations and notions introduced in [110, 29]. The *extended transition map* $S_x^k \colon \mathsf{W}^k \to \mathsf{X}$ can be defined inductively via

$$S_x^{k+1}(w_{1:k+1}) \coloneqq F(S_x^k(w_{1:k}), w_{k+1}), \quad S_x^0 \coloneqq x, \tag{1.21}$$

for $k \in \mathbb{N}$, $x \in \mathsf{X}$ and $w_{1:k+1} = (w_1, \dots, w_{k+1}) \in \mathsf{W}^{k+1}$. The value $S_x^k(w_{1:k})$ corresponds to the k^{th} iterate of the chain ϕ_k defined via (1.3), conditionally to $\phi_0 = x$ and $\alpha(\phi_t, U_{t+1}) = w_t$ for $t = 0, \dots, k-1$. Remark that, by composition, if F is continuous(ly locally Lipschitz), then so is $(x, w_{1:k}) \mapsto S_x^k(w_{1:k})$. Similarly, we define the extended probability density p_x^k via

$$p_x^{k+1}(w_{1:k+1}) \coloneqq p_x^k(w_{1:k}) p_{S_x^k(w_{1:k})}(w_{k+1}), \quad p_x^1(w_1) \coloneqq p_x(w_1). \tag{1.22}$$

The function p_x^k is then the density of the random variable $(\alpha(\phi_0, U_1), \dots, \alpha(\phi_{k-1}, U_k))$, with $\phi_0 = x$, w.r.t. the product measure $\zeta_{\mathsf{W}}^{\otimes k}$. If $(x, w) \mapsto p_x(w)$ is l.s.c., then $(x, w_{1:k}) \mapsto p_x^k(w_{1:k})$ is l.s.c. as well. In this case, the control sets

$$\mathcal{O}_x^k \coloneqq \{w_{1:k} \in \mathsf{W}^k \mid p_x^k(w_{1:k}) > 0\} \tag{1.23}$$

are nonempty open subsets of W^k. The control set $\mathcal{O}_x^k$ corresponds to the set of paths $w_{1:k}$ starting at x which have positive density $p_x^k(w_{1:k})$.

Moreover, for $x \in \mathsf{X}$, A a measurable subset of X, and $k > 0$, we say that $w_{1:k} \in \mathsf{W}^k$ is a *k-steps path* from x to A if $w_{1:k} \in \mathcal{O}_x^k$ and $S_x^k(w_{1:k}) \in A$, implying that A is then reachable by P from x. A point $x^* \in \mathsf{X}$ is said to be a *globally attracting state* if for any $y \in \mathsf{X}$ and any neighborhood U of x^*, there exist $k > 0$ and a k-steps path between y and U (the original definition of a globally attracting state is actually given in (1.35) and we show in Proposition 1.6 the equivalence with this

latter statement). It is said to be *steadily attracting* if for any $y \in \mathsf{X}$ and any neighborhood U of x^*, there exists $T > 0$ such that for every $k \geqslant T$, we can find a k-steps path between y and U. Note that any steadily attracting state is in particular globally attracting. These two notions are illustrated in Figure 1.1. Assuming **H**1 and **H**2, as emphasized in Theorem 1.5, we show that the kernel P defined via (1.3) is φ-irreducible exhibiting the existence of a globally attracting state (it is in fact an equivalence). On a related note, we deduce in Theorem 1.7 that the existence of a steadily attracting state is equivalent to the φ-irreducibility and aperiodicity of P.

We introduce now the notation ∂f for the *Clarke's generalized Jacobian* of a locally Lipschitz function $f : \mathsf{X} \to \mathsf{Y}$ between two manifolds X and Y. These Jacobians have been defined in [30], and we recall the definition in the Euclidean case in Definition 1.5. For the sake of completeness, we define here and give basic properties in Section B of the Clarke's Jacobian for functions defined on manifolds.

Proposition and Definition 1 (Clarke's generalized Jacobian on manifolds)**.** Let X and Y be two manifolds and $f : \mathsf{X} \to \mathsf{Y}$ be locally Lipschitz at $x_0 \in \mathsf{X}$. Let (φ, U) be a local chart of X around x_0 and (ψ, V) be a local chart of Y around $f(x_0)$. Define $g = \psi \circ f \circ \varphi^{-1}$. Then $g : \varphi(U) \to \psi(V)$ is locally Lipschitz at $\varphi(x_0)$, and we can define

$$\partial f(x_0) = \left\{ \mathcal{D}\psi^{-1}(g \circ \varphi(x_0)) \circ h \circ \mathcal{D}\varphi(x_0) \mid h \in \partial g(\varphi(x_0)) \right\}, \tag{1.24}$$

where $\mathcal{D}$ denotes the usual differential operator, and ∂ the Clarke differential operator. This definition does not depend on the choice of the charts (φ, U) and (ψ, V).

Proof. See Section B. □

In the case of a differentiable function f, the definition of Clarke's generalized Jacobian corresponds to the definition of the Jacobian, i.e., $\partial f(x) = \{\mathcal{D}f(x)\}$. The notion of Clarke's generalized Jacobian is used to formulate the controllability condition for an element $x \in \mathsf{X}$:

$$\text{there exists } w_{1:k} \in \overline{\mathcal{O}_x^k} \text{ such that } \partial S_x^k(w_{1:k}) \text{ is of maximal rank.} \tag{C_x}$$

Note that here, $\partial S_x^k(w_{1:k})$ is of maximal rank, is understood as any element of $\partial S_x^k(w_{1:k})$ is of rank n, the dimension of X. In Corollary 1.3, we show that P is a T-chain assuming that condition (C_x) holds for every state $x \in \mathsf{X}$. In comparison to **H**3 below, we do not assume that states x for which (C_x) holds are globally attracting. However, we show that if (C_{x^*}) holds for x^* a globally attracting state, then it holds for every state in X, see Proposition 1.11.

H3. *The controllability condition* (C_{x^*}) *is satisfied for a globally attracting state* x^*.

Alternatively, if we want to prove aperidocity on top of φ-irreducibility, we assume instead the following.

H4. *The controllability condition* (C_{x^*}) *is satisfied for a steadily attracting state* x^*.

Remark that **H**4 implies **H**3. Assumptions **H**3-**H**4 also appear in [29] but with the additional condition that the functions S_x^k are continuously differentiable for $x \in \mathsf{X}$ and $k > 0$, condition that we relax here. Globally and steadily attracting states are characterized by Proposition 1.6 and Proposition 1.8(ii) respectively below. In Section 3, we give one example of a smooth model on manifolds, and one example of a nonsmooth model on a Euclidean space, for which we show that **H**4 holds.

2.2 *Main results*

Before stating our main results, we introduce concepts that are needed for their statements. Given P a Markov kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$, we define $P^1 = P$ and for $k \geqslant 1$, $x \in \mathsf{X}$ and $\mathsf{A} \in \mathcal{B}(\mathsf{X})$, $P^{k+1}(x, \mathsf{A}) = \int P(y, \mathsf{A}) P^k(x, \mathrm{d}y)$. We say that P is φ-irreducible when there exists a nontrivial measure φ on $\mathcal{B}(\mathsf{X})$ such that for any $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ with $\varphi(\mathsf{A}) > 0$, we have $\sum_{k \geqslant 1} P^k(x, \mathsf{A}) > 0$ for every $x \in \mathsf{X}$. Let b be a probability distribution on $\mathbb{N}$, and let K_b be the transition kernel defined by $K_b(x, \mathsf{A}) := \sum_{k \geqslant 0} b(k) P^k(x, \mathsf{A})$. A substochastic transition kernel T with $K_b \geqslant T$ such that $x \mapsto T(x, \mathsf{A})$ is lower semicontinuous for every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ is called a *continuous component* of K_b. If P admits a distribution b such that there exists a continuous component T of K_b with $T(\cdot, \mathsf{X}) > 0$, then P is called a *T-chain*.

A set $C \in \mathcal{B}(\mathsf{X})$ is called *petite* if there exist a probability distribution b on $\mathbb{N}$ and a nontrivial measure ν_b on $\mathcal{B}(\mathsf{X})$ such that $K_b(x, \mathsf{A}) \geqslant \nu_b(\mathsf{A})$ for every $x \in \mathsf{X}$ and $\mathsf{A} \in \mathcal{B}(\mathsf{X})$. If moreover $b = \delta_a$ the Dirac distribution at some $a \in \mathbb{N}$, then C is called a-small.

If P is φ-irreducible, then the family $(D_i)_{i=1,\dots,d} \in \mathcal{B}(\mathsf{X})^d$ is called a d-*cycle* when

$$\begin{cases} P(x, D_{i+1}) = 1 \text{ for } x \in D_i \text{ and } i = 0, \dots, d-1 \bmod d \\ \varphi((\cup_{1 \leqslant i \leqslant d} D_i)^c) = 0 \text{ for any irreducibility measure } \varphi. \end{cases} \tag{1.25}$$

By [110, Theorem 5.4.4 and Proposition 5.2.4], if P is φ-irreducible, then there exist $d \geqslant 1$ and a d-cycle. The *period* of P is the largest integer d for which there exists a d-cycle. If the period of P is equal to 1, then P is said to be *aperiodic*.

If P is φ-irreducible, then P is said to be recurrent when for every $A \in \mathcal{B}(\mathsf{X})$ such that $\varphi(A) > 0$ and for every $x \in A$, we have $\sum_{k=1}^{\infty} P^k(x, A) = +\infty$. We say that P is positive when it admits an invariant probability measure. In practice, in order to show that P is positive recurrent, it is sufficient to establish a Foster-Lyapunov condition, see e.g., [110, Theorem 15.0.1], i.e., the existence of a function $V : \mathsf{X} \to [1, +\infty]$ finite at least at one point of X, of a petite set C and of constants $b < \infty$ and $\rho \in (0, 1)$, such that, for any $x \in \mathsf{X}$, we have

$$\int V(y) P(x, \mathrm{d}y) \leqslant \rho V(x) + b \mathbb{1}\{x \in C\}. \tag{1.26}$$

We have now all the tools to state our main contribution.

Theorem 1.2. Assume **H**1-**H**2 and **H**3. Then, the Markov kernel P defined via (1.3) is a φ-irreducible T-chain, and any compact set is petite. If moreover **H**4 holds, then, the Markov kernel P is aperiodic, and any compact set is small. In addition, if P is positive recurrent, then $\{\phi_k\}_{k \in \mathbb{N}}$ is ergodic, i.e., P admits a unique stationary distribution π and for π-almost every $x \in \mathsf{X}$,

$$\lim_{k \to +\infty} \|\delta_x P^k - \pi\|_{\mathrm{TV}} = 0 \,. \tag{1.27}$$

In addition to the assumptions of Theorem 1.2, if we suppose that P is Harris recurrent, then a Law of Large Numbers holds, see [110, Theorem 17.0.1]. For any π-integrable function g, a Markov chain $\{\phi_k\}_{k \in \mathbb{N}}$ associated to the kernel P satisfies

$$\lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} g(\phi_k) = \int g \, \mathrm{d}\pi. \tag{1.28}$$

In particular, a Foster-Lyapunov condition (1.26) implies Harris recurrence.

The proof of Theorem 1.2 is postponed to Section 4.2. It relies on intermediary results that to a great extent are generalizations of results in [29]. In particular, Propositions 1.6 to 1.8 characterize globally attracting states, reachable states and steadily attracting states respectively. Propositions 1.10, 1.11, 1.13 and 1.14 provide consequences of the assumption of controllability (C_x). Lemma 1.1 is a generalization of [109, Lemma 3.0], which turns out to be useful to prove that the controllability condition (C_x) implies that the Markov kernel P is a T-chain, as stated in Proposition 1.15 and Corollary 1.3. Proposition 1.16 characterizes the support of the irreducibility measures of P, while Theorems 1.5 to 1.8 end the proof of Theorem 1.2.

In contrast to [29], we assume here F to be locally Lipschitz instead of continuously differentiable. This changes the assumption of controllability (C_x), which consists here of a maximal rank condition for every element in the Clarke's generalized Jacobian of the extended transition map at some point, instead of a maximal rank condition of the Jacobian. Furthermore, we assume the sets X and W to be manifolds instead of open subsets of Euclidean spaces. Hence, while in [29] the Jacobian matrix of the extended transition map can be identified to a rectangular matrix, here, the Clarke's generalized Jacobian consists in a set of linear applications between tangent spaces.

3 Applications

3.1 CMA-ES

We consider in this section the process $\{(z_k, \mathbf{\Sigma}_k)\}_{k\in\mathbb{N}}$ defined in Section 2.1.1. It defines a time-homogeneous Markov chain when the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ is scaling-invariant, see (1.8). As observed in Section 2.1.1, the Markov chain $\{(z_k, \mathbf{\Sigma}_k)\}_{k\in\mathbb{N}}$ obeys (1.3), with F and α defined by (1.10) and (1.11) respectively.

Besides, we assume in this section that f has Lebesgue-negligible level sets, i.e., $\mathrm{Leb}(\mathcal{L}_t) = 0$, with

$$\mathcal{L}_t \coloneqq \{x \in \mathbb{R}^d \mid f(x) = t\} \quad \text{for } t \in \mathbb{R}. \tag{1.29}$$

Stability of Markov chains defined in the context of ES with step-size adaptation has been proven [17, 141], yielding to linear convergence. We complement these results applying now Theorem 1.2 to show the stability of $\{(z_k, \mathbf{\Sigma}_k)\}_{k\in\mathbb{N}}$. First, observe that the assumption **H**2 is automatically satisfied, since F is continuously differentiable. As for **H**1, we use the following result.

Proposition 1.1. Suppose that f has Lebesgue-negligible level sets. Define for any $\theta = (z, \mathbf{\Sigma}) \in \mathsf{X}$ and $v = (v_1, \dots, v_\mu) \in \mathsf{W}$,

$$p_\theta(v) = \frac{\lambda!}{(\lambda-\mu)!}\mathbb{1}\left\{f\left(z+\sqrt{\mathbf{\Sigma}}v_1\right) < \cdots < f\left(z+\sqrt{\mathbf{\Sigma}}v_\mu\right)\right\} \times \left(1 - Q^f_\theta(v_\mu)\right)^{\lambda-\mu} \gamma^d(v_1)\dots\gamma^d(v_\mu) \tag{1.30}$$

with $Q^f_\theta(u) = \int \mathbb{1}\{f(z+\sqrt{\mathbf{\Sigma}}\xi) < f(z+\sqrt{\mathbf{\Sigma}}u)\}\gamma^d(\xi)\mathrm{d}\xi$ and where γ^d is the density of the d-dimensional standard normal distribution w.r.t. Lebesgue. Then, p_θ defines a density (w.r.t. Lebesgue in $\mathbb{R}^{d\mu}$) of the random variable $\alpha(\theta, U_1)$.

If f has Lebesgue-negligible level sets and is continuous, it follows that **H**1 holds.

The proof of Proposition 1.1 mimics the one of [29, Proposition 5.2], but is given for completeness in Section C. Then, it remains to prove **H**4 and in particular to find a steadily attracting state

$\theta^* = (z^*, \mathbf{\Sigma}^*)$ for which there exist $k > 0$ and $v^*_{1:k} \in \overline{\mathcal{O}^k_{\theta^*}}$ such that $\partial S^k_{\theta^*}(v^*_{1:k})$ is of maximal rank. This is achieved in the following proposition proven in Section 4.3.

Proposition 1.2. Suppose that f is continuous, scaling-invariant with Lebesgue-negligible level sets. Then,

(i) the state $\theta^* = (0, I_d)$ is steadily attracting ;

(ii) there exists $k > 0$ and $v^*_{1:k} \in \overline{\mathcal{O}^k_{\theta^*}}$ such that $\mathcal{D}S^k_{\theta^*}(v^*_{1:k})\colon \mathsf{W}^k \to T_{S^k_{\theta^*}(v^*_{1:k})}\mathsf{X}$ is surjective, hence is full rank, where

$$T_{S^k_{\theta^*}(v^*_{1:k})}\mathsf{X} = \mathbb{R}^d \times \ker\left(\mathcal{D}\det(I_d)\right),$$

where $\det$ is the determinant map on the set of symmetric matrices $\mathcal{S}^d$, and $\ker$ denotes the kernel of a linear application.

Then, by applying Theorem 1.2, the φ-irreducibility and aperiodicity of the chain $\{\theta_k\}_{k\in\mathbb{N}}$ follow.

Theorem 1.3. Suppose that f is continuous, scaling-invariant with Lebesgue-negligible level sets. Then the Markov chain $\{\theta_k\}_{k\in\mathbb{N}}$ defines a time-homogeneous φ-irreducible aperiodic T-chain, for which any compact subset of X is small.

3.2 The step-size adaptive ES with nonsmooth update

Here, we consider the process $\{z_k\}_{k\in\mathbb{N}}$ defined in Section 2.1.2. As for CMA-ES, if f is supposed to be scaling-invariant, then this sequence defines a time-homogeneous Markov chain. In Section 2.1.2, we have established that this chain follows the model (1.3). Like CMA-ES, the following proposition gives a sufficient condition for assumption **H**1 to hold. The proof goes as for Proposition 1.1, which can be found in Section C.

Proposition 1.3. Suppose that f has Lebesgue-negligible level sets. Define for all $z \in \mathsf{X}$ and $v = (v_1, \ldots, v_\mu) \in \mathsf{W}$

$$p_z(v) = \frac{\lambda!}{(\lambda-\mu)!}\mathbb{1}\left\{f\left(z+v_1\right) < \cdots < f\left(z+v_\mu\right)\right\}\left(1 - Q^f_z(v_\mu)\right)^{\lambda-\mu}\gamma^d(v_1)\ldots\gamma^d(v_\mu) \tag{1.31}$$

with $Q^f_z(u) = \int \mathbb{1}\{f(z+\xi) < f(z+u)\}\gamma^d(\xi)\mathrm{d}\xi$ and where γ^d is the density of the d-dimensionnal standard normal distribution w.r.t. Lebesgue. Then, p_z defines a density (w.r.t. Lebesgue in $\mathbb{R}^{d\mu}$) of the random variable $\alpha(z, U_1)$. Moreover, if f is (a monotone transformation of) a continuous function, then $(z, v) \mapsto p_z(v)$ is l.s.c.

As for CMA-ES, assumption **H**2 holds since F, given in (1.19), is the composition of a continuously differentiable function with the Lipschitz function $x \mapsto \|x\|$. Regarding **H**4, the next proposition states the existence of a steadily attracting state. The proof follows the same lines as [29, Proposition 5.3], but is given for completeness.

Proposition 1.4. Suppose that f is continuous, scaling-invariant with Lebesgue-negligible level sets. Then, 0 is a steadily attracting state.

Proof. For $z_0 \in \mathbb{R}^d$, we set $v_1 = -[z_0, \dots, z_0] \in \mathbb{R}^{d\mu}$, and $v_k = [0, \dots, 0] \in \mathbb{R}^{d\mu}$. Note that, by Proposition 1.3, since f has Lebesgue-negligible level sets, $v_{1:k} \in \overline{\mathcal{O}^k_{z_0}}$. Moreover, we have $S^k_{z_0}(v_{1:k}) = 0$ for every $k \geqslant 1$, where $S^k_{z_0}$ is defined in (1.21). We conclude the proof by using Corollary 1.2 . □

To complete the verification of **H**4, we show in the next proposition that there exists $v_1 \in \overline{\mathcal{O}^1_0}$ such that S^1_0 is differentiable in v_1 and $\mathcal{D}S^1_0(v_1)$ is of maximal rank.

Proposition 1.5. Suppose that f is continuous, scaling-invariant with Lebesgue-negligible level sets. Then, S^1_0 is differentiable in $v_1 = (0, \dots, 0) \in \overline{\mathcal{O}^1_0}$ and $\mathcal{D}S^1_0(v_1)$ is of maximal rank.

Proof. Note that, by Proposition 1.3, v_1 belongs to $\overline{\mathcal{O}^1_0}$. Moreover, for $h = (h_1, \dots, h_\mu) \in \mathsf{W}$, we have by definition of F and of S^1_0, see (1.19) and (1.21) respectively, that

$$S^1_0(v_1 + h) = F(0, h) = \exp\left(\frac{1}{d_\sigma}\left(1 - \frac{\sqrt{\mu_{\text{eff}}}\|\sum_{i=1}^\mu w_i h_i\|}{\mathbb{E}\|\mathcal{N}(0, I_d)\|}\right)\right) \times \sum_{i=1}^\mu w_i h_i.$$

A simple Taylor expansion shows that

$$\lim_{h\to 0} \frac{\left\|S^1_0(v_1 + h) - S^1_0(v_1) - \exp\left(\frac{1}{d_\sigma}\right) \times \sum_{i=1}^\mu w_i h_i\right\|}{\|h\|} = 0, \tag{1.32}$$

ending the proof. □

Using Theorem 1.2, we deduce the φ-irreducibility and aperiodicity of the chain $\{z_k\}_{k\in\mathbb{N}}$.

Theorem 1.4. Suppose that f is continuous, scaling-invariant with Lebesgue-negligible level sets. Then, the Markov chain $\{z_k\}_{k\in\mathbb{N}}$ defines a time-homogeneous φ-irreducible aperiodic T-chain, for which compact subsets of X are small.

Note that in [141], it has been proven that the chain $\{z_k\}_{k\in\mathbb{N}}$ is φ-irreducible, aperiodic and positive recurrent, on the condition that the step-size obeys to a smooth update instead of (1.17). However, a smooth step-size update was required only to prove the φ-irreducibility and aperiodicity of the chain, since the derivation of these two results rely in [141] on results in [29]. Now that we have proven that the chain $\{z_k\}_{k\in\mathbb{N}}$ is φ-irreducible and aperiodic even when the step-size update is nonsmooth, we can prove that it is positive recurrent following the proofs of [141].

4 Proofs

4.1 Preliminary results

4.1.1 Accessibility, attracting and attainable states

In this section, we generalize characterizations of globally and steadily attracting states developed in [29]. In contrast to this reference, we relax assumptions on the sets X, U and W. Indeed, [29] supposed that these sets were open subsets of Euclidean spaces. Here, we only suppose that they are smooth connected manifolds, as formalized in Section 2.1. The proofs in this section are almost identical to those of [29], but are given for completeness in Section C.

For the rest of the paper, let us define $A_+^0(x) \coloneqq \{x\}$ and

$$A_+^k(x) \coloneqq \{S_x^k(w_{1:k}) \mid w_{1:k} \in \mathcal{O}_x^k\} \quad \text{for } k \geqslant 1. \tag{1.33}$$

The set $A_+^k(x)$ is the set of states that can be reached by ϕ_k conditionally to $\phi_0 = x$. We also define the set of states that can be reached by $\{\phi_k\}_{k\in\mathbb{N}}$ (in finite time) conditionally to $\phi_0 = x$ as

$$A_+(x) \coloneqq \bigcup_{k\in\mathbb{N}} A_+^k(x). \tag{1.34}$$

Then, we say that the control model associated to (1.3) is *forward accessible* if for every $x \in \mathsf{X}$, $A_+(x)$ has a nonempty interior in X. Moreover, with these notations, a point $x^* \in \mathsf{X}$ is a *globally attracting state*, if for every $y \in \mathsf{X}$ we have

$$x^* \in \bigcap_{T\geqslant 1} \overline{\bigcup_{k\geqslant T} A_+^k(y)}. \tag{1.35}$$

As shown in the next proposition, this definition is equivalent to the statement we used in Section 2.1 to introduce a globally attracting state that for any $y \in \mathsf{X}$ and any neighborhood U of x^*, there exists $k > 0$ and a k-steps path between y and U.

Proposition 1.6 (Characterization of globally attracting states)**.** Suppose **H**1. A point $x^* \in \mathsf{X}$ is globally attracting if and only if one of the following equivalent conditions holds.

(i) For any $y \in \mathsf{X}$, $x^* \in \overline{A_+(y)}$.

(ii) For any $y \in \mathsf{X}$ and any open subset U of X containing x^*, there exist $k > 0$ and a k-steps path from y to U.

(iii) For any $y \in \mathsf{X}$, there exists a sequence $\{y_k\}_{k>0}$ with $y_k \in A_+^k(y)$, from which we can extract a subsequence converging to x^*.

A point $x \in \mathsf{X}$ is said to be *reachable* by P [110, Section 6.1.2] if for any measurable neighborhood U of x in X, we have

$$\forall y \in \mathsf{X}, \quad \sum_{k\geqslant 1} P^k(y, U) > 0. \tag{1.36}$$

The equivalence between globally attracting states and reachable states relies on the following proposition.

Proposition 1.7 (Characterization of reachable states)**.** Consider the Markov kernel P defined via Eq. (1.3), and suppose **H**1 and that F is continuous. Then for any open subset U of X, any $x \in \mathsf{X}$ and $k > 0$, the following statements are equivalent.

(i) There exists a k-steps path from x to U.

(ii) $P^k(x, U) > 0$.

As an immediate consequence of Propositions 1.6 and 1.7, we get the following equivalence between states that are globally attracting by the control model associated to (1.3) and states that are reachable by P.

Corollary 1.1. Consider the Markov kernel P defined via (1.3), and suppose **H**1 and that F is continuous. Then $x \in \mathsf{X}$ is globally attracting if and only if it is reachable by P.

Recall that a state $x^* \in \mathsf{X}$ is *steadily attracting* [29] if for all $y \in \mathsf{X}$ and all open neighborhood U of x^* in X, there exists $T > 0$ such that for all $k \geqslant T$ there exists a k-steps path from y to U.

We now state two technical results related to steadily attracting states, which will be instrumental in the proofs of our main results. The first one is the equivalent of [29, Proposition 3.3].

Proposition 1.8. Suppose **H**1. The following statements hold.

(i) If $x^* \in \mathsf{X}$ is steadily attracting, then it is globally attracting.

(ii) A state $x^* \in \mathsf{X}$ is steadily attracting if and only if for every $y \in \mathsf{X}$ we can find a sequence $\{y_k\}_{k>0}$ with $y_k \in \overline{A_+^k(y)}$, which converges to x^*.

(iii) Assume F is continuous. If there exists a steadily attracting state, then every globally attracting state is steadily attracting.

Note however that the statement of [29, Proposition 3.3 (ii)] is slightly different as the element y_k belongs to $A_+^k(y)$ while in (ii) above y_k belongs to $\overline{A_+^k(y)}$. It is easy to see that both statements are equivalent.

In addition, we give the following corollary of Proposition 1.8 when F is assumed continuous.

Corollary 1.2. Suppose **H**1 and that F is continuous. Then, for any $x \in \mathsf{X}$ and $k \in \mathbb{N}$ we have the inclusion $\{S_x^k(w_{1:k}) \mid w_{1:k} \in \overline{\mathcal{O}_x^k}\} \subset \overline{A_+^k(x)}$. Consequently the following statements are equivalent.

(i) The point x^* is steadily attracting.

(ii) For every $x \in \mathsf{X}$ we can find a sequence $\{y_k\}_{k>0}$ satisfying $y_k \in \{S_x^k(w_{1:k}) \mid w_{1:k} \in \overline{\mathcal{O}_x^k}\}$ for any $k \geqslant 1$, and which converges to x^*.

(iii) For every $x \in \mathsf{X}$, for every neighborhood U of x^*, there exists $T > 0$ such that for any $k \geqslant T$ we can find $w_{1:k} \in \overline{\mathcal{O}_x^k}$ satisfying $S_x^k(w_{1:k}) \in U$.

The next result turns out to be useful later in order to prove the aperiodicity of the Markov kernel P, given that it is φ-irreducible. To this end, we need to introduce the notion of attainability, as considered in [110]. We say that a state $x^* \in \mathsf{X}$ is *attainable* if

$$\forall y \in \mathsf{X}, \quad x^* \in A_+(y). \tag{1.37}$$

Proposition 1.9. Consider the Markov kernel P defined via (1.3), and suppose **H**1. Let $x^* \in \mathsf{X}$ be attainable, and set

$$E \coloneqq \{a \in \mathbb{N}^* \mid \exists T \in \mathbb{N}, \forall k \geqslant T, x^* \in A_+^{ak}(x^*)\}. \tag{1.38}$$

Then, the following statements hold.

(i) E is nonempty and for every $a, b \in E$, the greatest common divider of a and b satisfies $\gcd(a, b) \in E$.

(ii) If $\gcd(E) = \max\{c \in \mathbb{N} \mid c \text{ divides } a, \forall a \in E\} = 1$, then x^* is steadily attracting.

(iii) If P is φ-irreducible, then there exists a d-cycle (as defined in (1.25)) with $d = \gcd(E)$.

4.1.2 *Controllability condition*

In this section, we generalize results of [29], more precisely [29, Propositions 3.5, 3.6 and 3.7], on the consequences of condition (C_x). The main challenge here is to deal with the condition that F is supposed to be locally Lipschitz only.

First, we generalize [29, Proposition 3.5] and prove that if the controllability condition (C_{x^*}) is satisfied for some x^* a globally attracting state, then is satisfied for every $y \in \mathsf{X}$.

Proposition 1.10. Suppose **H**1 and **H**2. Let $x^* \in \mathsf{X}$ be a globally attracting state. If (C_{x^*}) holds, then for every $y \in \mathsf{X}$, (C_y) holds.

Proof. By (C_{x^*}), there exist $k > 0$ and $w^*_{1:k} \in \overline{\mathcal{O}^k_{x^*}}$ such that $\partial S^k_{x^*}(w^*_{1:k})$ is of rank n, the dimension of X. See that, by Proposition 1.20, we can assume that $w^*_{1:k} \in \mathcal{O}^k_{x^*}$. Moreover, the function $S^k : (z, w_{1:k}) \mapsto S^k_z(w_{1:k})$ is locally Lipschitz (since F is locally Lipschitz), hence according to Proposition 1.19, $\lim_{z\to x^*} \partial S^k(z, w_{1:k}) \subset \partial S^k(x^*, w_{1:k})$ and since $\partial S^k(z, w_{1:k}) = \partial_z S^k(z, w_{1:k}) \times \partial S^k_z(w_{1:k})$, we obtain $\lim_{z\to x^*} \partial S^k_z(w^*_{1:k}) \subset \partial S^k_{x^*}(w^*_{1:k})$. Since rank is lower semicontinuous, we deduce that there exists an open neighborhood U of x^* such that for any $z \in U$, $\partial S^k_z(w^*_{1:k})$ is of rank n. Moreover, $z \mapsto p^k_z(w^*_{1:k})$ is lower semicontinuous, so, up to taking U smaller, we can suppose that for any $z \in U$, $p^k_z(w^*_{1:k}) > 0$, i.e., $w^*_{1:k} \in \mathcal{O}^k_z$.
Let $y \in \mathsf{X}$. Since x^* is a globally attracting state, then by Proposition 1.6, there exist $t_0 > 0$ and $u_{1:t_0}$ a t_0-steps path from y to $x \in U$, i.e., $u_{1:t_0} \in \mathcal{O}^{t_0}_y$ and $x = S^{t_0}_y(u_{1:t_0}) \in U$. Since $\mathcal{O}^{t_0}_y$ is open and $S^{t_0}_y$ is continuously locally Lipschitz, by Corollary 1.4 we can assume w.l.o.g. that $S^{t_0}_y$ is differentiable at $u_{1:t_0}$.
Since $x \in U$, then $\partial S^k_x(w^*_{1:t_0})$ is of maximal rank, using the chain rule, see Proposition 1.21, we deduce that, for $T = t_0 + k$ and $u_{t_0+1:t_0+k} = w^*_{1:k}$, we have that $\partial S^T_y(u_{1:T})$ is of maximal rank. □

The next proposition states that if we find a point x^*, $k > 0$ and $w^*_{1:k} \in \mathcal{O}^k_{x^*}$ which satisfy the forementionned controllability condition (C_{x^*}), that is, $\partial S^k_{x^*}(w^*_{1:k})$ is of maximal rank, then, using Proposition 1.12, we can find $u^*_{1:k} \in \mathcal{O}^k_{x^*}$ as closed as we want from $w^*_{1:k}$ such that $S^k_{x^*}$ is differentiable in $u^*_{1:k}$ and $\mathcal{D}S^k_{x^*}(u^*_{1:k})$ is of maximal rank. In other words, our controllability condition (C_{x^*}) implies a full rank condition as used in [29].

Proposition 1.11. Suppose **H**1 and **H**2. Let $x^* \in \mathsf{X}$ and suppose that (C_{x^*}) holds. Then, condition (R_{x^*}) stated below holds.

Proof. By (C_{x^*}) and by Proposition 1.20, there exist $k > 0$ and $w^*_{1:k} \in \mathcal{O}^k_{x^*}$ such that $\partial S^k_{x^*}(w^*_{1:k})$ is of maximal rank. By Proposition 1.12 below, for any neighborhood $W \subset \mathsf{W}$ of $w^*_{1:k}$, there exists $u^*_{1:k} \in W$, such that $S^k_{x^*}$ is differentiable in $u^*_{1:k}$, with $\operatorname{rank} \mathcal{D}S^k_{x^*}(u^*_{1:k}) = n$. However $\mathcal{O}^k_{x^*}$ is open, so we can take $W = \mathcal{O}^k_{x^*}$ and complete the proof. □

Proposition 1.12. Suppose that $f : \mathsf{X} \to \mathsf{Y}$ is locally Lipschitz at $x_0 \in \mathsf{X}$, and that $\partial f(x_0)$ is of maximal rank, i.e., any $h \in \partial f(x_0)$ is of maximal rank. Then, there exists a neighborhood U

of x_0 such that for any $y \in U$, $\partial f(y)$ is of maximal rank. Moreover, for every neighborhood $V \subset U$ of x_0, there exists $y_0 \in V$ such that f is differentiable at y_0 and $\mathcal{D}f(y_0)$ is of maximal rank.

Proof. Let $A = \{h \in \mathcal{L}(T_{x_0}\mathsf{X}, T_{f(x_0)}\mathsf{Y}) \mid h \text{ is not of maximal rank}\}$. Since the application rank is l.s.c., then A is a closed set. By Proposition 1.18, $\partial f(x_0)$ is compact, and disjoint from A since it is assumed to be of maximal rank. Thus $\operatorname{dist}(\partial f(x_0), A) > 0$, where dist is a metric induced by a norm on the finitely dimensioned affine space $\mathcal{L}(T_{x_0}\mathsf{X}, T_{f(x_0)}\mathsf{Y})$. Moreover, there exists $h^* \in \partial f(x_0)$ such that for every $h \in \partial f(x_0)$ we have

$$\operatorname{dist}(h, A) \geqslant \operatorname{dist}(h^*, A) = \operatorname{dist}(\partial f(x_0), A) > 0.$$

By [30, Proposition 2.6.2(c)], there exists a neighborhood U of x_0 such that for all $y \in U$, $\operatorname{dist}(\partial f(y), A) \geqslant \operatorname{dist}(h^*, A)/2 > 0$, thus $\partial f(y)$ is of maximal rank. The second part follows from Rademacher's theorem, see Corollary 1.4. □

From now on, we can assume a full rank condition, i.e.,

$$\text{there exists } w_{1:k} \in \mathcal{O}_x^k \text{ such that } \mathcal{D}S_x^k(w_{1:k}) \text{ exists and is of maximal rank,} \tag{R_x}$$

instead of the controllability condition (C_x). We can then use Proposition 1.11 to extend our results. The next proposition states that if we can find a globally attracting state x^* satisfying the maximal rank condition (R_{x^*}), then we can find an attainable state. It generalizes [29, Proposition 3.6].

Proposition 1.13. Suppose **H**1 and **H**2. Let $x^* \in \mathsf{X}$ and suppose that there exist $k > 0$ and $w_{1:k}^* \in \mathcal{O}_{x^*}^k$ such that (R_{x^*}) is satisfied with $w_{1:k}^*$.

(i) There exists U a neighborhood of x^* such that for any $x \in U$, there exists $w_{1:k} \in \mathcal{O}_x^k$ for which $S_x^k(w_{1:k}) = S_{x^*}^k(w_{1:k}^*)$.

(ii) If x^* is globally attracting, then $S_{x^*}^k(w_{1:k}^*)$ is attainable, see (1.37).

Proof. (i) Let (U, φ) be a local chart of X around x^*, (V, θ) a local chart of X around $S_{x^*}^k(w_{1:k}^*)$, and (W, ψ) a local chart of W^k around $w_{1:k}^*$, such that the following differentiable function is well-defined

$$\tilde{S}^k\colon \begin{array}{l} \varphi(U) \times \psi(W) \subset \mathbb{R}^n \times \mathbb{R}^{kp} \to V \subset \mathbb{R}^n \\ (x, w) = ((x_1, \dots, x_n), (w_1, \dots, w_{kp})) \mapsto \tilde{S}_x^k(w) := \theta \circ S_{\varphi^{-1}(x)}^k \circ \psi^{-1}(w). \end{array}$$

We recall that the positive integers n and p are the dimensions of X and W, respectively.
By composition, observe that $\mathcal{D}_w\tilde{S}^k(\varphi(x^*), \psi(w_{1:k}^*))$ is surjective. Hence, we can find coordinates $i_1, \dots, i_n$ of $\mathbb{R}^{kp}$ such that

$$\det\left[\mathcal{D}_{w_{i_1}}\tilde{S}^k \mid \dots \mid \mathcal{D}_{w_{i_n}}\tilde{S}^k\right](\varphi(x^*), \psi(w_{1:k}^*)) = n.$$

Note that, up to a permutation of indices in the chart ψ, we can assume w.l.o.g. that $i_1, \dots, i_n$ equal respectively $kp - n + 1, \dots, kp$. To ease the presentation, we use the following abuse of notation $(w_1^*, \dots, w_{kp}^*) = \psi(w_{1:k}^*)$. Then, by the implicit function theorem, see Theorem 1.10, there exist neighborhoods M of $(\varphi(x^*), w_1^*, \dots, w_{kp-n}^*)$ and N of $(w_{kp-n+1}^*, \dots, w_{kp}^*)$, and a $\mathcal{C}^1$ function $g\colon M \to N$ such that, for every $(x_1, \dots, x_n, w_1, \dots, w_{kp-n}) \in M$, we have

$$\tilde{S}_{(x_1,\dots,x_n)}^k(w_1, \dots, w_{kp-n}, g(x_1, \dots, x_n, w_1, \dots, w_{kp-n})) = \tilde{S}_{\varphi(x^*)}^k(w_1^*, \dots, w_{kp}^*).$$

This proves (i).
(ii) Suppose that x^* is globally attracting. Let $U \subset \mathsf{X}$ be a neighborhood of x^* satisfying (i), and let $y \in \mathsf{X}$. Then, by Proposition 1.6(ii), there exist $k_1 > 0$ and $w_{1:k_1} \in \mathcal{O}_y^{k_1}$ such that $S_y^{k_1}(w_{1:k_1}) \in U$. Since U satisfies (i), there exists $w_{k_1+1:k_1+k} \in \mathcal{O}^k_{S_y^{k_1}(w_{1:k_1})}$ with $S_y^{k_1+k}(w_{1:k_1+k}) = S_{x^*}^k(w^*_{1:k})$. □

We discuss in the next proposition the forward accessibility of the control model (1.3). We recall that it is said to be forward accessible if for every $x \in \mathsf{X}$, the subset $A_+(x) \subset \mathsf{X}$ defined in (1.34) of states that can be reached in finite time starting from x, has a nonempty interior. The next proposition generalizes [29, Proposition 3.7].

Proposition 1.14. Suppose **H**1 and **H**2. If for every $x \in \mathsf{X}$, (R_x) holds, then the control model associated to (1.3) is forward accessible.
Furthermore, if F is smooth (infinitely differentiable), the control model is forward accessible if and only if for every $x \in \mathsf{X}$, (R_x) holds.

Proof. We apply the Local Submersion Theorem [59, Chapter 1.4]. Since S_x^k is a submersion at $w_{1:k}$, there exist local charts (W, ψ) of W^k around $w_{1:k}$ and (V, φ) of X around $S_x^k(w_{1:k})$ such that

$$\varphi \circ S_x^k \circ \psi(u_1, \dots, u_{kp}) = (u_1, \dots, u_n) \quad \text{for all } (u_1, \dots, u_{kp}) \in \psi(W).$$

Therefore, since φ is a continuous bijection (by definition of a local chart), then there exists a neighborhood U of $S_x^k(w_{1:k})$ such that $S_x^k(W) = U$. Moreover, $\mathcal{O}_{x^*}^k$ is an open subset of W^k, so we can assume $W \subset \mathcal{O}_x^k$. Therefore, $U \subset A_+^k(x)$, which hence has a nonempty interior.
Suppose now that F is smooth and that the control model is forward accessible. Then, for every $x \in \mathsf{X}$, $\mathrm{int}(A_+(x)) \neq \emptyset$. Since $A_+(x) = \cup_{k \geqslant 0} A_+^k(x)$, we deduce that there exists $k \in \mathbb{N}$ such that $\mathrm{int}(A_+^k(x)) \neq \emptyset$. Since $\mathrm{int}(A_+^0(x)) = \mathrm{int}(\{x\}) = \emptyset$, we find that necessarily $k > 0$. By Sard's theorem [59, Appendix 1], we have that the set $N := \{\mathbf{w} \in \mathcal{O}_x^k \mid \mathrm{rank}\, \mathcal{D}S_x^k(\mathbf{w}) < n\}$ is of measure zero, that is, for all charts (φ, U) of X, we have $\mathrm{Leb}\, \varphi(N \cap U) = 0$, hence $\mathrm{int}(N) = \emptyset$. We deduce that there exists $w_{1:k} \in \mathcal{O}_x^k \setminus N$, i.e., such that $\mathrm{rank}\, \mathcal{D}S_x^k(w_{1:k}) = n$. □

4.2 *Proofs of the main results: verifiable conditions for irreducibility and aperiodicity*

4.2.1 *T-chain and irreducibility*

Here, we generalize the main results of [29] on φ-irreducibility of Markov kernels P defined via (1.3), as well as the T-chain property.

First, we slightly generalize [109, Lemma 3.0] to our context, that is, for a locally Lipschitz function between manifolds instead of a smooth function between open subsets of Euclidean spaces.

Lemma 1.1. Let X_1 be a n-dimensional manifold, $\tilde{\mathsf{W}}_1$ a m-dimensional manifold, $\hat{\mathsf{W}}_1$ a n-dimensional manifold, equipped with their respective Borelian σ-fields and with a measure ζ_{X} (resp. $\zeta_{\tilde{\mathsf{W}}}$, $\zeta_{\hat{\mathsf{W}}}$), which satisfies that for any $A \in \mathcal{B}(\mathsf{X}_1)$ (resp. of $\mathcal{B}(\tilde{\mathsf{W}}_1)$, $\mathcal{B}(\hat{\mathsf{W}}_1)$), $\zeta_{\mathsf{X}}(A) = 0$ (resp. $\zeta_{\tilde{\mathsf{W}}}(A) = 0$, $\zeta_{\hat{\mathsf{W}}}(A) = 0$) if and only if $\varphi(A \cap U)$ is Lebesgue-negligible for every chart (φ, U).
Let $G\colon (x, \tilde{w}, \hat{w}) \in \mathsf{X}_1 \times \tilde{\mathsf{W}}_1 \times \hat{\mathsf{W}}_1 \mapsto z \in \mathsf{X}_1$ be a locally Lipschitz map differentiable in $(x_0, \tilde{w}_0, \hat{w}_0)$ such that $\mathrm{rank}\, \mathcal{D}_{\hat{w}} G(x_0, \tilde{w}_0, \hat{w}_0) = n$. Then,

(i) There exists an open subset $\mathsf{X} \times \tilde{\mathsf{W}} \times \hat{\mathsf{W}} \subset \mathsf{X}_1 \times \tilde{\mathsf{W}}_1 \times \hat{\mathsf{W}}_1$ containing $(x_0, \tilde{w}_0, \hat{w}_0)$ such that for any $x \in \mathsf{X}$, the measure defined by

$$\nu(x, \cdot)\colon A \subset \mathsf{X}_1 \mapsto \int_{\tilde{\mathsf{W}}} \int_{\hat{\mathsf{W}}} \mathbb{1}_A\{G(x, \tilde{w}, \hat{w})\} d\zeta_{\tilde{\mathsf{W}}}(\tilde{w}) d\zeta_{\hat{\mathsf{W}}}(\hat{w}) \tag{1.39}$$

is equivalent to the measure ζ_{X} on an open subset $\mathcal{R}_x$ of X_1.

(ii) There exist $c > 0$, U_{x_0} an open subset of X_1 containing x_0, $V_{x_0}^{\tilde{w}_0, \hat{w}_0}$ an open subset of $\tilde{\mathsf{W}}_1$ containing $G(x_0, \tilde{w}_0, \hat{w}_0)$ such that for every $x \in \mathsf{X}$ and every measurable subset A of X_1, we have $\nu(x, A) \geqslant c \mathbb{1}_{U_{x_0}}(x) \times \zeta_{\mathsf{X}_1}(A \cap V_{x_0}^{\tilde{w}_0, \hat{w}_0})$.

Proof. First we prove the lemma when $\mathsf{X}_1, \tilde{\mathsf{W}}_1, \hat{\mathsf{W}}_1$ are open subsets respectively of $\mathbb{R}^n$, $\mathbb{R}^m$, $\mathbb{R}^n$, and $\zeta_{\mathsf{X}}, \zeta_{\tilde{\mathsf{W}}}, \zeta_{\hat{\mathsf{W}}}$ are assumed to be the Lebesgue measures on $\mathbb{R}^n$, $\mathbb{R}^m$, $\mathbb{R}^n$ respectively. Define the function

$$G^\star\colon (x, \tilde{w}, \hat{w}) \in \mathsf{X}_1 \times \tilde{\mathsf{W}}_1 \times \hat{\mathsf{W}}_1 \mapsto (x, \tilde{w}, G(x, \tilde{w}, \hat{w})) \in \mathbb{R}^{n+m+n}$$

Then, since $\mathcal{D}_{\hat{w}} G(x_0, \tilde{w}_0, \hat{w}_0)$ is of rank n, then $\mathcal{D}G^\star(x_0, \tilde{w}_0, \hat{w}_0)$ exists and is a full-rank squared matrix. Therefore, the inverse function theorem –as stated in Theorem 1.9– applies and we find a neighborhood $\mathsf{X} \times \tilde{\mathsf{W}} \times \hat{\mathsf{W}}$ of $(x_0, \tilde{w}_0, \hat{w}_0)$, a neighborhood $\mathcal{R}$ of $(x_0, \tilde{w}_0, z_0)$ (where $z_0 \coloneqq G(x_0, \tilde{w}_0, \hat{w}_0)$), and a locally Lipschitz function $H^\star\colon \mathcal{R} \to \mathsf{X} \times \tilde{\mathsf{W}} \times \hat{\mathsf{W}}$ such that

$$H^\star(G^\star(x, \tilde{w}, \hat{w})) = (x, \tilde{w}, \hat{w}) \quad \text{for every } (x, \tilde{w}, \hat{w}) \in \mathsf{X} \times \tilde{\mathsf{W}} \times \hat{\mathsf{W}}.$$

Thus, there exists a locally Lipschitz function $H\colon \mathcal{R} \to \hat{\mathsf{W}}$ such that

$$H(x, \tilde{w}, G(x, \tilde{w}, \hat{w})) = \hat{w} \quad \text{for every } (x, \tilde{w}, \hat{w}) \in \mathsf{X} \times \tilde{\mathsf{W}} \times \hat{\mathsf{W}}.$$

Then, by the chain rule, see [30, Theorem 2.6.6], for every $(x, \tilde{w}, \hat{w}) \in \mathsf{X} \times \tilde{\mathsf{W}} \times \hat{\mathsf{W}}$ at which G admits a partial derivative w.r.t. $\hat{w}$, we have that

$$\mathcal{D}_z H(x, \tilde{w}, G(x, \tilde{w}, \hat{w})) = [\mathcal{D}_{\hat{w}} G(x, \tilde{w}, \hat{w})]^{-1} \tag{1.40}$$

which is thus invertible. Moreover, by [30, Proposition 2.6.2(c)], $\mathcal{D}_z H$ is continuous at points on which it is defined (which is dense by Rademacher's theorem [43, Theorem 3.2]). Therefore, there exists $h_0 > 0$ such that in each of these points, by (1.40) we have

$$|\det \mathcal{D}_z H| \geqslant h_0. \tag{1.41}$$

Then, applying [62, Theorem 3] and Fubini's theorem, we get

$$\nu(x, A) = \int_A \left(\int \mathbb{1}_{\mathcal{R}}(x, \tilde{w}, z) \, |\det \mathcal{D}_z H| \, \mathrm{d}\tilde{w} \right) \mathrm{d}z, \tag{1.42}$$

so that

$$p(x, z) \coloneqq \int \mathbb{1}_{\mathcal{R}}(x, \tilde{w}, z) \, |\det \mathcal{D}_z H| \, \mathrm{d}\tilde{w} \tag{1.43}$$

defines a density w.r.t. Lebesgue for $\nu(x, \cdot)$. The rest of proof goes as in [109, Lemma 3.0], that we recall here for completeness.

Fix $x \in \mathsf{X}$ and let $\mathcal{R}_x$ be the open subset of $\mathbb{R}^n$ defined by

$$\mathcal{R}_x = \left\{ z \in \mathsf{X}_1 \mid \exists \tilde{w} \in \tilde{\mathsf{W}}, (x, \tilde{w}, z) \in \mathcal{R} \right\}.$$

Then, note that $p(x, z)$ is positive if and only if $z \in \mathcal{R}_x$, and zero otherwise. This proves (i). For (ii), observe that, since $\mathcal{R}$ is a neighborhood of $(x_0, \tilde{w}_0, z_0)$, then it contains a nonempty open subset $\mathsf{X}_0 \times \tilde{\mathsf{W}}_0 \times \mathsf{Z}_0$ containing $(x_0, \tilde{w}_0, z_0)$. We get then that $p(x, z) \geqslant h_0 \times \mathrm{Leb}(\tilde{\mathsf{W}}_0)$ for every $(x, z) \in \mathsf{X}_0 \times \mathsf{Z}_0$. Then,

$$\nu(x, A) \geqslant h_0 \mathrm{Leb}(\tilde{\mathsf{W}}_0) \mathbb{1}\{x \in \mathsf{X}_0\} \times \mathrm{Leb}(A \cap \mathsf{Z}_0)$$

which proves (ii).
Now suppose that $\mathsf{X}_1, \tilde{\mathsf{W}}_1, \hat{\mathsf{W}}_1$ are manifolds.
Let (φ, X_2) be a local chart of X_1 around x_0, $(\tilde{\psi}, \tilde{\mathsf{W}}_2)$ be a local chart of $\tilde{\mathsf{W}}_1$ around $\tilde{w}_0$, $(\hat{\psi}, \hat{\mathsf{W}}_2)$ be a local chart of $\hat{\mathsf{W}}_1$ around $\hat{w}_0$, and (η, X_3) be a local chart of X_1 around $z_0 = G(x_0, \tilde{w}_0, \hat{w}_0)$. Then, define the locally Lipschitz map

$$G^{\mathrm{loc}} \colon (x, \tilde{w}, \hat{w}) \in \varphi(\mathsf{X}_2) \times \tilde{\psi}(\tilde{\mathsf{W}}_2) \times \hat{\psi}(\hat{\mathsf{W}}_2) \mapsto z = \eta \circ G(\varphi^{-1}(x), \tilde{\psi}^{-1}(\tilde{w}), \hat{\psi}(\hat{w})) \in \mathbb{R}^n.$$

Thus, (i) and (ii) hold with G^{loc}, and

$$\nu^{\mathrm{loc}}(x, A) = \int_{\tilde{\mathsf{W}}_0} \int_{\hat{\mathsf{W}}_0} \mathbb{1}_A(x, \tilde{w}, \hat{w}) \eta \circ G(\varphi^{-1}(x), \tilde{\psi}^{-1}(\tilde{w}), \hat{\psi}^{-1}(\hat{w}))\, \mathrm{d}\tilde{w}\, \mathrm{d}\hat{w}$$

is equivalent to the Lebesgue measure, for all $x \in \mathsf{X}_0$, and $\mathsf{X}_0 \times \tilde{\mathsf{W}}_0 \times \hat{\mathsf{W}}_0$ being a neighborhood of $(\varphi(x_0), \tilde{\psi}(\tilde{w}_0), \hat{\psi}(\hat{w}_0))$. But, by assumption on the measures ζ_{X}, $\zeta_{\tilde{\mathsf{W}}}$ and $\zeta_{\hat{\mathsf{W}}}$, $\nu(x, \cdot)$ is locally equivalent to $\rho^{-1} \circ \nu^{\mathrm{loc}}(\varphi(x), \cdot)$ for all local chart (ρ, A) of X_1, thus is locally equivalent to $\eta^{-1} \circ \mathrm{Leb}_n$ where Leb_n is the Lebesgue measure of $\mathbb{R}^n$. Thus, $\nu(x, \cdot)$ is equivalent to ζ_{X}. This proves (i).
Now apply (ii) to G^{loc}, and find $c > 0$, U_{x_0} an open of $\varphi(\mathsf{X}_2)$ containing $\varphi(x_0)$, $V_{x_0}^{\tilde{w}_0, \hat{w}_0}$ an open of X_3 containing $\eta(z_0)$, such that

$$\nu^{\mathrm{loc}}(x, A) \geqslant c \mathbb{1}_{U_{x_0}}(x) \times \mathrm{Leb}_n(A \cap V_{x_0}^{\tilde{w}_0, \hat{w}_0}) \quad \text{for every } x \in \varphi(\mathsf{X}), A \subset \mathbb{R}^n.$$

But, by assumption on ζ_{X}, we find $L_1^\eta, L_2^\eta > 0$ such that

$$\begin{aligned}
\nu(x, A) &\geqslant L_1^\eta \times \nu^{\mathrm{loc}}(\varphi(x), \eta(A \cap \mathsf{X}_3)) \\
&\geqslant L_1^\eta \times c \mathbb{1}_{U_{x_0}}(\varphi(x)) \times \mathrm{Leb}_n(\eta(A \cap \mathsf{X}_3) \cap V_{x_0}^{\tilde{w}_0, \hat{w}_0}) \\
&\geqslant L_1^\eta \times c \mathbb{1}_{\varphi^{-1}(U_{x_0})}(x) \times L_2^\eta \times \zeta_{\mathsf{X}}(A \cap \eta^{-1}(V_{x_0}^{\tilde{w}_0, \hat{w}_0}))
\end{aligned}$$

for all $x \in \mathsf{X}$ and $A \subset \mathsf{X}_1$, which proves (ii). □

We can now state the following result.

Proposition 1.15. Consider the Markov kernel P defined via (1.3), and suppose **H**1-**H**2. Let $x \in \mathsf{X}$.

(i) If (R_x) holds for some $k > 0$ and $w_{1:k} \in \mathcal{O}_x^k$, then there exist $c > 0$, and open subsets U_x

and $V_x^{w_{1:k}}$ of X containing x and $S_x^k(w_{1:k})$ respectively, such that

$$P^k(y, A) \geqslant c\zeta_{\mathsf{X}}(A) \quad \text{for every } y \in U_x \text{ and } A \in \mathcal{B}(\mathsf{X}), \tag{1.44}$$

for some nontrivial measure ζ_{X} on $V_x^{w_{1:k}}$. That is, U_x is a k-small set.

(ii) If furthermore F is smooth (infinitely differentiable), and if there exist $k > 0$, $c > 0$ and (φ, V) a local chart of X such that

$$P^k(x, A) \geqslant c\mathrm{Leb} \circ \varphi(A \cap V) \quad \text{for every } A \in \mathcal{B}(\mathsf{X}), \tag{1.45}$$

then (R_x) holds.

Proof. Condition (R_x) implies that $\operatorname{rank} \mathcal{D}S_x^k(w_{1:k}) = n$ for some $k > 0$ and $w_{1:k} \in \mathcal{O}_x^k$. Since $(\bar{x}, \bar{w}_{1:k}) \mapsto p_{\bar{x}}^k(\bar{w}_{1:k})$ is l.s.c., and $p_x^k(w_{1:k}) > 0$, then there exist $p_0 > 0$ and a neighborhood $\mathsf{X}_1 \times \mathsf{W}_1$ of $(x, w_{1:k})$ such that $p_{\bar{x}}^k(\bar{w}_{1:k}) \geqslant p_0$ for every $\bar{x} \in \mathsf{X}_1$ and every $\bar{w}_{1:k} \in \mathsf{W}_1$. Then, for every $y \in \mathsf{X}_1$, we have

$$P^k(y, A) = \int_{\mathcal{O}_y^k} \mathbb{1}_A(S_y^k(\bar{w}_{1:k}))p_y^k(\bar{w}_{1:k})\mathrm{d}\zeta_{\mathsf{W}}^{\otimes k}(\bar{w}_{1:k}) \geqslant p_0 \int_{\mathsf{W}_1} \mathbb{1}_A(S_y^k(\bar{w}_{1:k}))\mathrm{d}\zeta_{\mathsf{W}}^{\otimes k}(\bar{w}_{1:k}). \tag{1.46}$$

Since S_x^k is a submersion at $w_{1:k}$, by the Local Submersion theorem, there exists a local chart (V, ψ) of W^k around $w_{1:k}$ and a local chart (U, φ) of X around $S_x^k(w_{1:k})$, such that for every $(\bar{w}_1, \dots, \bar{w}_{kp}) \in \psi(V)$, we have

$$\varphi \circ S_x^k \circ \psi^{-1}(\bar{w}_1, \dots, \bar{w}_{kp}) = (\bar{w}_1, \dots, \bar{w}_n). \tag{1.47}$$

Note that, up to taking V and W_1 smaller, we can assume $V = \mathsf{W}_1$, and that $\psi(\mathsf{W}_1) = \hat{\mathsf{W}}_1 \times \tilde{\mathsf{W}}_1$ is a rectangle of $\mathbb{R}^{kp}$, with $\hat{\mathsf{W}}_1 = \{(\bar{w}_1, \dots, \bar{w}_n) \in \mathbb{R}^n \mid \exists (\bar{w}_{n+1}, \dots, \bar{w}_{kp}) \in \mathbb{R}^{kp-n}, (\bar{w}_1, \dots, \bar{w}_{kp}) \in \psi(\mathsf{W}_1)\} \subset \mathbb{R}^n$ and likewise $\tilde{\mathsf{W}}_1 \subset \mathbb{R}^{kp-n}$. Hence, the function

$$\begin{aligned} G\colon \quad & \mathsf{X} \times \hat{\mathsf{W}}_1 \times \tilde{\mathsf{W}}_1 \to \mathsf{X} \\ & (\bar{x}, (\bar{w}_1, \dots, \bar{w}_n), (\bar{w}_{n+1}, \dots, \bar{w}_{kp})) \mapsto S_{\bar{x}}^k \circ \psi^{-1}(\bar{w}_1, \dots, \bar{w}_{kp}) \end{aligned} \tag{1.48}$$

satisfies, by (1.47), that $\operatorname{rank} \mathcal{D}_{(w_1,\dots,w_n)} G(x, w_1, \dots, w_{kp}) = n$. Then, by Lemma 1.1 and **H**1(ii), there exist $c > 0$, U_x an open subset of X containing x, $(\varphi, V_x^{w_{1:k}})$ a local chart of X around $S_x^k(w_{1:k})$ such that, for every $y \in \mathsf{X}$, we have

$$\int_{\mathsf{W}_1} \mathbb{1}_A(S_y^k(\bar{w}_{1:k}))\mathrm{d}\zeta_{\mathsf{W}}^k(\bar{w}_{1:k}) \geqslant c\mathbb{1}_{U_x}(y)\zeta_{\mathsf{X}}(A \cap V_x^{w_{1:k}}), \tag{1.49}$$

with $\zeta_{\mathsf{X}} = \mathrm{Leb} \circ \varphi(\cdot \cap V_x^{w_{1:k}})$ is a measure which satisfies the assumption required in Lemma 1.1 on $V_x^{w_{1:k}}$. Combining (1.46) and (1.49) gives $P^k(y, A) \geqslant cp_0\zeta_{\mathsf{X}}(A \cap V_x^{w_{1:k}})$ for every $y \in U_x \cap \mathsf{X}_1$, proving (i).

Suppose now that F is smooth, then $(\bar{x}, \bar{w}_{1:k}) \mapsto S_{\bar{x}}^k(\bar{w}_{1:k})$ is smooth for all $k > 0$. Take $k > 0$, $c > 0$ and (φ, V) a local chart such that (1.45) holds. Let $N = \{S_x^k(\bar{w}_{1:k}) \in \mathsf{X} \mid \bar{w}_{1:k} \in \mathcal{O}_x^k, \operatorname{rank} \mathcal{D}S_x^k(\bar{w}_{1:k}) < n\}$. By Sard's theorem, we know that $\mathrm{Leb} \circ \varphi(N \cap V) = 0$, implying that $P^k(x, V \setminus N) \geqslant c\mathrm{Leb} \circ \varphi(V \setminus N) = c\mathrm{Leb} \circ \varphi(V) > 0$. Hence there exists $w_{1:k} \in \mathcal{O}_x^k$ such that $S_x^k(w_{1:k}) \in V \setminus N$, i.e., $\operatorname{rank} \mathcal{D}S_x^k(w_{1:k}) = n$. □

Following [29, Corollary 4.1], we now deduce sufficient conditions for the Markov kernel P to define a T-chain.

Corollary 1.3. Consider the Markov kernel P defined via (1.3), and suppose **H**1 and **H**2. Suppose that for any $x \in \mathsf{X}$, (C_x) holds. Then X can be written as the union of open small sets and thus P is a T-chain.

Proof. First, using Proposition 1.11, for all $x \in \mathsf{X}$, (R_x) holds, i.e., there exist $k > 0$ and $u_{1:k} \in \mathcal{O}_x^k$ such that S_x^k is differentiable in $u_{1:k}$ and $\operatorname{rank} \mathcal{D}S_x^k(u_{1:k}) = n$.
Proposition 1.15 implies that for every $x \in \mathsf{X}$, there exists an open neighborhood U_x of x in X which is a k-small set. Denoting a the Dirac distribution in k, we find that U_x is ν_a-petite, hence, by [110, Proposition 6.2.3], K_a possesses a continuous component T which is nontrivial on U_x and in particular at x. Thus, by [110, Proposition 6.2.4], P is a T-chain. □

We now characterize the support of the maximal irreducibility measure of P. We recall that, by [110, Proposition 4.2.2], any φ-irreducible Markov kernel P admits a maximal irreducibility measure ψ, that is, P is ψ-irreducible and for every irreducibility measure φ of P, we have that $\operatorname{supp} \varphi \subset \operatorname{supp} \psi$. The proof mimics the one of [29, Proposition 4.2], and is given for completeness in Section C.

Proposition 1.16. Suppose that P is a ψ-irreducible Markov kernel, defined via (1.3), with ψ a maximal irreducibility measure, that **H**1 holds and that F is continuous. Then

$$\operatorname{supp} \psi = \{x^* \in \mathsf{X} \mid x^* \text{ is globally attracting}\}. \tag{1.50}$$

Furthermore, if $x^* \in \mathsf{X}$ is globally attracting, then

$$\operatorname{supp} \psi = \overline{A_+(x^*)}. \tag{1.51}$$

We now state our core results, from which we deduce Theorem 1.2. Assuming the controllability condition is satisfied at every x, there is equivalence between the irreducibility of P and the existence of a globally attracting state.

Theorem 1.5. Consider the Markov kernel P defined via (1.3), and suppose **H**1 and **H**2. Suppose (C_x) is satisfied for every $x \in \mathsf{X}$. Then P is φ-irreducible if and only if a globally attracting state exists.

Proof. By Proposition 1.11, we know that (R_x) holds for any $x \in \mathsf{X}$. If P is φ-irreducible, then by Proposition 1.16, any point of the support of the nontrivial measure φ is globally attracting, hence there exists a globally attracting state. Conversely, if $x^* \in \mathsf{X}$ is globally attracting, then, by Corollary 1.1, x^* is reachable by P, and by Corollary 1.3, P is a T-chain. As a result, by [110, Proposition 6.2.1], P is φ-irreducible. □

We deduce from this theorem our first practical result in order to prove the irreducibility, the T-chain property of a Markov kernel following the model investigated. If assumptions **H**1 and **H**2 are satisfied for a Markov kernel defined via (1.3), the theorem below implies that one needs to find a globally attracting state x^* where the controllability condition (C_{x^*}) is satisfied to obtain the φ-irreducible and T-chain property of the Markov kernel.

Theorem 1.6 (Practical condition for φ-irreducibility and T-chain property)**.** Consider the Markov kernel P defined via (1.3), and suppose **H**1-**H**3. Then P is a φ-irreducible T-chain, and thus every compact set of X is petite.

Proof. By Proposition 1.10, for any $x \in \mathsf{X}$, (C_x) holds. Then, by Corollary 1.3, P is a T-chain and by Theorem 1.5, $P\acute{\mathrm{I}}$ is φ-irreducible, and by [110, Theorem 6.2.5] all compact sets of X are petite. □

This latter theorem constitutes the first part of our main result stated in Theorem 1.2 while the second part relates to the aperiodicity of the kernel which is developed in the next section.

4.2.2 *Aperiodicity*

In this section, we provide conditions for P to be aperiodic. We start with the following characterization which is the counterpart to Theorem 1.5 for a kernel to be φ-irreducible aperiodic.

Theorem 1.7. Consider the Markov kernel P defined via (1.3), and suppose **H**1 and **H**2. If for every $x \in \mathsf{X}$, (C_x) holds, then P is a φ-irreducible aperiodic Markov kernel if and only if there exists a steadily attracting state.

Proof. First suppose that P is φ-irreducible and aperiodic. By Theorem 1.5, there exists a globally attracting state $x^* \in \mathsf{X}$. Besides, by Proposition 1.13, there exists an attainable state y^*, to which we apply Proposition 1.9 (iii), so that there exists a d-cycle. However, P is aperiodic, so $d = 1$. Thus, by Proposition 1.9 (ii), y^* is steadily attracting.
Conversely, suppose that there exists a steadily attracting state x^*. By Proposition 1.8 (i), x^* is globally attracting, so that by Theorem 1.5, P is φ-irreducible. It remains to prove that it is aperiodic. By Proposition 1.11, (R_{x^*}) holds for some $k > 0$ and $w^*_{1:k} \in \mathcal{O}^k_{x^*}$. Therefore, we apply Proposition 1.13, hence $y^* := S^k_{x^*}(w^*_{1:k})$ is attainable. Let U be a neighborhood of x^* which satisfies Proposition 1.13 (i). Since x^* is steadily attracting, there exists $T > 0$, such that for every $t \geqslant T$, there exists $u_{1:t} \in \mathcal{O}^t_{y^*}$ such that $z := S^t_{y^*}(u_{1:t}) \in U$. As U satisfies Proposition 1.13 (i), then there exists $w_{1:k} \in \mathcal{O}^k_z$ such that $S^k_z(w_{1:k}) = S^k_{x^*}(w^*_{1:k}) = y^*$. All in all, we have that for every $t \geqslant T$, there exists $w_{1:k+t} \in \mathcal{O}^{k+t}_{y^*}$ such that $y^* = S^{k+t}_{y^*}(w_{1:k+t})$, hence $y^* \in A^{k+t}_+(y^*)$. By Proposition 1.9 (iii), there exists a 1-cycle, i.e., P is aperiodic. □

We now state our main practical condition to ensure that P is aperiodic.

Theorem 1.8 (Practical condition for φ-irreducibility and aperiodicity)**.** Consider the Markov kernel P defined via (1.3), and suppose **H**1-**H**2 and **H**4. Then P is a φ-irreducible aperiodic T-chain, and every compact set of X is small.

Proof. By Proposition 1.10, (C_x) holds for any $x \in \mathsf{X}$. Thus, by Theorems 1.6 and 1.7, we have that P is a φ-irreducible aperiodic T-chain for which compact sets of X are petite. Note that, by [110, Theorem 5.5.7], any petite set is small. □

4.3 *Proofs for the application to CMA-ES*

Proof of Proposition 1.2 (i). By Corollary 1.2, it is sufficient to find, for any $\theta_0 = (z_0, \mathbf{\Sigma}_0) \in \mathsf{X}$, a sequence $\{v_k\}_{k\geqslant 1}$ such that $v_{1:k} \in \overline{\mathcal{O}^k_{\theta_0}}$ for every $k \geqslant 1$, and $\lim_{k\to\infty} S^k_{\theta_0}(v_{1:k}) = (0, I_d)$. Since f has Lebesgue negligible level sets, we have, by Proposition 1.1, that for every $(z, \mathbf{\Sigma}) \in \mathsf{X}$, and for every $u \in \mathbb{R}^d$, the element $v = (u, \dots, u)$ of $\mathsf{W} = \mathbb{R}^{d\mu}$ belongs to $\overline{\mathcal{O}^1_{(z,\mathbf{\Sigma})}}$.

Then, set $v^1_1 = \cdots = v^\mu_1 = -\mathbf{\Sigma}_0^{-1/2} z_0$ so that $v_1 = (v^1_1, \dots, v^\mu_1) \in \overline{\mathcal{O}^1_{\theta_0}}$ and $z_1 = 0$. Note that $S^1(v_1) = \theta_1 := (0, \mathbf{\Sigma}_1)$, for some $\mathbf{\Sigma}_1 \in \mathcal{S}^d_{++}$. Next, consider $(e_1, \dots, e_d)$ an orthogonal basis of eigenvectors of the positive definite matrix $\mathbf{\Sigma}_1$, with $\mathbf{\Sigma}_1 e_i = \lambda_i(\mathbf{\Sigma}_1) e_i$, where $\lambda_i(\cdot)$ denotes the function that maps a symmetric matrix to its i-th largest eigenvalue (counted with multiplicity).

Let then $\kappa \geqslant 0$ and define $v^1_2 = \cdots = v^\mu_2 = -\mathbf{\Sigma}_1^{-1/2} \kappa e_2$ and $\theta_2 = (z_2, \mathbf{\Sigma}_2) = S^1_{\theta_1}(v_2)$. Then let $v^1_3 = \cdots = v^\mu_3 = -\mathbf{\Sigma}_2^{-1/2} z_2$ and $\theta_3 = (z_3, \mathbf{\Sigma}_3) = S^1_{\theta_2}(v_3)$. Then, we have $z_2 = \kappa e_2$ and $z_3 = 0$. Besides, $\mathbf{\Sigma}_2 = r_2^{-1} \times \left((1-c)\mathbf{\Sigma}_1 + c\kappa^2 e_2 e_2^\top\right)$ with $r_2 = \det((1-c)\mathbf{\Sigma}_1 + c\kappa^2 e_2 e_2^\top)^{1/d}$ depends continuously on the choice of $\kappa \geqslant 0$. Moreover, we have $1 - c \leqslant r_2 \leqslant 1 - c + c\kappa^2$. Then,

$$r_3 \mathbf{\Sigma}_3 = \mathbf{\Sigma}_1 + c(1-c)^{-2}\kappa^2 \times (r_2 + 1 - c) e_2 e_2^\top =: K_3 \tag{1.52}$$

for some $r_3 > 0$. However, the eigenvalue of the matrix K_3 associated to the eigenvector e_1 equals $\lambda_1(\mathbf{\Sigma}_1)$ for any value of κ, while the eigenvalue of K_3 associated to the eigenvector e_2 depends continuously on $\kappa \geqslant 0$ and tends to $+\infty$ when $\kappa \to \infty$ and to $\lambda_2(\mathbf{\Sigma}_1) \leqslant \lambda_1(\mathbf{\Sigma}_1)$ when $\kappa \to 0$. Hence, there exists a value of $\kappa \geqslant 0$ such that the eigenvalues of K_3 associated respectively to the eigenvectors e_1 and e_2 are equal. Setting κ to this value, we get then that $\lambda_1(\mathbf{\Sigma}_3) = \lambda_2(\mathbf{\Sigma}_3)$.

Repeating eventually these steps $(d-1)$ times, we find $v_4, \dots, v_{1+2(d-1)} \in \mathsf{W}$ such that, denoting $\theta_k = (z_k, \mathbf{\Sigma}_k) = S^k_{\theta_0}(v_{1:k})$, $z_{1+2(d-1)} = 0$, and $\lambda_1(\mathbf{\Sigma}_{1+2(d-1)}) = \cdots = \lambda_d(\mathbf{\Sigma}_{1+2(d-1)})$, with $v_k \in \overline{\mathcal{O}^1_{\theta_{k-1}}}$ for each $k > 0$. However, $\det(\mathbf{\Sigma}_k) = 1$ for every $k \in \mathbb{N}$, thus $\mathbf{\Sigma}_{1+2(d-1)} = I_d$. For the next steps $k \geqslant 2d$, we choose $v_k = 0 \in \overline{\mathcal{O}^1_{\theta_{k-1}}}$, so that, by induction, we obtain $\theta_k = (0, I_d)$. By Corollary 1.2, we find that $(0, I_d)$ is a steadily attracting state. □

Proof of Proposition 1.2 (ii). Let $k > 0$ and $v_{1:k} \in \overline{\mathcal{O}^k_{\theta_0}}$, with $\theta_0 = (z_0, \mathbf{\Sigma}_0) := (0, I_d)$. We find here values for $k > 0$ and $v_1, \dots, v_k \in \mathsf{W}$ such that the map

$$\mathcal{D} S^k_{(0,I_d)}(v_{1:k}) : T_{(v_{1:k})}\mathsf{W}^k \to T_{S^k_{(0,I_d)}(v_{1:k})}\mathsf{X}$$

is full-rank, i.e., is surjective. We remind that $\mathsf{W} = (\mathbb{R}^d)^\mu$, hence $T_{(v_{1:k})}\mathsf{W}^k = \mathsf{W}^k = \mathbb{R}^{d\times\mu\times k}$. Moreover, we have $\mathsf{X} = \mathbb{R}^d \times \det^{-1}(\{1\})$, therefore, by [100, Proposition 5.38]

$$T_{S^k_{(0,I_d)}(v_{1:k})}\mathsf{X} = \mathbb{R}^d \times \ker\, \mathcal{D}\det(\mathbf{\Sigma}_k),$$

where $\theta_t = (z_t, \mathbf{\Sigma}_t) = S^t_{\theta_0}(v_{1:t})$ for each $t = 0, \dots, k$.

We define then inductively the covariance matrix before normalization as

$$K_{t+1} = (1-c)K_t + c\sqrt{K_t}\sum_{i=1}^{\mu} w_i \left(v^i_{t+1}\right)\left(v^i_{t+1}\right)^\top \sqrt{K_t}$$

with $K_0 = \mathbf{\Sigma}_0 = I_d$, so that, by induction, we have for every $t = 0, \dots, k$, $\mathbf{\Sigma}_t = \frac{K_t}{\det(K_t)^{1/d}}$. Let us introduce (small) perturbations $h_t = (h_t^1, \dots, h_t^\mu) \in \mathsf{W}$ for $t = 1, \dots, k$, and let us denote the perturbed process as

$$\theta_t^h = (z_t^h, \mathbf{\Sigma}_t^h) = S_{\theta_0}^t(v_{1:t} + h_{1:t}).$$

Define $K_t^h \in \mathcal{S}_{++}^d$ similarly. Set $k_0 = d(d+1)/2$ the dimension of $\mathcal{S}^d$, and set $k = k_0(k_0 - 1) + 1$. Then, set $v_1, \dots, v_{k_0}$ as follows. Define $\psi_1, \dots, \psi_{k_0}$ nonzero vectors of $\mathbb{R}^d$, such that $(\psi_1\psi_1^\top, \dots, \psi_{k_0}\psi_{k_0}^\top)$ forms a basis of $\mathcal{S}^d$.

For $t = 1, \dots, k_0$, using Proposition 1.1, we set $v_t = (K_{t-1}^{-1/2}\psi_t, \dots, K_{t-1}^{-1/2}\psi_t) \in \overline{\mathcal{O}^1_{\theta_{t-1}}}$, so that $K_t = (1-c)K_{t-1} + c\psi_t\psi_t^\top$. Fix then $\kappa_t^1 \in \mathbb{R}$, and let $\varepsilon_1 > 0$ be an arbitrary small positive quantity. Set $h_t^1 = \cdots = h_t^\mu = \frac{1}{2}\kappa_t^1\varepsilon_1 K_{t-1}^{-1/2}\psi_t$ and then

$$K_t^h = (1-c)K_{t-1}^h + c\psi_t\psi_t^\top + \varepsilon_1\kappa_t^1 c\psi_t\psi_t^\top + \varepsilon_1 A_t^1(\varepsilon_1),$$

where $A_t^1(\varepsilon_1) \in \mathcal{S}^d$ tends to 0 when $\varepsilon_1 \to 0$. Then, we get by induction,

$$K_{k_0}^h = K_{k_0} + \varepsilon_1 \sum_{t=1}^{k_0} \kappa_t^1 (1-c)^{k_0 - t} c\psi_t\psi_t^\top + \varepsilon_1 A_{k_0}^1(\varepsilon_1).$$

Likewise, $A_{k_0}^1(\varepsilon_1)$ defines a symmetric matrix which then tends to 0 when ε_1 tends to 0. Repeat these steps $k_0 - 1$ times with $\varepsilon_2, \dots, \varepsilon_{k_0-1} > 0$ instead of $\varepsilon_1 > 0$ and $\kappa_t^2, \dots, \kappa_t^{k_0-1} \in \mathbb{R}$ instead of $\kappa_t^1 \in \mathbb{R}$. All in all, we have finally, since $k = k_0(k_0-1)+1$,

$$K_{k-1}^h = K_{k-1} + \sum_{s=1}^{k_0-1} \left[\varepsilon_s \sum_{t=1}^{k_0} \kappa_t^s (1-c)^{k_0(k_0-1) - sk_0 + k_0 - t} c\psi_t\psi_t^\top + \varepsilon_s A_{k-1}^s(\varepsilon_s) \right].$$

Again, for each $s = 1, \dots, k_0 - 1$, $A_{k-1}^s(\varepsilon_s)$ defines a symmetric matrix which tends to 0 when ε_s tends to 0. Now, consider $(S_1, \dots, S_{k_0-1})$ a basis of $\ker \mathcal{D}\det(\mathbf{\Sigma}_{k-1})$. For $s = 1, \dots, k_0 - 1$, we set now the real values κ_t^s, $t = 1, \dots, k_0$ such that we have

$$\sum_{t=1}^{k_0} \kappa_t^s (1-c)^{k-1-sk_0+k_0-t} c\psi_s\psi_s^\top = S_s.$$

This is possible since $(\psi_1\psi_1^\top, \dots, \psi_{k_0}\psi_{k_0}^\top)$ is a basis of $\mathcal{S}^d$. Then,

$$K_{k-1}^h = K_{k-1} + \sum_{t=1}^{k_0-1} \varepsilon_t S_t + \varepsilon_t A_{k-1}^t(\varepsilon_t).$$

Yet, since $S_t \in \ker \mathcal{D}\det(\mathbf{\Sigma}_{k-1})$, we have then

$$\begin{aligned}\mathbf{\Sigma}_{k-1}^h = \frac{K_{k-1}^h}{\det(K_{k-1}^h)^{1/d}} &= \frac{K_{k-1} + \sum_{t=1}^{k_0-1} \varepsilon_t S_t + \varepsilon_t A_{k-1}^t(\varepsilon_t)}{\det\left(K_{k-1} + \sum_{t=1}^{k_0-1} \varepsilon_t S_t + \varepsilon_t A_{k-1}^t(\varepsilon_t)\right)^{1/d}} \\ &= \mathbf{\Sigma}_{k-1} + r\sum_{t=1}^{k_0-1} \varepsilon_t S_t + \varepsilon_t B^t(\varepsilon_t),\end{aligned}$$

where we set $r = \det(K_{k-1})^{-1/d}$, and the symmetric matrices $B^t(\varepsilon_t)$ tend to 0 when $\varepsilon_t \to 0$, for $t = 1, \dots, k_0 - 1$. Lastly, set $v_k = 0 \in \overline{\mathcal{O}^1_{\theta_{k-1}}}$, and let $h_k^1 = \cdots = h_k^\mu = \mathbf{\Sigma}_{k-1}^{-1/2}\xi_{k_0}$, for some arbitrary small vector $\xi_{k_0} \in \mathbb{R}^d$. Then, $\mathbf{\Sigma}_k = \mathbf{\Sigma}_{k-1}$ and

$$z_k^h = z_k + (1-c)^{-1/2}\xi_{k_0} + l(\varepsilon_1, \dots, \varepsilon_{k_0-1}) + \|(\varepsilon_1, \dots, \varepsilon_{k_0-1}, \xi_{k_0})\| h(\varepsilon_1, \dots, \varepsilon_{k_0-1}, \xi_{k_0}),$$

where $l\colon \mathbb{R}^{k_0-1} \to \mathbb{R}^d$ is linear and $h(\varepsilon_1,\dots,\varepsilon_{k_0-1},\xi_{k_0})$ tends to 0 when $\|(\varepsilon_1,\dots,\varepsilon_{k_0-1},\xi_{k_0})\| \to 0$. Furthermore,

$$\mathbf{\Sigma}_k^h = \mathbf{\Sigma}_k + r\sum_{t=1}^{k_0-1}\varepsilon_t S_t + \varepsilon_t B^t(\varepsilon_t) + c\xi_{k_0}\xi_{k_0}^\top.$$

Finally,

$$\frac{S^k_{(0,I_d)}(v_{1:k}+h_{1:k}) - S^k_{(0,I_d)}(v_{1:k}) - \begin{pmatrix} (1-c)^{-1/2}\xi_{k_0} + l(\varepsilon_1,\dots,\varepsilon_{k_0-1}) \\ r\sum_{t=1}^{k_0-1}\varepsilon_t S_t \end{pmatrix}}{\|(\varepsilon_1,\dots,\varepsilon_{k_0-1},\xi_{k_0})\|} \tag{1.53}$$

tends to 0 when $\|(\varepsilon_1,\dots,\varepsilon_{k_0-1},\xi_{k_0})\| \to 0$. Therefore,

$$\mathcal{D}S^k_{(0,I_d)}(v_{1:k})h_{1:k} = \begin{pmatrix} (1-c)^{-1/2}\xi_{k_0} + l(\varepsilon_1,\dots,\varepsilon_{k_0-1}) \\ r\sum_{t=1}^{k_0-1}\varepsilon_t S_t \end{pmatrix} \tag{1.54}$$

defines a surjective map from W^k to $\mathbb{R}^d \times \ker \mathcal{D}\det(\mathbf{\Sigma}_k)$. Indeed, if $H_{\mathbf{\Sigma}} \in \ker \mathcal{D}\det(\mathbf{\Sigma}_k)$ and $h_z \in \mathbb{R}^d$, then there exist $\varepsilon_1,\dots,\varepsilon_{k_0-1} \in \mathbb{R}$ such that $r\sum_{t=1}^{k_0-1}\varepsilon_t S_t = H_{\mathbf{\Sigma}}$, and then there exists $\xi_{k_0} \in \mathbb{R}^d$ such that $\mathcal{D}S^k_{(0,I_d)}(v_{1:k})h_{1:k} = (h_z; H_{\mathbf{\Sigma}})$. □

A Background on manifolds

We call throughout the paper a manifold a smooth manifold. We recall below the definition of a manifold, and we refer to [9] for more details.

Definition 1.2 (Manifolds)**.** A topological space X is said to be a topological manifold of dimension n if it is a second countable Hausdorff space that is locally Euclidean of dimension n.
Note that X is said to be a Hausdorff space if for every pair of distinct points $x, y \in \mathsf{X}$, there exist neighborhoods U of x and V of y that are disjoint. Moreover, X is said to be second countable if there exists a countable basis, that is, a countable collection $\mathcal{B}$ of open subsets of X such that any open subset of X can be written as the union of sets in $\mathcal{B}$.
Finally, X is locally Euclidean when for every $x \in \mathsf{X}$, there exists a neighborhood U of x, an open set V of $\mathbb{R}^n$ and a homeormophism (i.e., a continuous bijection with a continuous reciprocal function) $\varphi\colon U \to V$. We call (φ, U) a chart around x.
Besides, a manifold X is said to be smooth if it is topological, locally Euclidean, and if every charts (φ, U) and (ψ, V) around any point $x \in \mathsf{X}$ are such that $\varphi \circ \psi^{-1}$ is continuously differentiable.

Given X a (n-dimensional) manifold, and $x \in \mathsf{X}$, we denote by $T_x\mathsf{X}$ the tangent space of X in x. We refer to [9, Chapter XI] or to [59, Chapter 1, Section 2] for a formal definition of tangent spaces.

We introduce the measurability on a smooth manifold via the following definition. We refer to [9, Chapter XII] for further details.

Definition 1.3. A subset $\mathsf{A} \subset \mathsf{X}$ is said to be measurable if for all $x \in \mathsf{X}$, there exists a chart around x denoted (φ, U) such that $\varphi(\mathsf{A} \cap U)$ is measurable (in $\mathbb{R}^n$).

B Clarke's generalized derivative of locally Lipschitz functions on manifolds

Clarke's generalized Jacobian is defined for locally Lipschitz functions $g\colon \mathbb{R}^n \to \mathbb{R}^m$ [30]. We define here the Clarke's derivative for locally Lipschitz functions $f\colon \mathsf{X} \to \mathsf{Y}$ where X and Y are smooth manifolds. First, let us define formally what a locally Lipschitz function between manifolds is.

Definition 1.4. Let X and Y be two manifolds, equipped with their distance functions d_{X} and d_{Y} respectively, and $f\colon \mathsf{X} \to \mathsf{Y}$ a function.

(i) f is said to be Lipschitz if there exists $L > 0$ such that for all $x, y \in \mathsf{X}$ we have $d_{\mathsf{Y}}(f(x), f(y)) \leqslant L \times d_{\mathsf{X}}(x, y)$.

(ii) f is said to be locally Lipschitz at $x \in \mathsf{X}$ if there exists a neighborhood U of x in X such that the restriction of f to U is Lipschitz.

As stated below, a function is locally Lipschitz if and only if it is locally Lipschitz in the charts.

Proposition 1.17. If $f\colon \mathsf{X} \to \mathsf{Y}$ is locally Lipschitz at $x \in \mathsf{X}$, then for all local charts (φ, U) of X around x and (ψ, V) of Y around $f(x)$, the function $\psi \circ f \circ \varphi^{-1}$ is locally Lipschitz at $\varphi(x)$.

Proof. See that both φ^{-1} and ψ are $\mathcal{C}^1$ hence are locally Lipschitz at all points of their domains. By composition we find that $\psi \circ f \circ \varphi^{-1}$ is locally Lipschitz at $\varphi(x)$. □

Rademacher's theorem [43, Theorem 3.2], states that a locally Lipschitz function is almost everywhere differentiable. This is easily extended to locally Lipschitz functions on manifolds.

Corollary 1.4 (Rademacher's theorem)**.** Let ζ_{X} be a measure on X, which is locally equivalent to the Lebesgue measure, that is, for any measurable subset A of X, then $\zeta_{\mathsf{X}}(A) = 0$ if and only if for every charts (φ, U) of X, $\mathrm{Leb} \circ \varphi(A \cap U) = 0$. Then, any function $f\colon \mathsf{X} \to \mathsf{Y}$ locally Lipschitz at every $x \in \mathsf{X}$, is differentiable ζ_{X}-almost everywhere.

Proof. Consider local charts (φ, U) of X around x and (ψ, V) of Y around $f(x)$. Let us prove that for ζ_{X}-almost every point y of U, f is differentiable at y. See that by Proposition 1.17, $g \coloneqq \psi \circ f \circ \varphi^{-1}$ is locally Lipschitz on $\varphi(U)$. Thus, by [43, Theorem 3.2], we have that g is differentiable Leb-almost everywhere on $\varphi(U)$. Thus, since φ and ψ^{-1} are $\mathcal{C}^1$, and since the measures ζ_{X} and $\mathrm{Leb} \circ \varphi$ are equivalent on U, then $f = \psi^{-1} \circ g \circ \varphi$ is differentiable ζ_{X}-almost everywhere. □

We give now the definition of Clarke's Jacobian for locally Lipschitz functions on Euclidean spaces.

Definition 1.5 (Clarke's generalized Jacobian)**.** Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz at $x_0 \in \mathbb{R}^n$. Define

$$\partial f(x_0) = \mathrm{conv}\left\{ \lim_{t\to\infty} \mathcal{D}f(x_t) \mid x_t \to x_0,\ f \text{ is differentiable in all } x_t \right\} \tag{1.55}$$

where $\mathcal{D}f(x_t) \in \mathbb{R}^{n\times m}$ is the Jacobian matrix of f at x_t (when defined) and conv denotes the convex hull.

We generalize now this definition to locally Lipschitz functions on manifolds.

Proposition and Definition 6 (Clarke's generalized Jacobian on manifolds)**.** Let X and Y be two manifolds. Let $f\colon \mathsf{X} \to \mathsf{Y}$ be locally Lipschitzian at $x_0 \in \mathsf{X}$. Let (φ, U) be a local chart of X around x_0 and (ψ, V) be a local chart of Y around $f(x_0)$. Define $g = \psi \circ f \circ \varphi^{-1}$. Then $g\colon \varphi(U) \to \psi(V)$ is locally Lipschitz at $\varphi(x_0)$, and we can define

$$\partial f(x_0) = \left\{ \mathcal{D}\psi^{-1}(g \circ \varphi(x_0)) \circ h \circ \mathcal{D}\varphi(x_0) \mid h \in \partial g(\varphi(x_0)) \right\}. \tag{1.56}$$

Proof. The maps ψ and φ^{-1} are by definition continuously differentiable, hence are locally Lipschitz. Therefore, by composition, g is locally Lipschitz. Furthermore, note that the expression (1.56) does not depend on the choice of the charts. Indeed, let (φ_1, U) and (φ_2, U) be two charts of X at x_0 and (ψ_1, V) and (ψ_2, V) be two charts of Y at $f(x_0)$, such that $g_1 = \psi_1 \circ f \circ \varphi_1^{-1}$ and $g_2 = \psi_2 \circ f \circ \varphi_2^{-1}$ are well defined. Then, note that $g_2 = \psi_2 \circ \psi_1^{-1} \circ g_1 \circ \varphi_1 \circ \varphi_2^{-1}$. Apply then the chain rule [30, Corollary of Theorem 2.6.6] to g_2 and get

$$\begin{aligned}&\partial g_2(\varphi_2(x_0)) \\ &\quad = \left\{ \mathcal{D}\psi_2(f(x_0))\mathcal{D}\psi_1^{-1}(g_1(\varphi_1(x_0))) \circ h \circ \mathcal{D}\varphi_1(x_0)\mathcal{D}\varphi_2^{-1}(\varphi_2(x_0)) \mid h \in \partial g_1(\varphi_1(x_0)) \right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\left\{ \mathcal{D}\psi_1^{-1}(g_1 \circ \varphi_1(x_0)) \circ h \circ \mathcal{D}\varphi_1(x_0) \mid h \in \partial g_1(\varphi_1(x_0)) \right\} \\ &\qquad = \left\{ \mathcal{D}\psi_2^{-1}(g_2 \circ \varphi_2(x_0)) \circ h \circ \mathcal{D}\varphi_2(x_0) \mid h \in \partial g_2(\varphi_2(x_0)) \right\}.\end{aligned}$$

□

We also state the next result, which would be useful to prove Proposition 1.12.

Proposition 1.18. If $f\colon \mathsf{X} \to \mathsf{Y}$ is locally Lipschitzian at $x_0 \in \mathsf{X}$, then $\partial f(x_0)$ is nonempty, compact and convex.

Proof. This follows from [30, Proposition 2.6.2(a)] and Proposition and Definition 1. □

We now transpose the uppercontinuity of Clarke's Jacobians to the context of locally Lispchitz functions between manifolds.

Proposition 1.19. Let $f\colon \mathsf{X} \to \mathsf{Y}$ be locally Lipschitz at x_0. Then, $\lim_{x \to x_0} \partial f(x) \subset \partial f(x_0)$.

Proof. Let (φ, U) and (ψ, V) be two local charts respectively of X and Y at x_0 and $f(x_0)$, such that $\tilde{f} = \psi \circ f \circ \varphi^{-1}$ is well defined. Then, by Proposition and Definition 6, we have, for any $x \in U$

$$\partial f(x) = \left\{ \mathcal{D}\psi^{-1}(\tilde{f} \circ \varphi(x)) \circ h \circ \mathcal{D}\varphi(x) \mid h \in \partial \tilde{f}(\varphi(x)) \right\}.$$

By applying [30, Proposition 2.6.2] to $\tilde{f}$, we find that $\lim_{x \to x_0} \partial \tilde{f}(\varphi(x)) \subset \partial \tilde{f}(x_0)$, which ends the proof. □

The next proposition is actually a very important requirement for our analysis. It states that if we

can find a point for which the generalized differential of a locally Lipschitz function in this point is of maximal rank, then we can find a point closed to it in which the function is differentiable and the derivative is full rank.

Proposition 1.20. Suppose that $f\colon \mathsf{X} \to \mathsf{Y}$ is locally Lipschitzian at $x_0 \in \mathsf{X}$, and that $\partial f(x_0)$ is of maximal rank, i.e., all $h \in \partial f(x_0)$ is of maximal rank. Then, there exists a neighborhood U of x_0 such that for all $y \in U$, $\partial f(y)$ is of maximal rank. Moreover, for all neighborhood $V \subset U$ of x_0, there exists $y_0 \in V$ such that f is differentiable at y_0 and $\mathcal{D}f(y_0)$ is of maximal rank.

Proof. Let $A = \{h \in \mathcal{L}(T_{x_0}\mathsf{X}, T_{f(x_0)}\mathsf{Y}) \mid h \text{ is not of maximal rank}\}$. Since the application rank is l.s.c., then A is a closed set. By Proposition 1.18, $\partial f(x_0)$ is compact, and disjoint from A since it is assumed to be of maximal rank. Thus $d(\partial f(x_0), A) > 0$, where d is a metric induced by some norm on the affine space $\mathcal{L}(T_{x_0}\mathsf{X}, T_{f(x_0)}\mathsf{Y})$ of finite dimension. Moreover, there exits $h^* \in \partial f(x_0)$ such that for all $h \in \partial f(x_0)$ we have

$$d(h, A) \geqslant d(h^*, A) = d(\partial f(x_0), A) > 0,$$

By [30, Proposition 2.6.2(c)], there exists a neighborhood U of x_0 such that for all $y \in U$, $d(\partial f(y), A) \geqslant d(h^*, A)/2 > 0$, thus $\partial f(y)$ is of maximal rank. The second part follows from Corollary 1.4. □

Next, we state a chain rule for the generalized Jacobian on manifolds.

Proposition 1.21 (Chain rule)**.** Let X, Y and Z be three manifolds. If $f\colon \mathsf{X} \to \mathsf{Y}$ is locally Lipschitz at $x_0 \in \mathsf{X}$, and if $g\colon \mathsf{Y} \to \mathsf{Z}$ is differentiable at $f(x_0)$, then we have

$$\partial(g \circ f)(x_0) = \{\mathcal{D}g(f(x_0))h \mid h \in \partial f(x_0)\}.$$

Proof. Let (φ, U), (ψ, V) and (ν, W) be local charts respectively of X, Y and Z around x_0, $f(x_0)$ and $g \circ f(x_0)$. Define $\tilde{f} = \psi \circ f \circ \varphi^{-1}\colon \varphi(U) \to \psi(v)$ and $\tilde{g} = \nu \circ g \circ \psi^{-1}\colon \psi(V) \to \nu(W)$. Then, by Proposition and Definition 6, we obtain

$$\partial(g \circ f)(x_0) = \left\{\mathcal{D}\nu^{-1}(g \circ f(x_0)) \circ H \circ \mathcal{D}\varphi(x_0) \mid H \in \partial(\tilde{g} \circ \tilde{f})(\varphi(x_0))\right\}.$$

Now we apply the chain rule from [30, Corollary of Theorem 2.6.6]. Since $\tilde{g}$ is differentiable at $\tilde{f}(\varphi(x_0))$ and $\tilde{f}$ is locally Lipschitz at $\varphi(x_0)$, we have then

$$\partial(\tilde{g} \circ \tilde{f})(\varphi(x_0)) = \left\{\mathcal{D}\tilde{g}(\tilde{f}(\varphi(x_0))) \circ H \mid H \in \partial\tilde{f}(\varphi(x_0))\right\}.$$

Then,

$$\begin{aligned}\partial(g \circ f)(x_0) &= \left\{\mathcal{D}\nu^{-1}(g \circ f(x_0)) \circ \mathcal{D}\tilde{g}(\tilde{f}(\varphi(x_0))) \circ H \circ \mathcal{D}\varphi(x_0) \mid H \in \partial\tilde{f}(\varphi(x_0))\right\} \\ &= \left\{\mathcal{D}g(f(x_0)) \circ \mathcal{D}\psi^{-1}(f(x_0)) \circ H \circ \mathcal{D}\varphi(x_0) \mid H \in \partial\tilde{f}(\varphi(x_0))\right\} \\ &= \{\mathcal{D}g(f(x_0)) \circ H \mid H \in \partial f(x_0)\},\end{aligned}$$

the last line being obtained by applying Proposition and Definition 6 to f. □

Lastly, the next two theorems are extensions of the inverse function theorem and of the implicit function theorem to our context.

Theorem 1.9 (Inverse function theorem)**.** Let X and Y be two manifolds of dimension n. Let $f\colon \mathsf{X} \to \mathsf{Y}$ be locally Lipschitzian at $x_0 \in \mathsf{X}$. Suppose that $\partial f(x_0)$ is of maximal rank, i.e., for all $h \in \partial f(x_0)$, we have $\operatorname{rank} h = n$. Then, there exist a neighborhood of x_0 in X, a neighborhood V of $f(x_0)$ in Y and a Lipschitzian function $g\colon V \to U$ such that

(i) $g(f(u)) = u$ for all $u \in U$;

(ii) $f(g(v)) = v$ for all $v \in V$.

Proof. Let (φ, U) be a local chart of X around x_0 and (ψ, V) a local chart of Y around $f(x_0)$. Define then $\tilde{f} = \psi \circ f \circ \varphi^{-1}$. Since $\partial f(x_0)$ is of maximal rank, by the chain rule, using Proposition and Definition 1, then $\partial \tilde{f}(\varphi(x_0))$ is of maximal rank. Then, up to taking U and V smaller, by the Inverse function theorem applied to $\tilde{f}$ as stated in [30, Theorem 7.1.1], then there exists a Lipschitz function $\tilde{g}\colon \psi(V) \to \varphi(U)$ such that $\tilde{g}(\tilde{f}(\tilde{u})) = \tilde{u}$ for $\tilde{u} \in \varphi(U)$ and $\tilde{f}(\tilde{g}(\tilde{v})) = \tilde{v}$ for $\tilde{v} \in \varphi(V)$. Define then $g = \varphi^{-1} \circ \tilde{g} \circ \psi\colon U \to V$ to get

$$g(f(u)) = \varphi^{-1} \circ \tilde{g} \circ \psi(f(u)) = \varphi^{-1} \circ g(\tilde{f}(\varphi(u))) = \varphi^{-1} \circ \varphi(u) = u$$

for all $u \in U$, and

$$f(g(v)) = f \circ \varphi^{-1} \circ \tilde{g} \circ \psi(v) = \psi^{-1} \circ \tilde{f}(\tilde{g}(\psi(v))) = \psi^{-1} \circ \psi(v) = v$$

for all $v \in V$. □

Theorem 1.10 (Implicit function theorem)**.** Let X, Y and Z be manifolds of dimensions respectively m, k and k. Let $f\colon \mathsf{X} \times \mathsf{Y} \to \mathsf{Z}$ be locally Lipschitzian at $(x_0, y_0) \in \mathsf{X} \times \mathsf{Y}$. Moreover, assume that the partial generalized differential $\partial_y f(x_0, y_0)$ is of maximal rank. Then there exists a neighborhood U of x_0 and a Lipschitz function $g\colon U \to \mathsf{Y}$ such that $g(x_0) = y_0$, and for all $x \in U$,

$$f(x, g(x)) = f(x_0, y_0). \tag{1.57}$$

Proof. Define $F(x, y) = (x, f(x, y))$ a function $\mathsf{X} \times \mathsf{Y} \to \mathsf{X} \times \mathsf{Z}$, which is locally Lipschitz at (x_0, y_0). Define $n = m + k$ and note that the dimensions of $\mathsf{X} \times \mathsf{Y}$ and $\mathsf{X} \times \mathsf{Z}$ both equal n. Besides, since $\partial_y f(x_0, y_0)$ is of maximal rank, we find that $\partial F(x_0, y_0)$ is of maximal rank. Thus we can apply the inverse function theorem to F and find neighborhoods U, V, and W respectively of x_0 in X, y_0 in Y and $f(x_0, y_0)$ in Z, as well as a Lipschitz function $G\colon U \times W \to U \times V$ such that for all $(x, z) \in U \times W$ we have

$$F(G(x, z)) = (x, z).$$

Note that then $G(x, z) = (x, \tilde{G}(x, z))$ for some $\tilde{G}(x, z) \in V$, so that $f(x, \tilde{G}(x, z)) = z$. Therefore, define $g(x) = \tilde{G}(x, f(x_0, y_0))$ to get

$$f(x, g(x)) = f(x_0, y_0).$$

□

C Additional proofs

Proposition 1.22. Consider an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ which is scaling-invariant w.r.t. $x^* = 0$. Then, the sequence $\{z_k, \mathbf{\Sigma}_k\}_{k\in\mathbb{N}}$ defined by (1.9) is a time-homogeneous Markov chain which follows (1.3), with functions F and α defined by (1.10) and (1.11) respectively.

Proof. Let $i = 1, \dots, \lambda$. Then, we have

$$f\left(z_k + \sqrt{\mathbf{\Sigma}_k} U_{k+1}^i\right) = f\left(R(\mathbf{C}_k)^{-1/2} \times \left[m_k + \sqrt{\mathbf{C}_k} U_{k+1}^i\right]\right).$$

Since f is scaling-invariant, this implies that the permutation s_{k+1} satisfies almost surely that

$$f\left(z_k + \sqrt{\mathbf{\Sigma}_k} U_{k+1}^{s_{k+1}(1)}\right) \leqslant \cdots \leqslant f\left(z_k + \sqrt{\mathbf{\Sigma}_k} U_{k+1}^{s_{k+1}(\lambda)}\right).$$

Let $k \geqslant 1$, and observe that

$$\mathbf{K}_{k+1} \coloneqq \mathbf{\Sigma}_k + \sqrt{\mathbf{\Sigma}_k} \sum_{i=1}^{\mu} w_i \left[U_{k+1}^{s_{k+1}(i)}\right] \left[U_{k+1}^{s_{k+1}(i)}\right]^\top \sqrt{\mathbf{\Sigma}_k} = R(\mathbf{\Sigma}_k)^{-1} \mathbf{\Sigma}_{k+1} = R(\mathbf{C}_k)^{-1} \mathbf{C}_{k+1}.$$

Since $R = \det^{1/d}(\cdot)$ is (positively) homogeneous $R(\mathbf{K}_{k+1}) = R(\mathbf{C}_k)^{-1} R(\mathbf{C}_{k+1})$. Furthermore, we have that

$$\begin{aligned} z_{k+1} &= \left(R(\mathbf{C}_{k+1})\right)^{-1/2} \times m_{k+1} \\ &= R(\mathbf{C}_k)^{1/2} R(\mathbf{C}_{k+1})^{-1/2} R(\mathbf{C}_k)^{-1/2} \times \left[m_k + \sqrt{\mathbf{C}_k} \sum_{i=1}^{\mu} w_i U_{k+1}^{s_{k+1}(i)}\right] \\ &= R(\mathbf{K}_{k+1})^{-1/2} \times \left[m_k + \sqrt{\mathbf{\Sigma}_k} \sum_{i=1}^{\mu} w_i^m U_{k+1}^{s_{t+1}(i)}\right]. \end{aligned}$$

Moreover,

$$\mathbf{\Sigma}_{k+1} = R(\mathbf{C}_{k+1})^{-1} \mathbf{C}_{k+1} = R(\mathbf{K}_{k+1})^{-1} \mathbf{K}_{k+1}$$

All in all, we have

$$(z_{k+1}, \mathbf{\Sigma}_{k+1}) = F\left((z_k, \mathbf{\Sigma}_k), \alpha\left((z_k, \mathbf{\Sigma}_k), \left(U_{k+1}^1, \dots, U_{k+1}^\lambda\right)\right)\right),$$

ending the proof. □

Proof of Proposition 1.1 and Proposition 1.3. Let $\theta = (z, \mathbf{\Sigma}) \in \mathsf{X}$, consider i.i.d. random variables $U^1, \dots, U^\lambda \sim \mathcal{N}(0, I_d)$, and let $U = (U^1, \dots, U^\lambda)$. Then $W = \alpha(\theta, U)$ satisfies a.s.

$$W = \sum_{\sigma\in\mathfrak{S}_\lambda} \mathbb{1}\left\{f\left(z + \sqrt{\mathbf{\Sigma}} U^{\sigma(1)}\right) < \cdots < f\left(z + \sqrt{\mathbf{\Sigma}} U^{\sigma(\lambda)}\right)\right\} \times \left(U^{\sigma(1)}, \dots, U^{\sigma(\mu)}\right)$$

where $\mathfrak{S}_\lambda$ is the set of permutations of $\{1, \dots, \lambda\}$.

Hence, by symmetry,

$$W = \frac{1}{(\lambda-\mu)!}\sum_{\sigma\in\mathfrak{S}_\lambda} \mathbb{1}\left\{f\left(z+\sqrt{\Sigma}U^{\sigma(1)}\right) < \cdots < f\left(z+\sqrt{\Sigma}U^{\sigma(\mu)}\right)\right\} \times \prod_{k=\mu+1}^{\lambda} \mathbb{1}\left\{f\left(z+\sqrt{\Sigma}U^{\sigma(\mu)}\right) < f\left(z+\sqrt{\Sigma}U^{\sigma(k)}\right)\right\} \times \left(U^{\sigma(1)},\dots,U^{\sigma(\mu)}\right).$$

Let $\eta\colon \mathsf{W} \to \mathbb{R}_+$ be a smooth map with compact support. We obtain

$$\mathbb{E}\left[\eta(W)\right] = \frac{1}{(\lambda-\mu)!}\sum_{\sigma\in\mathfrak{S}_\lambda}\int \mathbb{1}\left\{f\left(z+\sqrt{\Sigma}u_{\sigma(1)}\right) < \cdots < f\left(z+\sqrt{\Sigma}u_{\sigma(\mu)}\right)\right\} \times \prod_{k=\mu+1}^{\lambda} \mathbb{1}\left\{f\left(z+\sqrt{\Sigma}u_{\sigma(\mu)}\right) < f\left(z+\sqrt{\Sigma}u_{\sigma(k)}\right)\right\} \times \eta\left(u_{\sigma(1)},\dots,u_{\sigma(\mu)}\right)\gamma^d(u_1)\dots\gamma^d(u_\lambda)\mathrm{d}u_1\dots\mathrm{d}u_\lambda.$$

However, observe that, for each $k = \mu+1,\dots,\lambda$, we have

$$\int \mathbb{1}\left\{f\left(z+\sqrt{\Sigma}u_{\sigma(\mu)}\right) < f\left(z+\sqrt{\Sigma}u_{\sigma(k)}\right)\right\}\gamma^d(u_{\sigma(k)})\mathrm{d}u_{\sigma(k)} = 1 - Q_\theta^f\left(u_{\sigma(\mu)}\right).$$

We deduce then the desired result. Note that Proposition 1.3 is obtained by taking $\Sigma = I_d$. □

Proof of Proposition 1.6. First observe that (1.35) is equivalent to (iii). Indeed, if (iii) holds, then for every $y \in \mathsf{X}$ there exists a sequence $\{y_k\}_{k\in\mathbb{N}}$ with $y_k \in A_+^k(y)$, and with a subsequence converging to x^*. In that case, for every $T \geqslant 1$, and for every neighborhood N of x^*, there exist infinitely many indices $k \geqslant T$ such that $y_k \in N$. Thus, every neighborhood N of x^* intersects $\cap_{T\geqslant 1}\cup_{k\geqslant T} A_+^k(y)$, which proves that (1.35) holds. Conversely, assume that (1.35) holds. Then, for every $T \geqslant 1$, there exists $k \geqslant T$ such that $x^* \in \overline{A_+^k(y)}$. Then, consider $y_k \in A_+^k(y)$ such that $\mathrm{dist}_{\mathsf{X}}(x^*, y_k) \leqslant 1/k$. This proves (iii).
Now, suppose (iii) and let us prove (i). Let $y \in \mathsf{X}$. By (iii), we know that there exists a sequence $\{y_k\}_{k>0}$ such that $y_k \in A_+^k(y)$ and with a subsequence which converge to x^*. However, $A_+^k(y) \subset A_+(y)$, therefore $\{y_k\}_{k>0}$ is a sequence with values in $A_+(y)$ admitting x^* as an accumulation point, which proves (i).
Next, assume that (i) holds, and let us prove that this implies (ii). Let $y \in \mathsf{X}$. By (i), $x^* \in \overline{A_+(y)}$. In other words, for any open U of X containing x^*, we have $U \cap A_+(y) \neq \emptyset$. Let U be such an open subset. Since $A_+(y) = \cup_{k\in\mathbb{N}} A_+^k(y)$, then there exists $k \in \mathbb{N}$ such that $A_+^k(y)$ intersects U. If $k \geqslant 1$, this proves (ii). Else, if $k = 0$, we do the same reasoning with $z = S_y^1(w_1)$ for some $w_1 \in \mathcal{O}_y^1$, which proves (ii).
Last, let us prove that (ii) implies (iii). Suppose (ii), let $y \in \mathsf{X}$, and let $\{k_t\}_{t\geqslant 1}$ be an increasing sequence of $\mathbb{N}_{>0}$ which satisfies that, for every $t \geqslant 1$, there exists a k_t-steps path from y to $B(x^*, 1/t)$. Hence, let $\{y_k\}_{k>0}$ be a sequence such that $y_k \in A_+^k(y)$ for every $k > 0$, and with $y_{k_t} \in B(x^*, 1/t)$. Then, the subsequence $\{y_{k_t}\}_{t>0}$ converges to x^*, proving (iii). □

Proof of Proposition 1.7. Let U be an open subset of X, $x \in \mathsf{X}$ and $k > 0$. First let us assume

that $P^k(x,U) > 0$. However, note that we have

$$P^k(x,U) = \int_{\mathcal{O}_x^k} \mathbb{1}\{S_x^k(w_{1:k}) \in U\} p_x^k(w_{1:k}) \mathrm{d}\zeta_{\mathsf{W}}(w_1) \dots \mathrm{d}\zeta_{\mathsf{W}}(w_k),$$

which implies that there exists $w_{1:k} \in \mathcal{O}_x^k$ such that $S_x^k(w_{1:k}) \in U$, hence $w_{1:k}$ is a k-steps path from x to U. Conversely, assume that there exists $w_{1:k}$ a k-steps path from x to U, i.e., that $S_x^k(w_{1:k}) \in U$. Since F is continuous, then S_x^k is continuous as well. Therefore there exists an open subset V of W^k such that for all $v_{1:k} \in V$, $S_x^k(v_{1:k})$. Then we obtain

$$P^k(x,U) \geqslant \int_{\mathcal{O}_x^k \cap V} \mathbb{1}\{S_x^k(w_{1:k}) \in U\} p_x^k(w_{1:k}) d\zeta_{\mathsf{W}}(w_1) \dots d\zeta_{\mathsf{W}}(w_k) > 0,$$

since $\mathcal{O}_x^k \cap V$ is open by intersection, and p_x^k is l.s.c. □

Proof of Proposition 1.8. The statement (i) is a consequence of Proposition 1.6 (ii).
For (ii), let $y \in \mathsf{X}$, and for every integer $s \geqslant 1$, consider the open subset $U_s = B(x^*, 1/s)$ of X. Then, there exists a nondecreasing sequence $\{T_s\}_{s\geqslant 1}$, such that for every $k \geqslant T_s$, there exists a k-steps path $w_{1:k}^k \in \mathcal{O}_y^k$ from y to U_s. For $k \in \mathbb{N}$, define $y_k = S_y^k(w_{1:k}^k)$, and observe that $y_k \in A_+^k(y)$. Moreover, we have $y_k \in U_s$ for every $k \in \{T_s, \dots, T_{s+1} - 1\}$. Then, the sequence $\{y_k\}_{k\in\mathbb{N}}$ converges to x^*. Conversely, suppose that for every $y \in \mathsf{X}$, there exists a sequence $\{y_k\}_{k\in\mathbb{N}}$ converging to x^* with $y_k \in \overline{A_+^k(y)}$. Hence, for every $k > 0$ there exists $z_k \in A_+^k(y) \cap B(y_k, 1/k)$. By definition of $A_+^k(y)$, there exists then $w_{1:k}^k \in \mathcal{O}_y^k$ such that $z_k = S_y^k(w_{1:k}^k)$. Besides, since y_k tends to x^*, then z_k tends to x^* as well. Let U be a neighborhood of x^*, so that there exists $T \in \mathbb{N}$ with $z_k \in U$ when $k \geqslant T$. Then, for $k \geqslant T$, $w_{1:k}^k$ is a k-steps path from y to U. Thus x^* is steadily attracting, proving (ii).
For (iii), suppose that there exist x^* a steadily attracting state and y^* a globally attracting state. Let us prove that y^* is steadily attracting. Let U be a neighborhood of y^* in X, and let $z \in \mathsf{X}$. Since y^* is globally attracting, there exist $k > 0$ and a k-steps path $w_{1:k} \in \mathcal{O}_{x^*}^k$ from x^* to U, i.e. such that $S_{x^*}^k(w_{1:k}) \in U$. Since F is continuous, then $S_{x^*}^k$ is continuous, and thus there exists a neighborhood V of x^* such that for every $x \in V$, we have $S_x^k(w_{1:k}) \in V$. Moreover, $x \mapsto p_x^k(w_{1:k})$ is lower semicontinuous, so up to taking V a smaller neighborhood of x^*, we can assume that $w_{1:k} \in \mathcal{O}_x^k$. Last, x^* is steadily attracting, so there exists $T > 0$ such that for every $t \geqslant T$, there exists a t-steps path $v_{1:t}$ from z to V, i.e., $S_z^t(v_{1:t}) \in V$. All in all, for every $t \geqslant T$, there exists a $(t+k)$-steps path $[v_{1:t}, w_{1:k}]$ from z to U, ending the proof. □

Proof of Corollary 1.2. The inclusion $\{S_x^k(w_{1:k}) \mid w_{1:k} \in \overline{\mathcal{O}_x^k}\} \subset \overline{A_+^k(y)}$ follows directly from the definition of $A_+^k(y)$ and the continuity of F. Let $x^* \in \mathsf{X}$ and assume that (i) x^* is steadily attracting. Then, by definition, for every $x \in \mathsf{X}$ and every neighborhood U of x^*, there exists $T > 0$ such that for every $k \geqslant T$ there is a k-steps path from x to U, hence (iii) holds.
Next, assume that (iii) for every $x \in \mathsf{X}$ and every neighborhood U of x^*, there exists $T > 0$ such that for every $k \geqslant T$ we can find $w_{1:k} \in \overline{\mathcal{O}_x^k}$ with $S_x^k(w_{1:k}) \in U$. Then, as in the previous proof, for every integer $s \geqslant 1$, consider the open subset $U_s = B(x^*, 1/s)$ of X. Therefore there exists a nondecreasing sequence $\{T_s\}_{s\geqslant 1}$, such that for every $k \geqslant T_s$, there exists $w_{1:k} \in \overline{\mathcal{O}_x^k}$ with $S_x^k(w_{1:k}) \in U_s$. So we find a sequence $\{y_k\}_{k\in\mathbb{N}}$ such that $y_k \in \{S_x^k(w_{1:k}) \mid w_{1:k} \in \overline{\mathcal{O}_x^k}\}$ for $k \in \mathbb{N}$, and with $y_k \in U_s$ for $k \in \{T_s, \dots, T_{s+1} - 1\}$, which proves (ii).
Last, observe that the implication '(ii) implies (i)' follows directly from Proposition 1.8(ii) and

the inclusion $\{S_x^k(w_{1:k}) \mid w_{1:k} \in \overline{\mathcal{O}_x^k}\} \subset \overline{A_+^k(y)}$. □

Proof of Proposition 1.9. First, we prove (i). Observe that E is nonempty. Indeed, x^* is attainable, so there exist $a \in \mathbb{N}^*$ and $w_{1:a} \in \mathcal{O}_{x^*}^a$ with $x^* = S_{x^*}^a(w_{1:a})$, and so $x^* = S_{x^*}^{ak}(w_{1:a}, \dots, w_{1:a})$. Consider now a and b two elements of E and let us prove that $d := \gcd(a,b) \in E$. By definition, there exist $T_a, T_b > 0$, such that for every $k \geqslant T_a$, $x^* \in A_+^{ak}(x^*)$, and for every $k \geqslant T_b$, $x^* \in A_+^{bk}(x^*)$. Let $T \in \mathbb{N}$ be larger than a/d, so that for every $k \geqslant 0$, the Euclidean division of $(T+k)d$ by a provides us q_a and r two integers such that $(T+k)d = q_a a + r$. Besides, by Bézout's theorem, we find that $r = q_b b$ for some $q_b \in \mathbb{N}$, hence $(T+k)d = q_a a + q_b b$. However, by definition of T_a and T_b, we have $x^* \in A_+^{(q_a+T_a)a}(x^*)$ and $x^* \in A_+^{(q_b+T_b)b}(x^*)$. All in all, $x^* \in A_+^{(T+k)d+T_a a+T_b b}$, proving (i) since d divides a and b.
To prove (ii), observe that if $\gcd(E) = 1$, then, by (i), we have $1 \in E$. Then, there exists $T \in \mathbb{N}$ such that for all $k \geqslant T$, $x^* \in A^k(x^*)$. Let $y \in \mathsf{X}$. Since x^* is attainable, there exists $t \in \mathbb{N}$ such that $x^* \in A^t(y)$, so that for all $k \geqslant T + t$, $x^* \in A^k(y)$. Thus, x^* is steadily attracting.
For (iii), define $d = \gcd(E)$, and let $D_i = \cup_{r \geqslant 0} A_+^{rd+1}(x^*)$ for $i \in \{0, \dots, d-1\}$. First observe that the D_i are disjoint sets. Indeed, $A_+^i(x^*)$ intersects $A_+^j(x^*)$ for some integers i, j, then there exists y in their intersection. As x^* is attainable, there exists $k > 0$ such that $x^* \in A_+^k(y)$, hence $x^* \in A_+^{r(k+i)}(x^*)$ and $x^* \in A_+^{r(k+j)}(x^*)$ for all $r \geqslant 0$. This implies that d divides both $k+i$ and $k+j$ hence d divides $i-j$. This shows that the sets D_i, $i \in \{0, \dots, d-1\}$, are disjoint. Moreover, by construction, we have $P(y, D^{i+1}) = 1$ for all $i \in \mathbb{Z}/d\mathbb{Z}$. Finally, observe that the union of the D_i, $i \in \{0, \dots, d-1\}$, is equal to $A_+(x^*)$. Since P is φ-irreducible, and $P^k(x^*, A_+(x^*)) = 1$ for all $k \in \mathbb{N}$, then the support of φ is included in $A_+(x^*)$. All in all, we have that $\{D_i\}_{0 \leqslant i \leqslant d-1}$ is a d-cycle. □

Proof of Proposition 1.16. First we prove (1.50). Let $x^* \in \operatorname{supp} \psi$, and let U be a neighborhood of x^*. Then, $\psi(U) > 0$, which implies that for every $y \in \mathsf{X}$, $\sum_{k \geqslant 0} P^k(y, U) > 0$. This is true for every neighborhood U of x^*, hence, by Proposition 1.7 and Proposition 1.6, x^* is globally attracting.
Conversely, let $x^* \in \mathsf{X}$ be a globally attracting state. Then, by Proposition 1.7 and Proposition 1.6, for every neighborhood U of x^* and for every $y \in \mathsf{X}$, there exists $k > 0$ such that $P^k(y, U) > 0$, hence $\psi(U) > 0$. This implies that $x^* \in \operatorname{supp} \psi$. All in all, we obtain (1.50).
Now, let us prove (1.51). Consider $x^* \in \mathsf{X}$ a globally attracting state, and let $y^* \in \operatorname{supp} \psi$. By (1.50), y^* is then a globally attracting state. Therefore, by Proposition 1.6, $y^* \in \overline{A_+(x^*)}$.
Conversely, let $y^* \in \overline{A_+(x^*)}$ and let us prove that y^* is globally attracting. Let U be a neighborhood of y^*, so that U intersects $A_+(x^*)$. This implies that there exist $k \geqslant 1$ and $w_{1:k} \in \mathcal{O}_{x^*}^k$ such that $S_{x^*}^k(w_{1:k}) \in U$. Since F is continuous, then $S_x^k(w_{1:k}) \in U$ for every x in a neighborhood V of x^*. Besides, $x \mapsto p_x^k(w_{1:k})$ is l.s.c., so, up to taking V smaller, we can assume that $w_{1:k} \in \mathcal{O}_x^k$ for every $x \in V$. Furthermore, x^* is globally attracting, so, for every $z \in \mathsf{X}$, there exist $t \in \mathbb{N}$ and $v_{1:t} \in \mathcal{O}_z^t$ such that $S_z^t(v_{1:t}) \in V$, hence such that $[v_{1:t}, w_{1:k}] \in \mathcal{O}_z^{t+k}$ and $S_z^{t+k}(v_{1:t}, w_{1:k})$, proving that y^* is globally attracting. This ends the proof, using (1.50). □

Chapter 2

Irreducibility of nonsmooth state-space models with an application to CMA-ES

Comments on Chapter 2: This chapter includes the paper "Irreducibility of nonsmooth state-space models with an application to CMA-ES" (Armand Gissler, Shan-Conrad Wolf, Anne Auger, Nikolaus Hansen) submitted in 2024 to the journal *Stochastic Processes and Their Applications* [53]. The main goals of this chapter is to introduce a normalized Markov chain underlying CMA-ES, that we study throughout the manuscript, and to prove that this chain is an irreducible and aperiodic T-chain. This result is exposed in Theorem 2.1. Chapter 4 will complete the stability analysis of the normalized Markov chain by the proof of geometric ergodicity when the objective function is ellipsoidal.

To prove the irreducibility, aperiodicity and T-chain property, we rely on the methodology introduced in Chapter 1. Indeed, we formulate the update of the normalized Markov chain as a nonsmooth state-space model. However, we encounter two limitations with this methodology. First, in order to include several hyperparameter settings of CMA-ES, we have to study different Markov chains. Thus, we introduce in this chapter the notion of redundant and projected Markov chains. Specifically, we show in Theorem 2.3 how similar assumptions than those in Chapter 1 on the redundant model yields to the irreducibility of the projected Markov chain. Second, when the normalization of CMA-ES is a nonsmooth function (as we will have in Chapter 4), the state space of the normalized Markov chain is then a nonsmooth manifold. However, one of the assumptions of Chapter 1 is that the state space is a smooth manifold. We explain in Theorem 2.4 how we can use the results of Chapter 1 on nonsmooth state spaces.

Abstract

We analyze a stochastic process resulting from the normalization of states in the zeroth-order optimization method CMA-ES. On a specific class of minimization problems where the objective function is scaling-invariant, this process defines a time-homogeneous Markov chain whose convergence at a geometric rate can imply the linear convergence of CMA-ES. However, the analysis of the intricate updates for this process constitute a great mathematical challenge. We establish that this Markov chain is an irreducible and aperiodic T-chain. These contributions represent a first major step for the convergence analysis towards a stationary distribution. We rely for this analysis on conditions for the irreducibility of nonsmooth state-space models on manifolds. To obtain our results, we extend these conditions to address the irreducibility in different hyperparameter settings that define different Markov chains, and to include nonsmooth state spaces.

1 Introduction . . . 58
2 Definition of Markov chains arising from a normalization of CMA-ES . . . 60
2.1 Presentation of CMA-ES . . . *61*
2.2 Assumptions . . . *63*
2.3 Proving the stability of a normalized Markov chain leads to linear convergence . . . *64*
3 Main Results I: Irreducibility, aperiodicity, and T-chain property of normalized Markov chains underlying the CMA-ES algorithm . . . 69
4 Main results II: Extension of the analysis of nonlinear state-space models . . . 71
4.1 Deterministic control model and sufficient conditions for irreducibility and aperiodicity *71*
4.2 Irreducibility and aperiodicity of a projected Markov chain . . . *73*
4.3 Homeomorphic transformation of an irreducible aperiodic T-chain . . . *76*
5 Proof of Theorem 2.1 . . . 77
5.1 Definition of normalized chains underlying CMA-ES following (2.25) *and satisfying* **H**1-**H**2 . . . *77*
5.2 Finding steadily attracting states . . . *82*
5.3 Controllability condition . . . *86*
5.4 Proof of Theorem 2.1 . . . *90*
6 Conclusion and perspectives . . . 91
A Proofs in Section 5.2 . . . 92
A.1 Proof of Lemma 2.8 . . . *92*
B Proofs in Section 5.3 . . . 94
B.1 Proof of Lemma 2.12 . . . *94*
B.2 Proof of Lemma 2.13 . . . *99*
B.3 Proof of Proposition 2.12 . . . *104*

1 Introduction

The convergence of stochastic processes is at the core of many algorithms in various domains. Well-known examples include Markov chain Monte-Carlo (MCMC) algorithms [25] like the Metropolis-Hastings algorithm [106, 78] that aim to sample a target distribution π by generating a Markov chain with stationary probability measure π. Fast convergence of the Markov chain towards π is one important property for the underlying algorithms. It can be described qualitatively as the geometric ergodicity of the Markov chain, i.e., convergence at a geometric rate towards π, a question that has been widely studied [47, 126]. We focus here on an application of stochastic processes to the domain of numerical stochastic optimization which is closely connected to MCMC. We analyze indeed a Markov chain underlying the so-called covariance matrix adaptation evolution strategy (CMA-ES) [76, 72], a widely used stochastic derivative-free optimization algorithm [129, 31, 22, 46, 2, 104, 137, 1][1] that can tackle difficult optimization problems which are notably nonconvex, multimodal and ill-conditioned. The algorithm minimizes a function $f : \mathbb{R}^n \to \mathbb{R}$ by sampling Gaussian vectors whose mean and covariance matrix are adapted iteratively. The adaptation of the parameters of the Gaussian distribution has been carefully designed, combining several independent principles [74, 76, 73, 120]. Ample empirical evidence shows that the algorithm converges geometrically fast [76, 73, 70, 68]—in optimization referred to as linear convergence—towards the optimum on large classes of functions and the covariance matrix learns the inverse Hessian [69] up to a scalar factor on strictly convex quadratic problems. Yet, establishing a convergence proof of CMA-ES that reflects its working

[1] As of March 2024, two Python implementations of CMA-ES received together more than 60 millions downloads.

principle (i.e., without modifying the algorithm to enforce convergence) is still an open and difficult theoretical question.

In this context, we extend a methodology that was already successful to analyze stepsize adaptive algorithms [14, 16, 23, 17, 141] to prove the convergence of CMA-ES by exploiting its mechanisms and reflecting its working principle, including the learning of second-order information. The methodology is based on the definition of a normalized Markov chain that models the algorithm when minimizing a scaling-invariant function, a function class that includes non quasi-convex functions [142]. As we will explain, if this Markov chain is stable—in the sense that it converges to a stationary distribution geometrically fast and satisfies a Law of Large Numbers—then the linear convergence of the algorithm follows. With more work, the learning of the inverse Hessian on strictly convex-quadratic functions should follow as well. In order to obtain such stability properties, the irreducibility of the process (the definitions will be recalled in the paper) is a necessary condition. On top of establishing the irreducibility of this Markov chain, we prove that it is an aperiodic T-chain, paving the way to a convergence analysis by means of a geometric drift condition.

Because of the intricacy of the algorithm, the irreducibility cannot be easily established by simply investigating the transition kernel of the Markov chain. Instead, we rely on recent results connecting the irreducibility of a Markov chain defined on a smooth manifold to the stability of an underlying control model. More precisely, we view the Markov chain as a nonlinear state-space model

$$\phi_{t+1} = F(\phi_t, \alpha(\phi_t, U_{t+1})) \tag{2.1}$$

where $\{U_{t+1}\}_{t\in\mathbb{N}}$ is an independent and identically distributed (i.i.d.) process valued in a measured space U, $F\colon \mathsf{X}\times\mathsf{V}\to\mathsf{X}$ is a locally Lipschitz update function between smooth manifolds X, V and $\alpha\colon \mathsf{X}\times\mathsf{U}\to\mathsf{V}$ is a measurable, possibly discontinuous function. When F is nonsmooth, we call (2.1) a nonsmooth state-space model. The connections that we rely on between the irreducibility, aperiodicity and T-chain property of the Markov chain and an underlying deterministic control model have been recently established [52], relaxing the assumptions in previous work [29] that the state space of the chain is an open subset of an Euclidean space and F is continuously differentiable. This latter work was already a generalization of the case where $\alpha(\phi_t, U_{t+1}) = U_{t+1}$ and F is smooth, i.e., infinitely differentiable [109].

While part of the methodology we follow relies on the results presented in [52], we introduce here two other generic and central techniques for the analysis.

Like in many practically used algorithms (in contrast to toy algorithms), different update mechanisms can be turned on and off in CMA-ES by some specific hyperparameter settings (like learning rates) resulting in different algorithm variants with varying number of state variables. Our aim is to analyze all algorithm variants without repeating the similar mathematical analysis for each of them. Hence, in order to have a single proof, we introduce the notions of *projected* and *redundant* Markov chains. Specifically, we consider a Markov chain $\{(\phi_t, \xi_t)\}_{t\in\mathbb{N}}$ valued in the manifold $\mathsf{X}\times\mathsf{Y}$ with

$$(\phi_{t+1}, \xi_{t+1}) = \tilde{F}((\phi_t, \xi_t), \tilde{\alpha}((\phi_t, \xi_t), U_{t+1})) \tag{2.2}$$

where $\{\phi_t\}_{t\in\mathbb{N}}$ obeys (2.1), and $\tilde{F}\colon \mathsf{X}\times\mathsf{Y}\times\mathsf{V}\to\mathsf{X}\times\mathsf{Y}$ and $\tilde{\alpha}\colon \mathsf{X}\times\mathsf{Y}\times\mathsf{U}\to\mathsf{X}\times\mathsf{Y}$ satisfy the same assumptions as F and α, respectively. We suppose then that

$$\Pi_{\mathsf{X}} \circ \tilde{F}\left((\phi, \xi), \tilde{\alpha}((\phi, \xi), u)\right) = F(\phi, \alpha(\phi, u)) \tag{2.3}$$

for every $(\phi, \xi, u) \in \mathsf{X}\times\mathsf{Y}\times\mathsf{U}$, where $\Pi_{\mathsf{X}}\colon \mathsf{X}\times\mathsf{Y}\to\mathsf{X}$ is the canonical projection of $\mathsf{X}\times\mathsf{Y}$ on X. The Markov chain $\{(\phi_t, \xi_t)\}_{t\in\mathbb{N}}$ is said to be redundant, whereas $\{\phi_t\}_{t\in\mathbb{N}}$ is said to be projected. We derive similar tools as in [52] to analyze the projected Markov chain $\{\phi_t\}_{t\in\mathbb{N}}$ by investigating the redundant control model (2.2).

Contributions Overall the contributions of this paper are twofold.

On the one hand, we provide two generic tools to analyze the irreducibility, aperiodicity and topological properties of complex nonsmooth state-space models. First, we extend the methodology to investigate Markov chains following (2.1) with locally Lipschitz updates on smooth manifolds in order to be able to deduce irreducibility, aperiodicity and T-chain property from a redundant chain to a projected chain. Second, we show how to transfer the analysis of nonsmooth state-space models following (2.1) from smooth manifolds to *nonsmooth* manifolds, as long as they can be continuously transformed into smooth manifolds.

On the other hand, using the developed tools, we establish the irreducibility, aperiodicity, and T-chain property of a Markov chain defined by the normalization of states of CMA-ES when minimizing a scaling-invariant function. Our results include most of the relevant hyperparameter settings, some of them described by separate Markov chains. The proven properties constitute an essential step for a proof of linear convergence of CMA-ES.

Organization In Section 2, we present the update equations behind CMA-ES and define a class of normalized Markov chains associated to the algorithm when minimizing scaling-invariant functions. In Section 3, we state our first main result that these Markov chains are irreducible, aperiodic T-chains. In Section 4, we state and prove results on the irreducibility, aperiodicity and topological properties of nonlinear state-space models. In Section 5, we apply the results exposed in Section 4 to the normalized Markov chain defined earlier ant prove the main result of Section 3. For the sake of readability, some proofs are delayed and presented in Section A and Section B.

Notations Throughout this paper, we use the following notations: $\mathbb{N}$, $\mathbb{N}^*$, $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}_{++}$ for the sets of nonnegative integers, positive integers, real numbers, nonnegative real numbers, and positive real numbers, respectively. Unless stated otherwise, for $n \in \mathbb{N}^*$ and any vector $x \in \mathbb{R}^n$, $\|x\|$ denotes the Euclidean norm of x. The set of real symmetric matrices of size $d \times d$ is denoted $\mathcal{S}^d$, and its subsets of positive semi-definite matrices and of positive definite matrices are denoted $\mathcal{S}^d_+$ and $\mathcal{S}^d_{++}$, respectively. Given a positive integer n, $\mathfrak{S}_n$ represents the set of permutations of $\{1, \dots, n\}$, and its cardinality is denoted $n!$. The differential application of a function F at a point x is denoted $\mathcal{D}F(x)$, and the Clarke derivative of F at x is denoted $\partial F(x)$. We use the notations $\operatorname{Arg\,min} f$ and $\operatorname{Arg\,max} f$ for the sets of global minima and global maxima of f, respectively. When unique global minimum and maximum exist, we denote them as $\arg\min f$ and $\arg\max f$, respectively. For any sequence $\{v_k\}_{k\in\mathbb{N}^*}$ and any $k \in \mathbb{N}^*$, we set $v_{1:k} = (v_1, \dots, v_k)$. For a topological space X, we denote $\mathcal{B}(\mathsf{X})$ the Borel σ-field of X, which makes X a measured space. If μ is a measure on $\mathcal{B}(\mathsf{X})$ and ν is a measure on $\mathcal{B}(\mathsf{Y})$, we denote $\mu \otimes \nu$ the product measure of μ and ν, which is a measure on $\mathcal{B}(\mathsf{X} \times \mathsf{Y})$. Likewise, for $k \in \mathbb{N}^*$, we denote $\mu^{\otimes k}$ the measure product of μ by itself k times, as a measure on $\mathcal{B}(\mathsf{X}^k)$.

2 Definition of Markov chains arising from a normalization of CMA-ES

We present in this section the CMA-ES algorithm and define normalized Markov chains—candidates to be stable—associated to the algorithm. We explain the connection between the stability of these Markov chains and the convergence of the algorithm, motivating thus why the irreducibility, aperiodicity and topological properties of the Markov chains that we study in the paper are an important part for obtaining a convergence proof of CMA-ES.

2.1 Presentation of CMA-ES

The covariance matrix adaptation evolution strategy (CMA-ES) is an iterative algorithm which aims to approximate a problem solution

$$x^* \in \operatorname*{Arg\,min}_{x \in \mathbb{R}^d} f(x) \tag{P}$$

where $d \in \mathbb{N}^*$ is the dimension of the problem, and $f \colon \mathbb{R}^d \to \mathbb{R}$ is the objective function. A vector $x^* \in \mathbb{R}^d$, solution to (P), is called a global minimum of f. The CMA-ES attempts to approach x^* by successively sampling, for iterations $t \in \mathbb{N}$, new candidate solutions from a multivariate normal probability distribution $\mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$. The vector $m_t \in \mathbb{R}^d$ is the current mean of the distribution and we specifically desire that $f(m_t)$ converges to the essential infimum of f. The positive real number $\sigma_t > 0$ is the current stepsize, and the symmetric positive definite matrix $\mathbf{C}_t \in \mathcal{S}_{++}^d$ is referred to as the current covariance matrix. For our analysis, we generalize the assumption that the distribution of the candidate solutions is multivariate normal.

The parameters of the sampling distribution are updated using two cumulation paths $p_t^\sigma, p_t^c \in \mathbb{R}^d$, which implement a weighted moving average of the steps followed by the mean.

More precisely, the algorithm works as follows. At iteration $t \in \mathbb{N}$, given $m_t \in \mathbb{R}^d$, $\sigma_t > 0$, $\mathbf{C}_t \in \mathcal{S}_{++}^d$, and $p_t^\sigma, p_t^c \in \mathbb{R}^d$, we generate independent identically distributed (i.i.d.) samples $U_{t+1}^1, \dots, U_{t+1}^\lambda$ following a sampling distribution ν_U^d in $\mathbb{R}^d$ and independently of $(m_t, \sigma_t, \mathbf{C}_t, p_t^\sigma, p_t^c)$. Usually, the distribution ν_U^d is the standard normal distribution in $\mathbb{R}^d$. However, throughout the paper we will refer to CMA-ES as the algorithm presented in this section with a general and abstract sampling distribution ν_U^d. We compute then λ candidate solutions

$$x_{t+1}^i \coloneqq m_t + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^i \text{ for } i = 1, \dots, \lambda \ , \tag{2.4}$$

and rank them with respect to their f-values. Formally, we define a permutation $s_{t+1} \in \mathfrak{S}_\lambda$ satisfying

$$f\left(x_{t+1}^{s_{t+1}(1)}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{s_{t+1}(\lambda)}\right) \ . \tag{2.5}$$

When $f\left(x_{t+1}^i\right) = f\left(x_{t+1}^j\right)$, we impose for uniqueness $s_{t+1}^{-1}(i) < s_{t+1}^{-1}(j)$ if $i < j$. We say that we have *neutral selection* when, instead of (2.5), the permutation s_{t+1} is independent of the samples U_{t+1}^i for all $t \in \mathbb{N}$. This is the case, for example, when the permutation is fixed for all t, or when $f(x_{t+1}^i)$ is independent of its argument or independent of U_{t+1}^i.

The mean is moved towards the best solutions, and is updated by applying the function $F_{c_m}^{\mathrm{m}}$ defined as

$$F_{c_m}^{\mathrm{m}} \colon (m, v) \in \mathbb{R}^d \times \mathbb{R}^d \ \mapsto m + c_m v \ , \tag{2.6}$$

given a fixed learning rate $c_m > 0$ (by default $c_m = 1$). Precisely, the mean obeys

$$m_{t+1} = F_{c_m}^{\mathrm{m}}\left(m_t, \sigma_t \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}\right) = m_t + c_m \sigma_t \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \ , \tag{2.7}$$

where $w_1^m \geqslant \cdots \geqslant w_\mu^m > 0$ are weights such that $\sum_{i=1}^\mu w_i^m = 1$, and $\sqrt{\mathbf{C}_t}$ is the symmetric positive definite square root of $\mathbf{C}_t$.

We introduce the function to update the paths $p_t^\sigma, p_t^c \in \mathbb{R}^d$. Given a decay factor $c \in (0, 1]$, F_c^{p} is defined as

$$F_c^{\mathrm{p}} \colon (p, v) \in \mathbb{R}^d \times \mathbb{R}^d \ \mapsto (1 - c)p + \sqrt{c(2 - c)\mu_{\mathrm{eff}}}\, v \tag{2.8}$$

where $\mu_{\text{eff}} = 1/\|\mathbf{w}_m\|^2$, with $\mathbf{w}_m = (w_1^m, \dots, w_\mu^m)^\top$. The closer the decay factor c is to zero, the more the updated path depends on the previous path due to the term $(1-c)p$. Conversely, when $c = 1$, the updated path is collinear to and only depends on v. We set two decay factors, $c_\sigma, c_c \in (0, 1]$, and use (2.8) to update two cumulation paths, one for updating the stepsize and the other for the rank-one update of the covariance matrix (see below). We update

$$p_{t+1}^\sigma = (1-c_\sigma)p_t^\sigma + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} = F_{c_\sigma}^{\text{p}}\left(p_t^\sigma, \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}\right) \quad , \tag{2.9}$$

and

$$p_{t+1}^c = (1-c_c)p_t^c + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} = F_{c_c}^{\text{p}}\left(p_t^c, \sqrt{\mathbf{C}_t}\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}\right) \quad . \tag{2.10}$$

The second argument in the RHS of (2.10) is the same as in (2.7) disregarding stepsize σ_t. Eq. (2.9) additionally drops $\sqrt{\mathbf{C}_t}$. Consequently, when p_0^σ and U_{t+1}^i are standard Gaussian, then, under neutral selection, p_{t+1}^σ is a standard Gaussian vector too and, in particular, the length of p_{t+1}^σ does not depend on its direction. The path p_{t+1}^c from (2.10) maintains under neutral selection the covariance matrix $\mathbf{C}_t$ when p_t^c has covariance matrix $\mathbf{C}_t$. The path is commensurable with updating $\mathbf{C}_t$ and its expected length can strongly depend on its direction.

The stepsize is updated using the path p_{t+1}^σ. Considering an abstract measurable function $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$ that we call stepsize change, the update reads

$$\sigma_{t+1} = \sigma_t \times \Gamma\left(p_{t+1}^\sigma\right) \quad . \tag{2.11}$$

A standard stepsize change used in CMA-ES is the cumulative stepsize adaptation (CSA) where Γ equals

$$\Gamma_{\text{CSA}}^1(p) = \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|p\|}{\mathbb{E}\|\nu_U^d\|} - 1\right)\right) \quad , \tag{2.12}$$

where $\mathbb{E}\|\nu_U^d\| := \mathbb{E}\|\xi\|$ for a random variable ξ distributed under ν_U^d. When ν_U^d is the standard normal distribution, (2.12) increases the stepsize when $\|p_t^\sigma\|$ is larger than to be expected under neutral selection (assuming that $p_0^\sigma \sim \nu_U^d$) and decreases the stepsize when $\|p_t^\sigma\|$ is smaller. When consecutive steps are taken in a similar direction, the expected path is long while the same progress could be made in fewer iterations with larger steps. When consecutive steps are negatively correlated, the expected path is short and a smaller stepsize is advisable. A smooth alternative to (2.12) implementing the same idea is [13]

$$\Gamma_{\text{CSA}}^2(p) = \exp\left(\frac{c_\sigma}{2d_\sigma}\left(\frac{\|p\|^2}{\mathbb{E}\|\nu_U^d\|^2} - 1\right)\right) \quad . \tag{2.13}$$

These two stepsize changes rely on the choice of damping parameter $d_\sigma > 0$, which is chosen $\approx 1 + 2\sqrt{\mu_{\text{eff}}/d}$ in the first case and $\approx 1 + 2\mu_{\text{eff}}/d$ in the second case. Empirically, Γ_{CSA}^1 and Γ_{CSA}^2 show similar performance when used with CMA-ES [49]. While the function Γ_{CSA}^1 is the default stepsize change, previous theoretical works on ES also have investigated Γ_{CSA}^2 [13, 141].

Last, we introduce the update function for the covariance matrix, which depends on the choice of learning rates $c_1, c_\mu \geqslant 0$ such that $c_1 + c_\mu \in [0, 1]$:

$$\begin{aligned} F_{c_1,c_\mu}^{\text{C}}\colon \mathcal{S}_{++}^d \times \mathbb{R}^d \times \mathcal{S}_+^d &\to \mathcal{S}_{++}^d \\ (\mathbf{C}, p, \mathbf{M}) &\mapsto (1-c_1-c_\mu)\mathbf{C} + c_1 pp^\top + c_\mu \mathbf{M} \quad , \end{aligned} \tag{2.14}$$

and the covariance matrix is updated via

$$\mathbf{C}_{t+1} = (1 - c_1 - c_\mu)\mathbf{C}_t + c_1 p_{t+1}^c [p_{t+1}^c]^\top + c_\mu \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right] \left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\mathbf{C}_t}$$
$$= F_{c_1,c_\mu}^{\mathrm{C}} \left(\mathbf{C}_t, p_{t+1}^c, \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right] \left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\mathbf{C}_t} \right) \quad , \quad (2.15)$$

where we define weights $w_1^c \geqslant \cdots \geqslant w_\mu^c > 0$ such that $\sum_{i=1}^{\mu} w_i^c = 1$. Moreover, we assume throughout the paper that $0 < c_1 + c_\mu < 1$. This assumption will be essential in the proofs of Lemma 2.8, Corollary 2.3 and Proposition 2.12. The setting $c_1 + c_\mu = 1$ is however used in practice when μ is large and we believe that with further work our results could be proven for this case as well.

The term $c_1 pp^\top$ is called the rank-one update, whereas the term $c_\mu \mathbf{M}$ is the rank-mu update since in (2.15) we replace $\mathbf{M}$ by a matrix of rank $\min(\mu, d)$ almost surely which satisfies a maximum likelihood condition for the best samples of the last iteration [70, Proposition 7]. In practice, also negative weights are used for the rank-mu update of the covariance matrix [91, 77]. However, since the updated covariance matrix must be positive definite, the norm of the vectors corresponding to negative weights must be controlled. We do not consider negative weights in the present paper.

2.2 *Assumptions*

Our analysis of CMA-ES relies on analyzing the stability of normalized Markov chains underlying the algorithm. The construction of these Markov chains assumes that the objective function is scaling-invariant. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *scaling-invariant with respect to* $x^* \in \mathbb{R}^d$ when for all $x, y \in \mathbb{R}^d$ and $\rho > 0$:

$$f(x + x^*) \leqslant f(y + x^*) \Leftrightarrow f(\rho x + x^*) \leqslant f(\rho y + x^*) \quad . \quad (2.16)$$

The class of scaling-invariant functions has been of interest for the convergence analysis of different variants of ES [17, 141], and is related to the class of positively homogeneous functions [142]. We make an additional technical assumption on the level sets of the objective function to avoid ties in (2.5), which will be useful to define lower semi-continuous density functions in Lemma 2.4. Overall, we will use the following assumptions:

F1. *The objective function* $f : \mathbb{R}^d \to \mathbb{R}$ *is a strictly increasing transformation of a continuous function with Lebesgue negligible level sets.*

F2. *The objective function* $f : \mathbb{R}^d \to \mathbb{R}$ *is scaling-invariant with respect to a point* $x^* \in \mathbb{R}^d$.

Instead of assuming **F**1, we can suppose without loss of generality that the function f is continuous with Lebesgue negligible level sets since CMA-ES is invariant to increasing transformations of the objective function [15]. Assumption **F**2 is central in this analysis since it is required to define an equivalent, normalized Markov chain via (2.17) below.

In order to go beyond scaling-invariant functions, it might be possible to adopt another approach, considering for instance recent works on the convergence of evolution strategies that prove a drift condition on the state variables of the algorithm and hence the convergence on composites of strongly convex functions with strictly increasing functions (for however so far the (1+1)-ES selection scheme only) [4, 113].

Furthermore, the sampling distribution ν_U^d should satisfy the following assumption that allows in particular to characterize a density for the ranked candidate solutions, see Lemma 2.4.

N1. *The probability distribution ν_U^d admits a continuous density $p_U^d(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}^d$ which is positive everywhere on $\mathbb{R}^d$.*

This assumption is satisfied by a multivariate standard normal distribution as used in CMA-ES. We have moreover the following assumptions on the stepsize change Γ.

Γ1. *The stepsize change $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$ is a locally Lipschitz map and is differentiable at every nonzero vector of $\mathbb{R}^d$.*

Γ2. *Given c_c the cumulation parameter for the path in (2.9), the function Γ satisfies $\liminf \Gamma(p) > (1-c_c)^{-1}$ for $\|p\|$ to $+\infty$.*

Γ3. *The function Γ satisfies $\Gamma(0) < 1$.*

Assumption **Γ**1 is required to apply the results stated in Section 4.1 and in particular to ensure that the analyzed process satisfies the condition **H**2 in Section 4.1 to obtain the irreducibility and aperiodicity of the Markov chain defined in (2.17). Assumptions **Γ**2 and **Γ**3 are used in Propositions 2.10 and 2.12, respectively.

Assumptions **Γ**1–**Γ**3 are satisfied by both stepsize changes, Γ^1_{CSA} and Γ^2_{CSA}, as stated in the following lemma.

Lemma 2.1. Assume that $c_\sigma \in (0,1]$. Then, the stepsize change functions Γ^1_{CSA} and Γ^2_{CSA}, defined by (2.12) and (2.13) respectively, satisfy the assumptions **Γ**1–**Γ**3.

Proof. The proof is simple and left to the reader. □

2.3 *Proving the stability of a normalized Markov chain leads to linear convergence*

Before stating our main results, we define a normalized Markov chain underlying the CMA-ES algorithm. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An event is an element $\mathsf{W} \in \mathcal{F}$, and the probability of W is $\mathbb{P}[\mathsf{W}]$. A random variable U valued in a measured space $(\mathsf{U}, \mathcal{U})$ is defined as a measurable function $U\colon \Omega \to U$, and for $\mathsf{A} \in \mathcal{U}$, we identify $\mathbb{P}[U \in \mathsf{A}]$ to $\mathbb{P}[\{\omega \in \Omega \mid U(\omega) \in \mathsf{A}\}]$. A transition kernel on a topological state space X equipped with its Borelian σ-field $\mathcal{B}(\mathsf{X})$ is an application $P\colon \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}_+$ such that, for every $x \in \mathsf{X}$, $\mathsf{A} \in \mathcal{B}(\mathsf{X}) \mapsto P(x, \mathsf{A})$ is a probability measure, and for every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$, $x \in \mathsf{X} \mapsto P(x, \mathsf{A})$ is a measurable map. Then, a (time-homogeneous) Markov chain with transition kernel P on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ and initial probability distribution ν on $\mathcal{B}(\mathsf{X})$ is a sequence of random variables $\Phi = \{\phi_t\}_{t\in\mathbb{N}}$ valued in X, satisfying for every $t \in \mathbb{N}$

$$\begin{aligned}&\mathbb{P}\left[(\phi_0, \ldots, \phi_t) \in \mathsf{A}_0 \times \cdots \times \mathsf{A}_t \mid \phi_0 \sim \nu\right] \\ &= \int_{\mathsf{X}^{t+1}} \mathbb{1}\left\{(x_0, \ldots, x_{t-1}) \in \mathsf{A}_0 \times \cdots \times \mathsf{A}_{t-1}\right\} P(x_{t-1}, \mathsf{A}_t) P(x_{t-2}, \mathrm{d}x_{t-1}) \ldots P(x_0, \mathrm{d}x_1)\nu(\mathrm{d}x_0)\end{aligned}$$

where for every probability measure ν on $\mathcal{B}(\mathsf{X})$, we have equipped $(\Omega, \mathcal{F})$ with a probability measure $\mathbb{P}[\cdot|\phi_0 \sim \nu]$. We define the t-step transition kernel by $P^t(x, \mathsf{A}) = \mathbb{P}[\phi_t \in \mathsf{A}|\phi_0 \sim \delta_x]$ for every $t \geqslant 0$ and $\mathsf{A} \in \mathcal{B}(\mathsf{X})$.

The sequence $\{(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$ introduced in Section 2.1 defines a time-homogeneous Markov chain on the state space $\mathbb{R}^{3d} \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d$. This is immediate from the observation that the definition of $(m_{t+1}, p_{t+1}^\sigma, p_{t+1}^c, \sigma_{t+1}, \mathbf{C}_{t+1})$ depends only on the previous state $(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)$ and the independent random input $U_{t+1}^1, \ldots, U_{t+1}^\lambda$. However, when the mean converges to the optimum of

the function, the stepsize σ_t, the covariance matrix $\mathbf{C}_t$ and the path p_t^c converge to 0. Therefore, this Markov chain is not Harris recurrent (it does not revisit every neighborhood of any state infinitely many times). Yet, as illustrated later in (2.21) and Proposition 2.3, our methodology to prove linear convergence [17] relies on a Law of Large Numbers which motivates to have a positive Harris recurrent Markov chain (with a stationary probability distribution) and, more generally, a chain stable enough to apply an ergodic theorem [110, Theorems 13.0.1] and satisfy a Law of Large Numbers [110, Theorem 17.0.1]. Therefore, we define a normalized process, candidate to have a stationary distribution, underlying the CMA-ES algorithm. Consider $R\colon \mathcal{S}_{++}^d \to \mathbb{R}_{++}$ a normalization function which is

R1. *(positively) homogeneous with degree 1, i.e., for every* $\mathbf{A} \in \mathcal{S}_{++}^d$ *and* $\rho > 0$, $R(\rho\mathbf{A}) = \rho R(\mathbf{A})$,

R2. *locally Lipschitz continuous,*

R3. *differentiable on a nonempty open subset of* $\mathcal{S}_{++}^d$.

Assumption **R**1 is required to define a normalized Markov chain, see (2.17) below, as proven in Proposition 2.2. Assumption **R**2 is used to prove irreducibility and aperiodicity of the normalized chain, notably for the verification of condition **H**2 introduced in Section 4.1. Later, Proposition 2.12 uses **R**3 to prove a maximal rank condition.

We give examples of normalization functions that satisfy these assumptions.

Proposition 2.1. The d-th root of the determinant, $\det(\cdot)^{1/d}$, and the i-th largest eigenvalue, $\lambda_i(\cdot)$, for $i \in \{1, \ldots, d\}$ counted with multiplicity, are functions defined on $\mathcal{S}_{++}^d$ that satisfy **R**1–**R**3.

Proof. The proof of the property **R**1 is immediate from the linearity of the determinant as a function of the columns of the matrix and the definition of an eigenvalue. For the properties **R**2 and **R**3, we know that the determinant of a matrix is a polynomial function of the coefficients of the matrix [79, Section 0.3], hence it is infinitely differentiable and in particular is locally Lipschitz [30, Proposition and Corollary 2.2.1]. For the eigenvalues, $\lambda_i(\cdot)$ is locally Lipschitz on $\mathcal{S}_{++}^d$, as a consequence of Weyl's theorem [79, Corollary 4.3.15]. Besides, $\lambda_i(\cdot)$ is infinitely differentiable on a neighborhood of any symmetric matrix with eigenvalues that have simple multiplicity [133, Theorem 5.3]. □

Given the CMA-ES Markov chain $\{(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$ defined in Section 2.1 and a normalization function R, we define the *normalized chain* $\Phi = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ as follows.[2] For all $t \geqslant 1$, we set

$$z_t = \frac{m_t - x^*}{\sigma_t \sqrt{R(\mathbf{C}_t)}},\ p_t = p_t^\sigma,\ q_t = \frac{p_t^c}{\sqrt{R(\mathbf{C}_{t-1})}},\ \mathbf{\Sigma}_t = \frac{\mathbf{C}_t}{R(\mathbf{C}_t)},\ r_t = \frac{R(\mathbf{C}_t)}{R(\mathbf{C}_{t-1})}\ . \tag{2.17}$$

We prove below that when the objective function is scaling-invariant, the normalized chain defined by (2.17) is a time-homogeneous Markov chain that can be defined independently of the original Markov chain $\{(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$. We establish first that on scaling-invariant functions, the permutation sorting the candidate solutions $m_t + \sigma_t\sqrt{\mathbf{C}_t}U_{t+1}^i$ also sorts the vectors $z_t + \sqrt{\mathbf{\Sigma}_t}U_{t+1}^i$, for $i = 1, \ldots, \lambda$.

[2] The definition of q_t in (2.17) suggests transforming p_t^c in (2.10) like $\mathbf{C}_t^{1/2}\mathbf{C}_{t-1}^{-1/2}p_t^c$ to avoid the time index $t-1$ in (2.17). Then, p_{t+1}^c would become equal to $\mathbf{C}_t^{1/2}p_{t+1}^\sigma$. We can prove unbiasedness for p^σ [70] and affine invariance with p^c and $c_\sigma = 1$ [15].

Lemma 2.2. Let $t \geqslant 1$ and suppose that the objective function f satisfies **F**2. Let $s_{t+1} \in \mathfrak{S}_\lambda$ be a (random) permutation that sorts the indices $i = 1, \dots, \lambda$ with respect to the f-values of $m_t + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^i$. Then, s_{t+1} also sorts the indices $i = 1, \dots, \lambda$ with respect to the f-values of $x^* + z_t + \sqrt{\mathbf{\Sigma}_t} U_{t+1}^i$. Moreover, we can ensure the uniqueness of s_{t+1} by imposing a tie-break (cf. Section 2.1).

Proof. Let $i = 1, \dots, \lambda$. By definition of z_t and $\mathbf{\Sigma}_t$, we obtain

$$f\left(x^* + z_t + \sqrt{\mathbf{\Sigma}_t} U_{t+1}^i\right) = f\left(x^* + R(\mathbf{C}_t)^{-1/2} \sigma_t^{-1} \times \left[m_t - x^* + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^i\right]\right).$$

We conclude by using the definition of a scaling-invariant function (2.16). □

From the previous lemma, we deduce in Proposition 2.2 below that the normalized chain defined in (2.17) is a time-homogeneous Markov chain that can be defined independently of the original Markov chain. Indeed, given R satisfying **R**1, denote with a slight abuse of notation (since we use the same notation as for (2.17) with however a different time index) the time-homogeneous Markov chain $\Phi = \{\phi_t\}_{t \geqslant 0} = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t \geqslant 0}$ defined via $\phi_0 \in \mathsf{Y} = (\mathbb{R}^d)^3 \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ (where $R^{-1}(\{1\}) = \{\mathbf{\Sigma} \in \mathcal{S}_{++}^d : R(\mathbf{\Sigma}) = 1\}$) and the following recursion

$$\begin{aligned}
z_{t+1} &= \frac{z_t + c_m \sqrt{\mathbf{\Sigma}_t} \mathbf{w}_m^\top U_{t+1}^{s_{t+1}}}{\sqrt{r_{t+1}}\, \Gamma(p_{t+1})} = \frac{F_{c_m}^{\mathrm{m}}(z_t, \sqrt{\mathbf{\Sigma}_t} \mathbf{w}_m^\top U_{t+1}^{s_{t+1}})}{\sqrt{r_{t+1}}\, \Gamma(p_{t+1})} \\
p_{t+1} &= (1 - c_\sigma) p_t + \sqrt{c_\sigma (2 - c_\sigma) \mu_{\mathrm{eff}}}\, \mathbf{w}_m^\top U_{t+1}^{s_{t+1}} = F_{c_\sigma}^{\mathrm{p}}(p_t, \mathbf{w}_m^\top U_{t+1}^{s_{t+1}}) \\
q_{t+1} &= F_{c_c}^{\mathrm{p}}(r_t^{-1/2} q_t, \sqrt{\mathbf{\Sigma}_t} \mathbf{w}_m^\top U_{t+1}^{s_{t+1}}) \\
\mathbf{\Sigma}_{t+1} &= r_{t+1}^{-1} F_{c_1, c_\mu}^{\mathrm{C}}\left(\mathbf{\Sigma}_t, q_{t+1}, \sqrt{\mathbf{\Sigma}_t} \sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right] \left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\mathbf{\Sigma}_t}\right) \\
r_{t+1} &= R \circ F_{c_1, c_\mu}^{\mathrm{C}}\left(\mathbf{\Sigma}_t, q_{t+1}, \sqrt{\mathbf{\Sigma}_t} \sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right] \left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\mathbf{\Sigma}_t}\right)
\end{aligned} \tag{2.18}$$

with $\mathbf{U} = \{U_{t+1}\}_{t \in \mathbb{N}}$ an i.i.d. process independent of ϕ_0 with $U_1 = (U_1^1, \dots, U_1^\lambda) \sim (\nu_U^d)^{\otimes \lambda}$, and s_{t+1} the (almost surely unique) permutation that sorts the $f(z_t + \sqrt{\mathbf{\Sigma}_t} U_{t+1}^i)$, $i = 1, \dots, \lambda$. Moreover, $U_{t+1}^{s_{t+1}}$ denotes the collection of vectors $(U_{t+1}^{s_{t+1}(1)}, \dots, U_{t+1}^{s_{t+1}(\lambda)})$. Remark that in (2.18), the update of the covariance matrix $\mathbf{\Sigma}_{t+1}$ writes

$$\mathbf{\Sigma}_{t+1} = \frac{\tilde{\mathbf{\Sigma}}_{t+1}}{R(\tilde{\mathbf{\Sigma}}_{t+1})}$$

where $\tilde{\mathbf{\Sigma}}_{t+1}$ is the covariance matrix to which we apply the rank-one and rank-mu updates, i.e.,

$$\begin{aligned}
\tilde{\mathbf{\Sigma}}_{t+1} &:= F_{c_1, c_\mu}^{\mathrm{C}}\left(\mathbf{\Sigma}_t, q_{t+1}, \sqrt{\mathbf{\Sigma}_t} \sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right] \left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\mathbf{\Sigma}_t}\right) \\
&= (1 - c_1 - c_\mu) \mathbf{\Sigma}_t + c_1 q_{t+1} (q_{t+1})^\top + c_\mu \sqrt{\mathbf{\Sigma}_t} \sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right] \left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\mathbf{\Sigma}_t}\ ,
\end{aligned} \tag{2.19}$$

where $F_{c_1, c_\mu}^{\mathrm{C}}$ is defined via (2.14). Similarly r_{t+1} can be expressed using $\tilde{\mathbf{\Sigma}}_{t+1}$ as

$$r_{t+1} = R(\tilde{\mathbf{\Sigma}}_{t+1})\ .$$

The update of ϕ_t in (2.18) defines a function F_Φ such that

$$\phi_{t+1} = F_\Phi(\phi_t, U_{t+1}^{s_{t+1}}) \ . \tag{2.20}$$

We prove in the next proposition that if f is scaling-invariant, the normalized chain defined in (2.17) can be defined independently of the original CMA-ES chain via the recursion (2.20) provided it is initialized properly. While the normalized process (2.17) imposes $t \geqslant 1$, the next proposition defines this process via the recursion (2.18) and thus allows to start with any time index.

Proposition 2.2. Suppose that the objective function f satisfies **F**2 and that the normalization function R satisfies **R**1. Let $\{(m_t, p_t^\sigma, p_t^c, \mathbf{C}_t, \sigma_t)\}_{t\in\mathbb{N}}$ be the chain associated to CMA-ES defined in Section 2.1 and $\Phi = \{\phi_t\}_{t\geqslant 1} = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ be the normalized process defined via (2.17) for $t \geqslant 1$. Then Φ is a time-homogeneous Markov chain valued in the state space $\mathsf{Y} = (\mathbb{R}^d)^3 \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ that satisfies

$$\phi_1 = \left(\frac{m_1}{\sigma_1 \sqrt{R(\mathbf{C}_1)}}, p_1^\sigma, \frac{p_1^c}{\sqrt{R(\mathbf{C}_0)}}, \frac{\mathbf{C}_1}{R(\mathbf{C}_1)}, \frac{R(\mathbf{C}_1)}{R(\mathbf{C}_0)} \right)$$

and for $t \geqslant 1$ we have

$$\phi_{t+1} = F_\Phi(\phi_t, U_{t+1}^{s_{t+1}})$$

where F_Φ is the function in (2.20) defined via the equations (2.18), $s_{t+1} \in \mathfrak{S}_\lambda$ is a permutation that sorts[a] the $f(x^* + z_t + \sqrt{\mathbf{\Sigma}_t} U_{t+1}^i)$, $i = 1, \ldots, \lambda$, and $\mathbf{U} = \{U_{t+1}\}_{t\geqslant 1}$ is the i.i.d. process used to define $\{(m_t, p_t^\sigma, p_t^c, \mathbf{C}_t, \sigma_t)\}_{t\in\mathbb{N}}$, thus independent of ϕ_1.

[a] We always sort increasing and, as explained in Section 2.1, in case of a tie between the f-values of the candidate solutions of indices i and j with $i < j$, we impose $s_{t+1}^{-1}(i) < s_{t+1}^{-1}(j)$ to ensure the uniqueness of the permutation s_{t+1}.

Proof. By Lemma 2.2, it is sufficient to show that (2.18) holds for every $t \geqslant 1$ in order to prove that Φ is a time-homogeneous Markov chain. Let $t \geqslant 1$, and consider the matrix $\tilde{\mathbf{\Sigma}}_{t+1}$ defined in (2.19). Since $F_{c_1,c_\mu}^{\mathrm{C}}$ is homogeneous with respect to its first variable, positively homogeneous of degree 2 with respect to the second variable, using (2.17) and the definition of $\mathbf{C}_{t+1}$ in (2.15) we find

$$\tilde{\mathbf{\Sigma}}_{t+1} = R(\mathbf{C}_t)^{-1} \mathbf{C}_{t+1} \ .$$

By the property **R**1 applied to the previous equation we obtain $R(\tilde{\mathbf{\Sigma}}_{t+1}) = R(\mathbf{C}_t)^{-1} R(\mathbf{C}_{t+1}) = r_{t+1}$. Furthermore, the following holds

$$\begin{aligned}
z_{t+1} =& R(\mathbf{C}_{t+1})^{-1/2} \sigma_{t+1}^{-1} \times (m_{t+1} - x^*) \\
=& r_{t+1}^{-1/2} R(\mathbf{C}_t)^{-1/2} \sigma_t^{-1} \Gamma(p_{t+1}^\sigma)^{-1} \times \left[m_t - x^* + c_m \sigma_t \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \right] \\
=& r_{t+1}^{-1/2} \Gamma(p_{t+1})^{-1} \times \left[z_t + c_m \sqrt{\mathbf{\Sigma}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \right] = \frac{F_{c_m}^{\mathrm{m}}(z_t, \sqrt{\mathbf{\Sigma}_t} \mathbf{w}_m^\top U_{t+1}^{s_{t+1}})}{\sqrt{r_{t+1}} \Gamma(p_{t+1})} \ ,
\end{aligned}$$

where $F^{\mathrm{m}}_{c_m}$ is defined via (2.6). Moreover,

$$\begin{aligned}
\boldsymbol{\Sigma}_{t+1} &= R(\mathbf{C}_{t+1})^{-1}\mathbf{C}_{t+1} \\
&= \frac{(1-c_1-c_\mu)\mathbf{C}_t + c_1(p^c_{t+1})(p^c_{t+1})^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \left(\sqrt{\mathbf{C}_t}U^{s_{t+1}(i)}_{t+1}\right)\left(\sqrt{\mathbf{C}_t}U^{s_{t+1}(i)}_{t+1}\right)^\top}{R(\mathbf{C}_{t+1})} \\
&= r_{t+1}^{-1} \times \left[(1-c_1-c_\mu)\boldsymbol{\Sigma}_t + c_1(q_{t+1})(q_{t+1})^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \left(\sqrt{\boldsymbol{\Sigma}_t}U^{s_{t+1}(i)}_{t+1}\right)\left(\sqrt{\boldsymbol{\Sigma}_t}U^{s_{t+1}(i)}_{t+1}\right)^\top\right] \\
&= r_{t+1}^{-1} F^{\mathrm{C}}_{c_1,c_\mu}\left(\boldsymbol{\Sigma}_t, q_{t+1}, \sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu} w_i^c \left[U^{s_{t+1}(i)}_{t+1}\right]\left[U^{s_{t+1}(i)}_{t+1}\right]^\top \sqrt{\boldsymbol{\Sigma}_t}\right)
\end{aligned}$$

Finally,

$$\begin{aligned}
q_{t+1} &= R(\mathbf{C}_t)^{-1/2} p^c_{t+1} \\
&= R(\mathbf{C}_t)^{-1/2}(1-c_c)p^c_t + \sqrt{\mu_{\mathrm{eff}} c_c(2-c_c)}R(\mathbf{C}_t)^{-1/2}\mathbf{C}_t^{1/2}\sum_{i=1}^{\mu} w_i^m U^{s_{t+1}(i)}_{t+1} \\
&= r_t^{-1/2}(1-c_c)q_t + \sqrt{\mu_{\mathrm{eff}} c_c(2-c_c)}\boldsymbol{\Sigma}_t^{1/2}\sum_{i=1}^{\mu} w_i^m U^{s_{t+1}(i)}_{t+1} = F^{\mathrm{p}}_{c_c}(r_t^{-1/2}q_t, \sqrt{\boldsymbol{\Sigma}_t}\mathbf{w}_m^\top U^{s_{t+1}}_{t+1}) \ ,
\end{aligned}$$

where $F^{\mathrm{p}}_{c_c}$ is defined via (2.8). □

Now that we formally prove that the normalized chain defined in (2.17) is a time-homogeneous Markov chain when the algorithm optimizes a scaling-invariant function, we recapitulate how its stability is connected to the linear convergence of CMA-ES on scaling-invariant functions. For $T \in \mathbb{N}$, using the definition of the normalized Markov chain in (2.17) and the definition of the stepsize change (2.11) we obtain

$$\begin{aligned}
\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} &= \frac{1}{T}\sum_{t=0}^{T-1}[\log\|m_{t+1}-x^*\| - \log\|m_t - x^*\|] \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\left(\log\left(\|z_{t+1}\|\sqrt{R(\mathbf{C}_{t+1})}\sigma_{t+1}\right) - \log\left(\|z_t\|\sqrt{R(\mathbf{C}_t)}\sigma_t\right)\right) \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\left(\log\|z_{t+1}\| - \log\|z_t\| + \log\frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2}\log\frac{R(\mathbf{C}_{t+1})}{R(\mathbf{C}_t)}\right) \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\left(\log\|z_{t+1}\| - \log\|z_t\| + \log\Gamma(p_{t+1}) + \frac{1}{2}\log r_{t+1}\right) \ . \qquad (2.21)
\end{aligned}$$

If the Law of Large Numbers applies to the RHS of (2.21), we obtain a limit of the LHS when T goes to infinity. If this limit is proven to be strictly negative, we have shown linear convergence of the underlying optimization algorithm. In order to apply limit theorems [110, Theorem 17.0.1] and obtain a Law of Large Numbers, we require the chain Φ to be geometrically ergodic. Key assumptions for ergodicity are irreducibility and aperiodicity of the Markov chain whose notions will be formally introduced in Section 3. We thus connected the stability of the normalized chain to the convergence of the underlying optimization algorithm. More formally the following proposition holds.

Proposition 2.3. Consider the CMA-ES algorithm defined in Section 2.1 optimizing a function f satisfying **F**2. Assume that the process Φ, obeying (2.18) with state space $\mathsf{Y} = (\mathbb{R}^d)^3 \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ (where $R^{-1}(\{1\}) = \{\mathbf{\Sigma} \in \mathcal{S}^d_{++} : R(\mathbf{\Sigma}) = 1\}$), is an irreducible, aperiodic and positive Harris-recurrent Markov chain with (unique) invariant probability measure π. Assume moreover that the functions

$$(z, p, q, \mathbf{\Sigma}, r) \in \mathsf{Y} \mapsto \log \|z\|, \log \Gamma(p), \log r \tag{2.22}$$

are π-integrable. Then the CMA-ES algorithm behaves globally asymptotically linearly almost surely:

$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}\right] = \int \left(\log \Gamma(p) + \frac{1}{2} \log r\right) \mathrm{d}\pi \ . \tag{2.23}$$

Proof. The almost sure limit of the LHS in (2.23) follows directly from (2.21), LLN for ergodic chains [110, Theorem 17.0.1]. The limit of the expectation in (2.23) follows from the ergodic theorem [110, Theorem 14.0.1], since

$$\log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|} = \log \|z_{t+1}\| - \log \|z_t\| + \log \Gamma(p_{t+1}) + \frac{1}{2} \log r_{t+1} \ .$$

□

The previous proposition illustrates that proving irreducibility and aperiodicity of the chain Φ is a stepping stone to establish linear convergence of CMA-ES. Proving these properties will occupy Section 5. Later, we intend to prove the geometric ergodicity by means of Foster-Lyapunov drift conditions [110, Theorem 15.0.1] that depend on small sets (as given in Theorem 2.1). Characterizing small sets is facilitated by the topological T-chain property as formalized in the next section.

3 Main Results I: Irreducibility, aperiodicity, and T-chain property of normalized Markov chains underlying the CMA-ES algorithm

We present in this section one of the two main results of this paper stating the irreducibility, aperiodicity and T-chain property of the normalized chains underlying the CMA-ES algorithm defined in (2.17). We start by introducing the definitions of irreducibility, aperiodicity and T-kernel. Let P be a transition kernel on a state space $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. We say that P is irreducible when there exists a nontrivial nonnegative measure φ on $\mathcal{B}(\mathsf{X})$ such that, for every $x \in \mathsf{X}$ and every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ with $\varphi(\mathsf{A}) > 0$, there exists a positive integer k satisfying $P^k(x, \mathsf{A}) > 0$. When a measure φ satisfies this definition, we say that P is φ-irreducible.

When P is irreducible, the period of P is the largest integer $k \geqslant 1$ such that there exist disjoint sets $\mathsf{D}_1, \ldots, \mathsf{D}_k \in \mathcal{B}(\mathsf{X})$ with

$$\begin{cases} \varphi((\mathsf{D}_1 \cup \cdots \cup \mathsf{D}_k)^c) = 0 \text{ for every irreducibility measure } \varphi \text{ of } P \\ P(x_i, \mathsf{D}_{i+1}) = 1 \text{ for } x_i \in \mathsf{D}_i \text{ and } i = 0, \ldots, k-1 \ (\mathrm{mod}\ k). \end{cases} \tag{2.24}$$

An irreducible transition kernel P always admits a period $k \geqslant 1$ [110, Theorem 5.4.4], and when $k = 1$, P is said to be aperiodic.

For any positive integer m, a set $\mathsf{C} \in \mathcal{B}(\mathsf{X})$ is called m-small when there exists a nontrivial measure ν_m on $\mathcal{B}(\mathsf{X})$ such that $P^m(x, \mathsf{A}) \geqslant \nu_m(\mathsf{A})$ for every $x \in \mathsf{C}$ and every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$.

Given a probability distribution b on $\mathbb{N}$, we define the transition kernel K_b on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ as $K_b(x, \mathsf{A}) = \sum_{k \geqslant 0} b(k) P^k(x, \mathsf{A})$.

A substochastic kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ is a function $T \colon \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}$ such that $T(\cdot, \mathsf{A})$ is measurable for every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ and $T(x, \cdot)$ is a finite measure on $\mathcal{B}(\mathsf{X})$ with $T(x, \mathsf{X}) \leqslant 1$ for every $x \in \mathsf{X}$. We say that the substochastic kernel T is a *continuous component* of the transition kernel K_b when $T(\cdot, A)$ is lower semicontinuous on X, $T(x, \mathsf{X}) > 0$ and $K_b(x, \mathsf{A}) \geqslant T(x, \mathsf{A})$ for every $x \in \mathsf{X}$ and $\mathsf{A} \in \mathcal{B}(\mathsf{X})$. A transition kernel P on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ is called a T-kernel when there exist a probability measure b on $\mathbb{N}$ and a substochastic kernel T which is a continuous component of the transition kernel K_b. Moreover, we say that a Markov chain is irreducible, respectively aperiodic, a T-chain, when its transition kernel is irreducible, respectively aperiodic, a T-kernel. We can now state our first main contribution presented in the next theorem and its corollary. They constitute a first milestone towards a linear convergence proof of CMA-ES. The complete proof of the following theorem is presented in Section 5 (cf. Theorem 2.5).

Theorem 2.1. Suppose that the objective function f, the normalization function R, the stepsize change Γ and the sampling distribution ν_U^d satisfy **F**1-**F**2, **R**1-**R**3, **Γ**1-**Γ**3 and **N**1, respectively. Let $\Phi = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t \geqslant 1}$ be the normalized Markov chain underlying CMA-ES defined via (2.17) and P its transition kernel. Assume that $0 < c_1 + c_\mu < 1$. Then,

(i) if $c_c, c_\sigma \in (0, 1)$, $c_\mu > 0$ and $1 - c_c \neq (1 - c_\sigma)\sqrt{1 - c_1 - c_\mu}$, then P is an irreducible aperiodic T-kernel, such that compact sets of $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ are small;

(ii) if $c_c \in (0, 1)$, $c_\sigma = 1$ and $c_\mu > 0$, then the process $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t \geqslant 1}$ is an irreducible aperiodic T-chain, such that compact sets of $\mathbb{R}^d \times \mathbb{R}^d \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ are small;

(iii) if $c_\sigma \in (0, 1)$ and $c_c = 1$, then the process $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t \geqslant 1}$ is an irreducible aperiodic T-chain, such that compact sets of $\mathbb{R}^d \times \mathbb{R}^d \times R^{-1}(\{1\})$ are small;

(iv) if $c_c = c_\sigma = 1$, then the process $\{(z_t, \mathbf{\Sigma}_t)\}_{t \geqslant 1}$ is an irreducible aperiodic T-chain, such that compact sets of $\mathbb{R}^d \times R^{-1}(\{1\})$ are small.

This result covers the entire range of eligible hyperparameter settings for CMA-ES except when $c_1 + c_\mu = 1$, or $c_\mu = 0$ and $c_c < 1$, or $1 - c_c = (1 - c_\sigma)\sqrt{1 - c_1 - c_\mu} > 0$. Most importantly, when cumulation is used in the rank-one update, we need for our proof the rank-mu update. Without cumulation however ($c_c = 1$), the rank-one update is already sufficient to prove irreducibility and aperiodicity.

We finally formulate a particular case of Theorem 2.1. Using Proposition 2.1, Lemma 2.1 and Theorem 2.1, we find that Markov chains obtained with some standard stepsize change of CMA-ES and normalized by its minimum eigenvalue (possibly expressed in a different coordinate system which would be more fitted to the objective function f) or $\det(\cdot)^{1/d}$ are irreducible, aperiodic T-chains.

Corollary 2.1. Let $\mathbf{H} \in \mathcal{S}_{++}^d$. Consider the process Φ defined via (2.17) with a normalization function $R = \det(\cdot)^{1/d}$ or $R = \lambda_{\min}(\mathbf{H}^{1/2} \times \cdot \times \mathbf{H}^{1/2})/\lambda_{\min}(\mathbf{H})$ and with the CSA stepsize change $\Gamma = \Gamma_{\text{CSA}}^1$ or $\Gamma = \Gamma_{\text{CSA}}^2$, see (2.12) or (2.13), respectively. Assume as in Theorem 2.1 that f satisfies **F**1-**F**2 and the sampling distribution satisfies **N**1, then under the same conditions

on the hyperparameters as in Theorem 2.1, Φ is an irreducible aperiodic T-chain.

4 Main results II: Extension of the analysis of nonlinear state-space models

We present in this section our methodological extensions of tools to analyze the irreducibility, aperiodicity and T-chain property of Markov chains. After reminding the basics in Section 4.1, we present two extensions.

First, some of the learning rate settings from Theorem 2.1(i), (ii), (iii) and (iv) give rise to a so-called *redundant Markov chain*, where one state variable can be dropped to define another Markov chain. We thus introduce redundant and projected Markov chains in Section 4.2 and explain how irreducibility, aperiodicity and T-chain property of the projected chain can be deduced from an analysis of the redundant chain. The main result of this section is Theorem 2.3.

The second methodological extension is motivated by the Markov chain (2.18) which is valued in a possibly nonsmooth manifold since the normalization $R(\cdot)$ may be not continuously differentiable, for instance when $R(\cdot) = \lambda_{\min}(\cdot)$. To analyze such a chain, we apply a homeomorphic transformation, thereby defining a Markov chain valued in a smooth manifold, and explain how irreducibility, aperiodicity and the T-chain property of the original Markov chain can be deduced from an analysis of the transformed Markov chain.

These results are applied in Section 5 for the proof of Theorem 2.1.

4.1 Deterministic control model and sufficient conditions for irreducibility and aperiodicity

We introduce in this section different definitions and theorems our analysis is based on the original article [52] to which we refer for more details. Let X and V be two smooth connected manifolds,[3] equipped with their Borel σ-fields, denoted $\mathcal{B}(\mathsf{X})$ and $\mathcal{B}(\mathsf{V})$, respectively. We later denote the dimension of X by n. The tangent space of X at a point $x \in \mathsf{X}$ is denoted $\mathrm{T}_x\mathsf{X}$, and we denote $\mathrm{dist}_{\mathsf{X}}$ and $\mathrm{dist}_{\mathsf{V}}$ the distance functions on X and V, respectively, which induce their respective topology. Consider a transition kernel P on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ associated to the Markov chain following the update equation

$$\phi_{t+1} = F(\phi_t, \alpha(\phi_t, U_{t+1})) \tag{2.25}$$

where $F\colon \mathsf{X} \times \mathsf{V} \to \mathsf{X}$ and $\alpha\colon \mathsf{X} \times \mathsf{U} \to \mathsf{V}$ are measurable functions, and $\{U_{t+1}\}_{t\in\mathbb{N}}$ is an i.i.d. process independent of ϕ_0 and valued in a measurable space $(\mathsf{U}, \mathcal{U})$, where $\mathcal{U}$ is a σ-field of U.[4] We consider additionally the following assumptions on the model.

H1. *For any $x \in \mathsf{X}$, the distribution μ_x of the random variable $\alpha(x, U_1)$ admits a density p_x with respect to a σ-finite measure ζ_{V} on V, such that*

(i) the function $(x, v) \mapsto p_x(v)$ is lower semicontinuous;

(ii) for $\mathsf{A} \in \mathcal{B}(\mathsf{V})$, $\zeta_{\mathsf{V}}(\mathsf{A}) = 0$ if and only if A is negligible, i.e., $\mathrm{Leb}(\varphi(\mathsf{A} \cap V)) = 0$ for every local chart (φ, V) of V.

H2. *The function $F : \mathsf{X} \times \mathsf{V} \to \mathsf{X}$ is locally Lipschitz (with respect to the metrics $\mathrm{dist}_{\mathsf{X}} \oplus \mathrm{dist}_{\mathsf{V}}$ and $\mathrm{dist}_{\mathsf{X}}$).*

[3] In the rest of the paper, manifolds will be considered as connected.

[4] Since we do not assume U to be a topological space, we consider a general σ-field $\mathcal{U}$ instead of its Borel σ-field.

Below, Proposition 2.8 provides the Markov chain (2.38) that follows the control model (2.25) and satisfies **H**1 and **H**2 under mild assumptions on the objective function f and the stepsize change Γ.

We define inductively the *extended transition map* $S_x^k \colon \mathsf{V}^k \to \mathsf{X}$ associated to (2.25) for any $k \in \mathbb{N}$, $x \in \mathsf{X}$ and $v_{1:k} = (v_1, \dots, v_k) \in \mathsf{V}^k$ as follows

$$\begin{cases} S_x^0 \coloneqq x \\ S_x^k(v_{1:k}) \coloneqq F\left(S_x^{k-1}(v_{1:k-1}), v_k\right) \quad \text{for } k \geqslant 1. \end{cases} \tag{2.26}$$

From this definition, we obtain that if F is locally Lipschitz (respectively differentiable), then $(x, v_{1:k}) \mapsto S_x^k(v_{1:k})$ is locally Lipschitz (respectively differentiable). Likewise, we define the *extended probability density* $p_x^k \colon \mathsf{V}^k \to \mathbb{R}_+$ by

$$\begin{cases} p_x^1(v_1) \coloneqq p_x(v_1) \\ p_x^k(v_{1:k}) \coloneqq p_x^{k-1}(v_{1:k-1}) \times p_{S_x^{k-1}(v_{1:k-1})}(v_k) \quad \text{for } k \geqslant 2. \end{cases} \tag{2.27}$$

Given the Markov chain $\{\phi_t\}_{t\in\mathbb{N}}$ defined via (2.25), the function p_x^k is a density associated to the random variable $(\alpha(\phi_0, U_1), \dots, \alpha(\phi_{k-1}, U_k))$, when $\phi_0 = x$. For $x \in \mathsf{X}$ and $k \in \mathbb{N}^*$, we define the *control sets* of (2.25) by

$$\mathcal{O}_x^k \coloneqq \left\{v_{1:k} \in \mathsf{V}^k \mid p_x^k(v_{1:k}) > 0\right\} \ . \tag{2.28}$$

Assumption **H**1(i) implies that these sets are open subsets of V^k. We define moreover

$$\mathcal{O}_x^\infty \coloneqq \left\{v_{1:\infty} \in \mathsf{V}^{\mathbb{N}} \mid \forall k \geqslant 1, v_{1:k} \in \mathcal{O}_x^k\right\} \ . \tag{2.29}$$

We say that $x^* \in \mathsf{X}$ is a *steadily attracting state*, when for every $x \in \mathsf{X}$ and every neighborhood U of x^*, there exists $T > 0$ such that for every $k \geqslant T$, there exists $v_{1:k} \in \mathcal{O}_x^k$ such that $S_x^k(v_{1:k}) \in U$. When F is continuous, $v_{1:k}$ can be taken in $\overline{\mathcal{O}_x^k}$ [52, Corollary 4.5], i.e., $x^* \in \mathsf{X}$ is steadily attracting if and only if for every neighborhood U of x^*, there exists $T > 0$ such that for every $k \geqslant T$, there exists $v_{1:k} \in \overline{\mathcal{O}_x^k}$ such that $S_x^k(v_{1:k}) \in U$. In particular, if for every $x \in \mathsf{X}$, there exists $v_{1:\infty} \in \overline{\mathcal{O}_x^\infty}$ such that $S_x^k(v_{1:k})$ tends to x^*, then x^* is a steadily attracting state [52, Corollary 4.5].

We formulate now the following controllability condition of a steadily attracting state.

H3. *There exist a steadily attracting state $x^* \in \mathsf{X}$, an integer $k > 0$, and a path $v_{1:k}^* \in \overline{\mathcal{O}_{x^*}^k}$, such that $\partial S_{x^*}^k(v_{1:k}^*)$ is of maximal rank.*

For a locally Lipschitz function $G \colon \mathsf{V}^k \to \mathsf{X}$, $\partial G(v)$ is the Clarke's derivative of G at a point $v \in \mathsf{V}^k$, which is a set of linear applications between $\mathrm{T}_v\mathsf{V}^k$ and $\mathrm{T}_{G(v)}\mathsf{X}$ [52, Appendix B]. If G is differentiable at v, then $\partial G(v) = \{\mathcal{D}G(v)\}$, where $\mathcal{D}G(v)$ denotes the usual differential application of G in v. We then say that $\partial G(v)$ is of maximal rank when its elements are of maximal rank, that is, of rank n (the dimension of X). When G is differentiable at v and $\mathcal{D}G(v)$ is of maximal rank, then $\partial G(v) = \{\mathcal{D}G(v)\}$ is of maximal rank. We base our analysis on the following statement.

Theorem 2.2 (Sufficient conditions for irreducibility and aperiodicity [52, Theorem 2.3])**.** Consider the Markov kernel P defined via (2.25) such that **H**1-**H**3 are satisfied. Then P is an irreducible, aperiodic T-kernel, and every compact set of X is small.

This theorem summarises the methodology we follow to analyze a normalized Markov chain underlying CMA-ES: we prove that the chain satisfies (2.25) as well as conditions **H**1-**H**3[5] under appropriate conditions on the learning rates, as well as on the functions f, Γ and R, and on the sampling distribution ν_U^d.

[5]Condition **H**1 introduced in Section 4.2 is required instead of **H**3 when $c_c = 1$ or $c_\sigma = 1$.

4.2 *Irreducibility and aperiodicity of a projected Markov chain*

The CMA-ES algorithm maintains two paths p_t^c and p_t^σ that do not parametrize the probability distribution for sampling candidate solutions but are used for (accelerating) the update of the covariance matrix and the stepsize, respectively. Yet, when no cumulation for the stepsize path is used, i.e., $c_\sigma = 1$, or no cumulation for the rank-one update path is used, i.e., $c_c = 1$, the CMA-ES algorithm typically still works properly while it is sometimes slower [49]. In these cases, the normalized Markov chain underlying CMA-ES can be described with fewer variables: p_t^c and p_t^σ boil down to random vectors that depend on the previous step only through the ranking permutation of candidate solutions. In order to analyze those algorithm variants without repeating proofs with small variations, we introduce here a method that allows to derive properties for a projected Markov chain from a redundant Markov chain with a specific parameter setting. We define a *redundant* Markov chain as a Markov chain $\{(\phi_t, \xi_t)\}_{t\in\mathbb{N}}$ valued in a topological product space $\mathsf{X} \times \mathsf{Y}$ such that the process $\{\phi_t\}_{t\in\mathbb{N}}$ also is a Markov chain, valued in X. In that case, we say that $\{\phi_t\}_{t\in\mathbb{N}}$ is a *projected* Markov chain of $\{(\phi_t, \xi_t)\}_{t\in\mathbb{N}}$.

Prior to that, we formalize the simplification of the normalized Markov chain of CMA-ES when at least one cumulation parameter is set to 1. The proof is a direct consequence of Proposition 2.2 and thus omitted.

Corollary 2.2. Suppose that the objective function f and that the normalization function R satisfy **F**2 and **R**1, respectively. Let $\{(m_t, p_t^\sigma, p_t^c, \mathbf{C}_t, \sigma_t)\}_{t\in\mathbb{N}}$ be the Markov chain associated to CMA-ES defined in Section 2.1 and $\Phi = \{\phi_t\}_{t\geqslant 1} = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ be the normalized process defined in (2.17).

(i) If $c_\sigma = 1$, then the process $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ defines a (time-homogeneous) Markov chain.

(ii) If $c_c = 1$, then the process $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ defines a (time-homogeneous) Markov chain.

(iii) If $c_\sigma = c_c = 1$, then the process $\{(z_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ defines a (time-homogeneous) Markov chain.

Corollary 2.2 motivates the introduction of the notion of a *projected chain* of a Markov chain $\tilde{\Phi}$, and to provide conditions for irreducibility and aperiodicity as in Theorem 2.2.

Define $\tilde{\Phi} = \{(\phi_t, \chi_t)\}_{t\in\mathbb{N}}$ a so-called *redundant* Markov chain on $(\mathsf{X} \times \mathsf{Y}, \mathcal{B}(\mathsf{X} \times \mathsf{Y}))$, with transition kernel $\tilde{P}$, such that

$$(\phi_{t+1}, \chi_{t+1}) = \tilde{F}(\phi_t, \chi_t, \tilde{\alpha}(\phi_t, \chi_t, U_{t+1})) \tag{2.30}$$

where $\tilde{F}\colon \mathsf{X} \times \mathsf{Y} \times \mathsf{V} \to \mathsf{X} \times \mathsf{Y}$ and $\tilde{\alpha}\colon \mathsf{X} \times \mathsf{Y} \times \mathsf{U} \to \mathsf{V}$ are measurable maps, $\mathsf{X}, \mathsf{Y}, \mathsf{V}$ are (smooth, connected) manifolds, $(\mathsf{U}, \mathcal{U})$ is a measurable space and $\{U_{t+1}\}_{t\in\mathbb{N}}$ is an i.i.d. process valued in U, independent of (ϕ_0, χ_0). Assume **H**1-**H**2, and denote $\tilde{S}^k_{(x,y)}$, $\tilde{p}^k_{(x,y)}$ and $\tilde{\mathcal{O}}^k_{(x,y)}$ the extended transition map, the extended probability density and the control sets associated to the control model (2.30), for every $(x, y) \in \mathsf{X} \times \mathsf{Y}$ and $k \in \mathbb{N}$, respectively. Besides, we suppose redundancy of the chain by assuming that the function $\tilde{\alpha}$ does not depend on the variable χ, i.e., there exists a function α such that

$$\tilde{\alpha}(\phi, \chi, u) = \alpha(\phi, u) \quad \text{for every } \phi \in \mathsf{X}, \chi \in \mathsf{Y}, u \in \mathsf{U}. \tag{2.31}$$

Furthermore, we suppose that $\Phi = \{\phi_t\}_{t\in\mathbb{N}}$ is a Markov chain on X with transition kernel denoted P, following the next deterministic control model

$$\phi_{t+1} = F(\phi_t, \alpha(\phi_t, U_{t+1})), \tag{2.32}$$

with $\{U_{t+1}\}_{t\in\mathbb{N}}$ being the i.i.d. process introduced to define the redundant chain via (2.30). Then, we say that Φ is a projected chain of $\tilde{\Phi}$. As above, we denote S_x^k, p_x^k and $\mathcal{O}_x^k$ the extended transition map, the extended probability density and the control sets associated to the control model (2.32), for every $x \in \mathsf{X}$ and $k \in \mathbb{N}$, respectively. The next proposition connects the assumptions required for the two deterministic control models (2.30) and (2.32) that are useful to show that $\tilde{\Phi}$ and Φ are irreducible aperiodic T-chains.

Proposition 2.4. Consider the control models associated to the redundant chain (2.30) and its associated projected chain (2.32). Then,

(i) if **H**1 (resp. **H**2) is satisfied for the redundant chain (2.30), then it is satisfied for its projected chain (2.32);

(ii) the closures of the control sets $\mathcal{O}_\phi^k$ of the projected chain (2.32) equal the closures of the control sets $\tilde{\mathcal{O}}_{(\phi,\chi)}^k$ of the redundant chain (2.30), that is,

$$\overline{\mathcal{O}_\phi^k} = \overline{\tilde{\mathcal{O}}_{(\phi,\chi)}^k} \quad \text{for every } \phi \in \mathsf{X}, \chi \in \mathsf{Y}, k \geqslant 1; \tag{2.33}$$

(iii) the extended transition maps S_ϕ^k and $\tilde{S}_{(\phi,\chi)}^k$, defined in (2.26), of the control models of the projected chain (2.32), and the redundant chain (2.30), respectively, satisfy

$$S_\phi^k = \Pi_{\mathsf{X}} \circ \tilde{S}_{(\phi,\chi)}^k \quad \text{for every } \phi \in \mathsf{X}, \chi \in \mathsf{Y}, k \geqslant 1, \tag{2.34}$$

where $\Pi_{\mathsf{X}} \colon \mathsf{X} \times \mathsf{Y} \to \mathsf{X}$ is the canonical projection of $\mathsf{X} \times \mathsf{Y}$ on X.

Proof. First, we prove (i). Suppose that the redundant Markov chain following (2.30) satisfies **H**1, i.e., for all $(\phi, \chi) \in \mathsf{X} \times \mathsf{Y}$, the random variable $\tilde{\alpha}(\phi, \chi, U_1)$ admits a density $\tilde{p}_{(\phi,\chi)}$ with respect to a σ-finite measure ζ_{V} satisfying **H**1, such that $(\phi, \chi, v) \mapsto \tilde{p}_{(\phi,\chi)}(v)$ is lower semicontinuous. Let $\phi \in \mathsf{X}$. By (2.31) we have $\alpha(\phi, U_1) = \tilde{\alpha}(\phi, \chi, U_1)$ for every $\chi \in \mathsf{Y}$. Let $\chi_0 \in \mathsf{Y}$. Then the random variable $\tilde{\alpha}(\phi, \chi_0, U_1)$ admits a density $\tilde{p}_{(\phi,\chi_0)}$ with respect to a measure ζ_{V} on V satisfying **H**1(ii), and $(\phi, v) \mapsto \tilde{p}_{(\phi,\chi_0)}(v)$ is lower semicontinuous. Hence, $\alpha(\phi, U_1)$ also admits a density with respect to ζ_{V} denoted p_ϕ, such that $p_\phi(v) = \tilde{p}_{(\phi,\chi_0)}(v)$ for every $v \in \mathsf{V}$. By uniqueness of the density up to a null set, we obtain that, for $\phi \in \mathsf{X}$ and $\chi \in \mathsf{Y}$

$$p_\phi(v) = \tilde{p}_{(\phi,\chi)}(v) \quad \text{for } \zeta_{\mathsf{V}}\text{-almost every } v \in \mathsf{V}. \tag{2.35}$$

Since $(\phi, v) \mapsto p_{(\phi,\chi_0)}(v)$ is lower semicontinuous, we obtain that $(\phi, v) \mapsto p_\phi(v) = p_{(\phi,\chi_0)}(v)$ is lower semicontinuousand thus the projected chain satisfies **H**1. For assumption **H**2, if $\tilde{F}$ is locally Lipschitz, then, since by definition we have $F(\phi, v) = \Pi_{\mathsf{X}} \circ \tilde{F}(\phi, \chi, v)$ for $\phi \in \mathsf{X}$, $\chi \in \mathsf{Y}$ and $v \in \mathsf{V}$, by composition F is locally Lipschitz (Y being nonempty).
For (ii), from (2.35) and by definition of the control sets in (2.28), for every $(\phi, \chi) \in \mathsf{X} \times \mathsf{Y}$, $\mathcal{O}_\phi^k$ and $\tilde{\mathcal{O}}_{(\phi,\chi)}^k$ only differ by a ζ_{V}-negligible set. Besides, both are open sets and ζ_{V} is ,by **H**1(ii), a Borel measure, so $\overline{\mathcal{O}_\phi^k} = \overline{\tilde{\mathcal{O}}_{(\phi,\chi)}^k}$ (as a direct consequence of Carathéodory's criterion of Borel measures [43, Theorem 1.9]).
In order to prove (iii), we proceed by induction. Indeed, $S_\phi^0 = \phi = \Pi_{\mathsf{X}}(\phi, \chi) = \Pi_{\mathsf{X}} \circ \tilde{S}_{(\phi,\chi)}^0$. Let $k \geqslant 0$ and assume $S_\phi^k = \Pi_{\mathsf{X}} \circ \tilde{S}_{(\phi,\chi)}^k$. Let $v_{1:k+1} \in \mathsf{V}^{k+1}$, we find $S_\phi^{k+1}(v_{1:k+1}) = F(S_\phi^k(v_{1:k}), v_{k+1}) = \Pi_{\mathsf{X}} \circ \tilde{F}(S_\phi^k(v_{1:k}), \chi_k, v_{k+1})$ where $\chi_k = \Pi_{\mathsf{Y}} \circ \tilde{S}_{(\phi,\chi)}^k(v_{1:k}) \in \mathsf{Y}$. By

induction hypothesis we find that $\Pi_{\mathsf{X}} \circ \tilde{F}(S^k_\phi(v_{1:k}), \chi_k, v_{k+1}) = \Pi_{\mathsf{X}} \circ \tilde{F}(\tilde{S}^k_{(\phi,\chi)}(v_{1:k}), v_{k+1}) = \Pi_{\mathsf{X}} \circ \tilde{S}^{k+1}_{(\phi,\chi)}(v_{1:k+1})$ and thus $S^{k+1}_\phi(v_{1:k+1}) = \Pi_{\mathsf{X}} \circ \tilde{S}^{k+1}_{(\phi,\chi)}(v_{1:k+1})$. Hence $S^{k+1}_\phi = \Pi_{\mathsf{X}} \circ \tilde{S}^{k+1}_{(\phi,\chi)}$. □

We deduce the following result, which characterizes the controllability condition for the projected Markov chain.

Proposition 2.5. Assume that F is continuous.[a] If $x^* = (\phi^*, \chi^*) \in \mathsf{X} \times \mathsf{Y}$ is a steadily attracting state of the redundant chain (2.30), then ϕ^* is a steadily attracting state of the projected chain (2.32).
Suppose moreover that there exists $k \geqslant 1$ and $v^*_{1:k} \in \overline{\tilde{\mathcal{O}}^k_{x^*}}$ such that $\tilde{S}^k_{x^*}$ is differentiable at $v^*_{1:k}$, and for every $h^\phi \in \mathrm{T}_{\phi_k}\mathsf{X}$, there exists $h^\chi \in \mathrm{T}_{\chi_k}\mathsf{Y}$, where $(\phi_k, \chi_k) = \tilde{S}^k_{x^*}(v^*_{1:k})$ and with $(h^\phi, h^\chi) \in \operatorname{rge} \mathcal{D}\tilde{S}^k_{x^*}(v^*_{1:k})$. Then $v^*_{1:k} \in \overline{\mathcal{O}^k_{\phi^*}}$ and $\mathcal{D}S^k_{\phi^*}(v^*_{1:k})$ exists and is of maximal rank.

[a] Alternatively, we can assume without loss of generality that for every $\phi \in \mathsf{X}$, the density functions $\tilde{p}_{(\phi,\chi)}$ are identical for $\chi \in \mathsf{Y}$.

Proof. Since F is continuous, we can use the definition of a steadily attracting set via taking the elements for the k-step paths within the closure of the control sets (see Section 4.1) instead of the control sets. According to the previous proposition $\overline{\mathcal{O}^k_\phi} = \overline{\tilde{\mathcal{O}}^k_{(\phi,\chi)}}$ for every $\phi \in \mathsf{X}, \chi \in \mathsf{Y}, k \geqslant 1$ and we can thus easily prove that ϕ^* is steadily attracting for (2.32) when (ϕ^*, χ^*) is steadily attracting for (2.30). Moreover, by Proposition 2.4(iii), we have that $S^k_{\phi^*} = \Pi_{\mathsf{X}} \circ \tilde{S}^k_{(\phi^*,\chi^*)}$. Then, by the chain rule [30, Corollary 2.6.6], we have

$$\begin{aligned}\mathcal{D}S^k_{\phi^*}(v^*_{1:k}) &= \mathcal{D}\Pi_{\mathsf{X}}(\tilde{S}^k_{(\phi^*,\chi^*)}(v^*_{1:k})) \circ \mathcal{D}\tilde{S}^k_{(\phi^*,\chi^*)}(v^*_{1:k}) \\ &= \Pi_{\mathrm{T}_{S^k_{\phi^*}(v^*_{1:k})}\mathsf{X}} \circ \mathcal{D}\tilde{S}^k_{(\phi^*,\chi^*)}(v^*_{1:k}).\end{aligned}$$

Therefore every $h^\phi \in \mathrm{T}_{S^k_{\phi^*}(v^*_{1:k})}\mathsf{X}$ belongs to the range of $\mathcal{D}S^k_{\phi^*}(v^*_{1:k})$ (we use here that by assumption there exists $h^\chi \in \mathrm{T}_{\Pi_{\mathsf{Y}} S_{(\phi^*,\xi^*)}(v^*_{1:k})}$), making it a surjective linear map, hence of maximal rank. □

As a consequence, we derive sufficient conditions for irreducibility and aperiodicity of the kernel P of the projected Markov chain Φ. We replace the controllability condition **H**3 by the following.

H1. *There exist a steadily attracting x^*, an integer $k > 0$ and a path $v^*_{1:k} \in \overline{\tilde{\mathcal{O}}^k_{x^*}}$ such that $S^k_{x^*}$ is differentiable at $v^*_{1:k}$, and for every $h^\phi \in \mathrm{T}_\phi\mathsf{X}$, there exists $h^\chi \in \mathrm{T}_\chi\mathsf{Y}$ with $(h^\phi, h^\chi) \in \operatorname{rge} \mathcal{D}\tilde{S}^k_{x^*}(v^*_{1:k})$, where $(\phi, \chi) = \tilde{S}^k_{x^*}(v^*_{1:k})$.*

Theorem 2.3 (Sufficient conditions for irreducibility and aperiodicity of a projected Markov chain)**.** Consider the control model of the redundant chain (2.30) and assume it satisfies conditions **H**1-**H**2 and **H**1. Then the kernel P of the projected chain defined via (2.32) is an irreducible aperiodic T-kernel, and every compact set of X is small.

Proof. Denote $x^* = (\phi^*, \chi^*)$. Since x^* is steadily attracting for (2.30) and since F is continuous by **H**2, then by Proposition 2.5 ϕ^* is steadily attracting for (2.32). Besides, by Proposition 2.4(ii), we have $v^*_{1:k} \in \overline{\mathcal{O}^k_{x^*}} = \overline{\tilde{\mathcal{O}}^k_{\phi^*}}$. By Proposition 2.5 again, we find that $\mathcal{D}S^k_{\phi^*}(v^*_{1:k})$ exists and is of maximal rank. We complete the proof by applying Theorem 2.2. □

Theorem 2.3 is used later to analyze the normalized chain defined in (2.17) when $c_c = 0$ or $c_\sigma = 0$. Even though Theorem 2.2 could be applied directly to the projected chain, this generalization allows to find a steadily attracting state and prove a controllability condition on the same chain for all settings without repeating the same proof.

4.3 *Homeomorphic transformation of an irreducible aperiodic T-chain*

The state space of the chain Φ defined via (2.17) is not a smooth manifold if the normalization function R is not continuously differentiable on $\mathcal{S}_{++}^d$. In order to include nonsmooth functions R in our analysis (for instance if $R(\cdot)$ is the minimal eigenvalue of a positive definite matrix), we apply Theorem 2.2 (or Theorem 2.3 when we do not have cumulation) to a Markov chain Θ, defined as a homeomorphic transformation of Φ, such that the state space of Θ is a smooth manifold. This is achieved in Sections 5.1 to 5.4. Now, we explain why it is sufficient to prove that the transformed chain Θ is an irreducible, aperiodic T-chain to have the same properties on Φ.

Theorem 2.4. Let $\xi \colon \mathsf{Y} \to \mathsf{X}$ be a homeomorphism between the topological spaces Y and X, equipped with their respective Borel σ-fields. Let $\Phi = \{\phi_t\}_{t\in\mathbb{N}}$ be a (time-homogeneous) Markov chain with state space Y, and define $\Theta = \{\xi(\phi_t)\}_{t\in\mathbb{N}}$. Then,

(i) Θ is a (time-homogeneous) Markov chain with state space X;

(ii) if Θ is irreducible (resp. aperiodic, a T-chain), then Φ is irreducible (resp. aperiodic, a T-chain).

Proof. First, we prove (i). Denote P the Markov kernel of Φ. If the distribution of $\xi(\phi_0)$ is δ_x for $x \in \mathsf{X}$, then ϕ_0 is distributed under $\delta_{\xi^{-1}(x)}$. For $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ and $x \in \mathsf{X}$, we have then

$$\mathbb{P}\left[\xi(\phi_1) \in \mathsf{A} \mid \xi(\phi_0) = x\right] = \mathbb{P}\left[\phi_1 \in \xi^{-1}(\mathsf{A}) \mid \phi_0 = \xi^{-1}(x)\right] = P(\xi^{-1}(x), \xi^{-1}(\mathsf{A}))$$

Moreover, $P(\xi^{-1}(\cdot), \xi^{-1}(\cdot))$ defines a Markov kernel for Θ. Indeed, since ξ is a homeomorphism, ξ^{-1} is continuous and thus measurable. In particular the k-step transition kernel of Θ equals $P^k(\xi^{-1}(\cdot), \xi^{-1}(\cdot))$ (where P^k is the k-step transition kernel of Φ). Thus Θ is a time-homogeneous Markov chain.
Now we prove (ii). Suppose that Θ is irreducible, i.e., the kernel $P(\xi^{-1}(\cdot), \xi^{-1}(\cdot))$ admits a nontrivial nonnegative measure ϑ on $\mathcal{B}(\mathsf{X})$ such that for $x \in \mathsf{X}$ and $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ with $\vartheta(\mathsf{A}) > 0$, there exists $k > 0$ with $P^k(\xi^{-1}(x), \xi^{-1}(\mathsf{A})) > 0$. Then, for every $\mathsf{B} \in \mathcal{B}(\mathsf{Y})$ such that $\vartheta(\xi(\mathsf{B})) > 0$ and for every $y \in \mathsf{Y}$, there exists $k > 0$ with $P^k(x, \mathsf{B}) > 0$, i.e., Φ is $\vartheta \circ \xi$-irreducible. Likewise, for every irreducibility measure φ of Φ, then $\varphi \circ \xi^{-1}$ is a irreducibility measure of Θ. In particular, every irreducibility measure ϑ of Θ can be defined as $\varphi \circ \xi^{-1}$ for some irreducibility measure φ of Φ. Moreover, denote $k \geqslant 1$ the period of Θ, i.e., k is the largest integer such that there exists disjoint sets $\mathsf{D}_1, \dots, \mathsf{D}_k \in \mathcal{B}(\mathsf{Y})$ with

$$\begin{cases} \varphi((\mathsf{D}_1 \cup \cdots \cup \mathsf{D}_k)^c) = 0 \quad \text{for any irreducibility measure } \varphi \text{ of } \Phi \\ P(\xi^{-1}(y_i), \xi^{-1}(\mathsf{D}_{i+1})) = 1 \text{ for } y_i \in \mathsf{D}_i \text{ and } i = 0, \dots, k-1 \mod k. \end{cases}$$

Therefore k is the largest integer such that there exists disjoint sets $\mathsf{C}_1, \dots, \mathsf{C}_k \in \mathcal{B}(\mathsf{X})$ with

$$\begin{cases} \varphi(\xi^{-1}(\mathsf{C}_1 \cup \cdots \cup \mathsf{C}_k)^c) = 0 \quad \text{for any irreducibility measure } \varphi \text{ of } \Phi \\ P(x_i, \mathsf{C}_{i+1}) = 1 \text{ for } x_i \in \mathsf{C}_i \text{ and } i = 0, \dots, k-1 \mod k. \end{cases}$$

Hence, the period of Φ equals k the period of Θ. In particular, if Θ is aperiodic, then Φ is aperiodic.
Suppose now that Θ is a T-chain and let $T\colon \mathsf{X}\times\mathcal{B}(\mathsf{X})\to\mathbb{R}_+$ be a substochastic kernel such that $K_b(\xi^{-1}(\cdot),\xi^{-1}(\cdot))\geqslant T$ for some probability distribution b on $\mathbb{N}$, $T(\cdot,\mathsf{X})>0$ and $x\mapsto T(x,\mathsf{A})$ is lower semicontinuousfor $\mathsf{A}\in\mathcal{B}(\mathsf{X})$. Then, if we define $T'(y,\mathsf{B})=T(\xi(y),\xi(\mathsf{B}))$ for $y\in\mathsf{Y}$ and $\mathsf{B}\in\mathcal{B}(\mathsf{Y})$, we obtain that T' is a substochastic kernel such that $K_b\geqslant T'$ for some probability distribution b on $\mathbb{N}$, $T'(\cdot,\mathsf{Y})>0$ and that $y\mapsto T(y,\mathsf{B})$ is lower semicontinuousfor every $\mathsf{B}\in\mathcal{B}(\mathsf{Y})$. Therefore, Φ is a T-chain. □

5 Proof of Theorem 2.1

The objective of this section is to prove Theorem 2.1. To do so, we investigate nonlinear state-space models associated to the recursion (2.18) using the theoretical tools presented in Section 4.

Since the normalization function R is not assumed to be smooth, we consider a transformed Markov chain—for which we can apply Theorem 2.4—valued in a smooth manifold and which can be transformed via a homeomorphism into the normalized Markov chain (2.18). In Section 5.1, we introduce the control model associated to this transformed process and verify the conditions **H**1, **H**2 reminded in Section 4.1.

Then, a last condition, **H**3 or **H**1,[6] is proven in two steps: Section 5.2 proves the existence of a steadily attracting state (defined in Section 4.1) and Section 5.3 shows that a required controllability condition is satisfied. We conclude the proof in Section 5.4, where the Markov chains associated to the different learning rate settings are analyzed, based on Theorem 2.3.

5.1 *Definition of normalized chains underlying CMA-ES following* (2.25) *and satisfying* **H**1-**H**2

In order to apply Theorem 2.2 to the normalized CMA-ES Markov chain defined via (2.17), we require the state space $\mathsf{Y}=\mathbb{R}^{3d}\times R^{-1}(\{1\})\times\mathbb{R}_{++}$ to be a smooth connected manifold. As mentioned in the previous section, this is not necessarily true unless we assume that the normalization R is continuously differentiable. Hence, we introduce a homeomorphic transformation of the normalized chain which lives on a smooth manifold. Consider $\rho\colon \mathcal{S}^d_{++}\to\mathbb{R}_{++}$ a map satisfying

***ρ*1.** *the function ρ is (positively) homogeneous and $\rho(\mathbf{I}_d)=1$,*

***ρ*2.** *the function ρ is smooth ($\mathcal{C}^\infty$) on $\mathcal{S}^d_{++}$.*

We keep this smooth normalization function abstract for the moment and will take it equal to $\rho(\cdot)=\det(\cdot)^{1/d}$ for proving Theorem 2.1. Define now

$$\begin{aligned}\xi\colon\ & \mathsf{Y}\to\mathsf{X}\\ & (z,p,q,\boldsymbol{\Sigma},r)\mapsto(z,p,q,\rho(\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma},r)\end{aligned}\tag{2.36}$$

where $\mathsf{X}=\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\rho^{-1}(\{1\})\times\mathbb{R}_{++}$. Then, as stated in the next proposition, X defines a smooth connected manifold.

Proposition 2.6. Suppose that the map $\rho\colon\mathcal{S}^d_{++}\to\mathbb{R}_{++}$ satisfies ***ρ*1**-***ρ*2**. Then, the set $\mathsf{X}=\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\rho^{-1}(\{1\})\times\mathbb{R}_{++}$ is a smooth connected manifold of dimension $3d+d(d+1)/2$.

[6] **H**3 for case (i) and **H**1 for cases (ii), (iii), (iv) of Theorem 2.1.

Proof. First note that the set $\mathsf{M} \coloneqq \mathbb{R}^{3d} \times \mathcal{S}_{++}^d \times \mathbb{R}_{++}$ is an open subset of the Euclidean space $\mathbb{R}^{3d} \times \mathcal{S}^d \times \mathbb{R}$, hence is a smooth submanifold of dimension $3d + d(d+1)/2 + 1$. Moreover, $\mathsf{N} \coloneqq \mathbb{R}$ is a smooth manifold of dimension 1. Define then the map

$$\begin{aligned} \bar{\varrho}\colon \ & \mathsf{M} \to \mathsf{N} \\ & (z, p, q, \boldsymbol{\Sigma}, r) \mapsto \rho(\boldsymbol{\Sigma}). \end{aligned}$$

Then, by $\boldsymbol{\rho}2$, $\bar{\varrho}$ is smooth. Moreover, it is a submersion at every point of M. Indeed, let $(z, p, q, \boldsymbol{\Sigma}, r) \in \mathsf{M}$, and let $\varepsilon \in (-1, 1)$. Then,

$$\bar{\varrho}\left((z, p, q, \boldsymbol{\Sigma}, r) + (0, 0, 0, \varepsilon\boldsymbol{\Sigma}, 0)\right) = \rho((1+\varepsilon)\boldsymbol{\Sigma}) = \rho(\boldsymbol{\Sigma}) + \varepsilon\rho(\boldsymbol{\Sigma})$$

by $\boldsymbol{\rho}1$. Therefore, by Taylor expansion and since the derivative $\mathcal{D}\bar{\varrho}(z, p, q, \boldsymbol{\Sigma}, r)$ is linear, we have

$$\mathcal{D}\bar{\varrho}(z, p, q, \boldsymbol{\Sigma}, r)(0, 0, 0, \kappa\boldsymbol{\Sigma}, 0) = \kappa\rho(\boldsymbol{\Sigma}),$$

for every $\kappa \in \mathbb{R}$, with $\rho(\boldsymbol{\Sigma}) > 0$. Hence, $\mathcal{D}\bar{\varrho}(z, p, q, \boldsymbol{\Sigma}, r)\colon \mathbb{R}^{3d} \times \mathcal{S}^d \times \mathbb{R} \to \mathbb{R}$ is surjective and thus $\bar{\varrho}$ is a submersion. Therefore, by the submersion level set theorem [100, Corollary 5.14], $\mathsf{X} = \bar{\varrho}^{-1}(\{1\})$ is a smooth manifold of dimension $3d + d(d+1)/2 + 1 - 1 = 3d + d(d+1)/2$. Let us prove now that X is connected. Since $\mathbb{R}^{3d} \times \mathbb{R}_{++}$ is connected, it is sufficient to prove that the manifold $\rho^{-1}(\{1\})$ is connected, and thus sufficient to prove that $\rho^{-1}(\{1\})$ is path-connected [100, Proposition 1.11]. Let $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1 \in \rho^{-1}(\{1\})$. Since $\mathcal{S}_{++}^d$ is connected, there exists a continuous path $\gamma\colon [0,1] \to \mathcal{S}_{++}^d$ with $\gamma(0) = \boldsymbol{\Sigma}_0$ and $\gamma(1) = \boldsymbol{\Sigma}_1$. Define then the path $\hat{\gamma}\colon [0,1] \to \rho^{-1}(\{1\})$ by $\hat{\gamma}(t) = \gamma(t)/\rho(\gamma(t))$ for $t \in [0,1]$. Since γ and ρ are continuous, then $\hat{\gamma}$ is continuous. Besides, $\hat{\gamma}(0) = \boldsymbol{\Sigma}_0/\rho(\boldsymbol{\Sigma}_0) = \boldsymbol{\Sigma}_0$ and $\hat{\gamma}(1) = \boldsymbol{\Sigma}_1/\rho(\boldsymbol{\Sigma}_1) = \boldsymbol{\Sigma}_1$, ending the proof. □

Moreover, the map ξ defined in (2.36) is a homeomorphism as stated below.

Proposition 2.7. Suppose that R and ρ are both continuous and satisfy $\mathbf{R}1$ and $\boldsymbol{\rho}1$. Then, the map ξ defined in (2.36) between the sets Y and X is a homeomorphism and

$$\begin{aligned} \xi^{-1}\colon \ & \mathsf{X} \to \mathsf{Y} \\ & (z, p, q, \hat{\boldsymbol{\Sigma}}, r) \mapsto \left(z, p, q, R(\hat{\boldsymbol{\Sigma}})^{-1}\hat{\boldsymbol{\Sigma}}, r\right). \end{aligned} \tag{2.37}$$

Proof. We can easily verify that the expression of the reciprocal function of ξ is (2.37). Then ξ^{-1} (resp. ξ) is continuous since R (resp. ρ) is continuous and takes value in $\mathbb{R}_{++}$. □

We formalize in the next lemma the update equations for $\{\theta_t\}_{t\in\mathbb{N}} = \{\xi(\phi_t)\}_{t\in\mathbb{N}}$.

Lemma 2.3. Suppose that the normalization function R satisfies $\mathbf{R}1$ and let $\Phi = \{\phi_t\}_{t\in\mathbb{N}}$ be the Markov chain defined via (2.18). Let ρ be a normalization function satisfying $\boldsymbol{\rho}1$ and let ξ be the homeomorphism defined in (2.36). Then the Markov chain $\Theta = \{\theta_t\}_{t\in\mathbb{N}} = \{\xi(\phi_t)\}_{t\in\mathbb{N}}$

satisfies

$$
\begin{aligned}
z_{t+1} &= \frac{F^{\mathrm{m}}_{c_m}(z_t, \sqrt{R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t}\mathbf{w}_m^\top U_{t+1}^{s_{t+1}})}{\sqrt{r_{t+1}\Gamma(p_{t+1})}} \\
p_{t+1} &= F^{\mathrm{p}}_{c_\sigma}(p_t, \mathbf{w}_m^\top U_{t+1}^{s_{t+1}}) \\
q_{t+1} &= F^{\mathrm{p}}_{c_c}(r_t^{-1/2} q_t, \sqrt{R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t}\mathbf{w}_m^\top U_{t+1}^{s_{t+1}}) \\
\hat{\boldsymbol{\Sigma}}_{t+1} &= \frac{F^{\mathrm{C}}_{c_1,c_\mu}\left(R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t, q_{t+1}, R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\sqrt{\hat{\boldsymbol{\Sigma}}_t}\sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\hat{\boldsymbol{\Sigma}}_t}\right)}{\rho \circ F^{\mathrm{C}}_{c_1,c_\mu}\left(R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t, q_{t+1}, R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\sqrt{\hat{\boldsymbol{\Sigma}}_t}\sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\hat{\boldsymbol{\Sigma}}_t}\right)} \\
r_{t+1} &= R \circ F^{\mathrm{C}}_{c_1,c_\mu}\left(R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t, q_{t+1}, R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\sqrt{\hat{\boldsymbol{\Sigma}}_t}\sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\hat{\boldsymbol{\Sigma}}_t}\right)
\end{aligned}
\tag{2.38}
$$

where $\{U_{t+1}\}_{t\in\mathbb{N}}$ is an i.i.d. process independent of $\theta_0 = (z_0, p_0, q_0, \hat{\boldsymbol{\Sigma}}_0, r_0) = \xi(\phi_0)$ distributed in the measured space $\mathsf{U} = (\mathbb{R}^d)^\lambda$ with $U_1 \sim \nu_U^d$, and s_{t+1} is a permutation of $\mathfrak{S}_\lambda$ that sorts the f-values of $x^* + z_t + \sqrt{R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t}U_{t+1}^i$, for $i = 1, \ldots, \lambda$.

Proof. Since for $t \in \mathbb{N}$, according to (2.37), we have that $\boldsymbol{\Sigma}_t = R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t$, the update equations for p_{t+1}, q_{t+1} ,r_{t+1} and z_{t+1} in (2.38) are deduced directly from Proposition 2.2 where we replace $\boldsymbol{\Sigma}_t$ by $R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t$. Moreover, we have, by Proposition 2.2 and using the definition of $\tilde{\boldsymbol{\Sigma}}_{t+1}$ in (2.19)

$$
\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{t+1} = \frac{\boldsymbol{\Sigma}_{t+1}}{\rho(\boldsymbol{\Sigma}_{t+1})} &= \frac{\tilde{\boldsymbol{\Sigma}}_{t+1}}{R(\tilde{\boldsymbol{\Sigma}}_{t+1})} \times \frac{R(\tilde{\boldsymbol{\Sigma}}_{t+1})}{\rho(\tilde{\boldsymbol{\Sigma}}_{t+1})} \\
&= \frac{F^{\mathrm{C}}_{c_1,c_\mu}\left(\boldsymbol{\Sigma}_t, q_{t+1}, \sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_t}\right)}{\rho \circ F^{\mathrm{C}}_{c_1,c_\mu}\left(\boldsymbol{\Sigma}_t, q_{t+1}, \sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_t}\right)} .
\end{aligned}
$$

The proof ends by replacing $\boldsymbol{\Sigma}_t$ by $R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t$. □

Using Theorem 2.4, we can transfer the irreducibility, aperiodicity and the T-chain property from the Markov chain Θ to the original normalized chain Φ we are interested in. Our objective from now on is to prove that the Markov chain Θ is an irreducible aperiodic T-chain. Our strategy for that is to apply Theorem 2.2 and verify that the required assumptions are satisfied. We first prove that Θ follows a deterministic control model of the form (2.25), described in Section 4.1.

Consider the smooth manifold $\mathsf{X} = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \rho^{-1}(\{1\}) \times \mathbb{R}_{++}$ (see Proposition 2.6) that defines the state space of the Markov chain Θ and let $\mathsf{V} := \mathbb{R}^{d\mu}$. We define

$$
\begin{aligned}
\alpha_\Theta \colon \mathsf{X} \times \mathsf{U} &\to \mathsf{V} \\
((z, p, q, \hat{\boldsymbol{\Sigma}}, r), (u^1, \ldots, u^\lambda)) &\mapsto \left[\sqrt{\frac{\hat{\boldsymbol{\Sigma}}}{R(\hat{\boldsymbol{\Sigma}})}}\, u^{s^u_{f(x^*+z+\sqrt{R(\hat{\boldsymbol{\Sigma}})^{-1}\hat{\boldsymbol{\Sigma}}}\cdot)}}\right]^{(i)}_{i=1,\ldots,\mu}
\end{aligned}
\tag{2.39}
$$

where given $g \colon \mathsf{A} \to \mathbb{R}$ a function and $v \in \mathsf{A}^\lambda$ we have used the notation s_g^v for a permutation that sorts increasingly the $g(v^i)$, $i = 1, \ldots, \lambda$. Consider the $(z^+, p^+, q^+, \hat{\boldsymbol{\Sigma}}^+, r^+)$ the update of

$\theta = (z, p, q, \hat{\mathbf{\Sigma}}, r)$ given the random input equals $v = \alpha_\Theta(\theta, u)$ that is

$$z^+ = \frac{z + c_m \mathbf{w}_m^\top v}{\sqrt{r^+}\Gamma(p^+)} \tag{2.40}$$

$$p^+ = (1 - c_\sigma)p + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}}} R(\hat{\mathbf{\Sigma}})^{1/2} \hat{\mathbf{\Sigma}}^{-1/2} \mathbf{w}_m^\top v \tag{2.41}$$

$$q^+ = r^{-1/2}(1 - c_c)q + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}} \mathbf{w}_m^\top v \tag{2.42}$$

$$\hat{\mathbf{\Sigma}}^+ = \frac{(1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}} + c_1 q^+ (q^+)^\top + c_\mu \sum_{i=1}^\mu w_i^c v_i v_i^\top}{\rho\left((1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}} + c_1 q^+ (q^+)^\top + c_\mu \sum_{i=1}^\mu w_i^c v_i v_i^\top\right)} \tag{2.43}$$

$$r^+ = R\left((1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}} + c_1 q^+ (q^+)^\top + c_\mu \sum_{i=1}^\mu w_i^c v_i v_i^\top\right). \tag{2.44}$$

This update defines a function $F_\Theta \colon \mathsf{X} \times \mathsf{V} \to \mathsf{X}$ such that

$$(z^+, p^+, q^+, \hat{\mathbf{\Sigma}}^+, r^+) = F_\Theta((z, p, q, \hat{\mathbf{\Sigma}}, r), \alpha_\Theta(\theta, u))$$

that can be expressed as

$$F_\Theta((z, p, q, \hat{\mathbf{\Sigma}}, r), (v^1, \ldots, v^\mu)) = \Big(F_z(z, p, q, R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}}, r; v), F_p(p, R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}}; v), F_q(q, r; v), F_{\mathbf{\Sigma}}(q, R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}}, r; v), F_r(q, R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}}, r; v)\Big)^\top \tag{2.45}$$

for $(z, p, q, \hat{\mathbf{\Sigma}}, r) \in \mathsf{X}$ and $(v^1, \ldots, v^\mu) \in \mathsf{V}$, and where F_z, F_p, F_q, $F_{\mathbf{\Sigma}}$ and F_r are defined as follows

$$F_z(z, p, q, \mathbf{\Sigma}, r; v) = F_r(q, \mathbf{\Sigma}, r; v)^{-1/2}\Gamma \circ F_p(p, \mathbf{\Sigma}; v)^{-1} F_{c_m}^{\text{m}}(z, \mathbf{w}_m^\top v) \tag{2.46}$$

$$F_p(p, \mathbf{\Sigma}; v) = F_{c_\sigma}^{\text{p}}(p, \mathbf{\Sigma}^{-1/2}\mathbf{w}_m^\top v) \tag{2.47}$$

$$F_q(q, r; v) = F_{c_c}^{\text{p}}(r^{-1/2}q, \mathbf{w}_m^\top v) \tag{2.48}$$

$$F_{\mathbf{\Sigma}}(q, \mathbf{\Sigma}, r; v) = \frac{F_{c_1, c_\mu}^{\text{C}}\left(\mathbf{\Sigma}, F_q(q, r; v), \sum_{i=1}^\mu w_i^c v_i v_i^\top\right)}{\rho \circ F_{c_1, c_\mu}^{\text{C}}\left(\mathbf{\Sigma}, F_q(q, r; v), \sum_{i=1}^\mu w_i^c v_i v_i^\top\right)} \tag{2.49}$$

$$F_r(q, \mathbf{\Sigma}, r; v) = R \circ F_{c_1, c_\mu}^{\text{C}}\left(\mathbf{\Sigma}, F_q(q, \mathbf{\Sigma}, r; v), \sum_{i=1}^\mu w_i^c v_i v_i^\top\right). \tag{2.50}$$

Then, as stated in the next proposition, Θ follows the model described in Section 4.1 with the functions F_Θ and α_Θ defined above.

Proposition 2.8. Suppose that the normalization functions R satisfies **R**1 and ρ satisfies $\boldsymbol{\rho}$1. Then, the Markov chain $\Theta = \{(z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t)\}_{t \in \mathbb{N}}$ defined by (2.38) satisfies

$$\theta_{t+1} = F_\Theta\left(\theta_t, \alpha_\Theta(\theta_t, U_{t+1})\right) \tag{2.51}$$

where F_Θ is defined in (2.45) and α_Θ in (2.39).

Proof. Straightforward by Lemma 2.3. □

Before to prove that the control model (2.51) associated to the Markov chain Θ satisfies the assumptions **H**1 and **H**2, we characterize in the next lemma the density of the random variable

$\alpha_\Theta(\theta, U)$ for $\theta \in \mathsf{X}$ and $U \sim (\nu_U^d)^{\otimes\lambda}$ assuming that the objective function is the composite of a strictly increasing function with a function with negligible level sets and the distribution ν_U^d is admits a density positive everywhere with respect to the Lebesgue measure. The latter assumption could be relaxed with more work, but for the purposes of this paper we only consider positive densities.

Lemma 2.4. Suppose that the objective function f satisfies **F**1 and that the probability distribution ν_U^d satisfies **N**1. Define, for any $\theta = (z, p, q, \hat{\Sigma}, r) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{S}_{++}^d \times \mathbb{R}_{++}$ and $v = (v_1, \dots, v_\mu) \in \mathbb{R}^{d\mu}$,

$$p_{z,\mathbf{\Sigma}}(v) = \frac{\lambda!}{(\lambda-\mu)!} \mathbb{1}\{f_*(z+\sqrt{\mathbf{\Sigma}}v_1) < \cdots < f_*(z+\sqrt{\mathbf{\Sigma}}v_\mu)\}(1-Q_{z,\mathbf{\Sigma}}^{f_*}(v_\mu))^{\lambda-\mu} p_U^d(v_1)\dots p_U^d(v_\mu) \tag{2.52}$$

with $\mathbf{\Sigma} = \hat{\mathbf{\Sigma}}/R(\hat{\mathbf{\Sigma}})$ where R is the normalization function used in (2.39), $Q_{z,\mathbf{\Sigma}}^{f_*}(u) = \int \mathbb{1}\{f_*(z+\sqrt{\mathbf{\Sigma}}\xi) < f_*(z+\sqrt{\mathbf{\Sigma}}u)\}\nu_U^d(\mathrm{d}\xi)$ for $u \in \mathbb{R}^d$, and $f_* = f(\cdot + x^*)$. Then, $p_{z,\mathbf{\Sigma}}$ defines a density (with respect to Lebesgue in $\mathbb{R}^{d\mu}$) of the random variable $\mathbf{\Sigma}^{-1/2}\alpha_\Theta(\theta, U)$, where $U \sim (\nu_U^d)^{\otimes\lambda}$ such that the density of $\alpha_\Theta(\theta, U)$ equals

$$v \mapsto \det \mathbf{\Sigma}^{-1/2} p_{z,\mathbf{\Sigma}}(\mathbf{\Sigma}^{-1/2}v) = \frac{1}{\sqrt{\det \mathbf{\Sigma}}} p_{z,\mathbf{\Sigma}}(\mathbf{\Sigma}^{-1/2}v) \ . \tag{2.53}$$

Besides, when R is continuous, the function $((z, p, q, \hat{\Sigma}, r), v) \in \mathbb{R}^{3d} \times \mathcal{S}_{++}^d \times \mathbb{R}_{++} \times \mathbb{R}^{d\mu} \mapsto p_{z,\hat{\mathbf{\Sigma}}/R(\hat{\mathbf{\Sigma}})}(v)$ is lower semicontinuous and thus the density function (2.53) is lower semicontinuous as well.

Proof. Let $U^1, \dots, U^\lambda$ be independent random vectors identically distributed under the probability distribution ν_U^d, and denote $U = (U^1, \dots, U^\lambda)$. Let $\theta = (z, p, q, \hat{\Sigma}, r) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{S}_{++}^d \times \mathbb{R}_{++}$. Since the objective function f satisfies **F**1, then the random vector $V = \mathbf{\Sigma}^{-1/2}\alpha_\Theta(\theta, U)$ satisfies almost surely

$$V = \sum_{\sigma\in\mathfrak{S}_\lambda} \mathbb{1}\left\{f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(1)}\right) < \cdots < f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(\lambda)}\right)\right\} \times \left(U^{\sigma(1)}, \dots, U^{\sigma(\mu)}\right).$$

where $\mathfrak{S}_\lambda$ is the set of permutations of $\{1, \dots, \lambda\}$. Hence, by symmetry,

$$\begin{aligned} V = \frac{1}{(\lambda-\mu)!} \sum_{\sigma\in\mathfrak{S}_\lambda} & \mathbb{1}\left\{f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(1)}\right) < \cdots < f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(\mu)}\right)\right\} \\ & \times \prod_{k=\mu+1}^{\lambda} \mathbb{1}\left\{f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(\mu)}\right) < f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(k)}\right)\right\} \times \left(U^{\sigma(1)}, \dots, U^{\sigma(\mu)}\right). \end{aligned}$$

Let $\eta\colon \mathbb{R}^{d\mu} \to \mathbb{R}_+$ be a smooth map with compact support. We have

$$\begin{aligned} \mathbb{E}\left[\eta(V)\right] = \frac{1}{(\lambda-\mu)!} \sum_{\sigma\in\mathfrak{S}_\lambda} \int & \mathbb{1}\left\{f_*\left(z+\sqrt{\mathbf{\Sigma}}u_{\sigma(1)}\right) < \cdots < f_*\left(z+\sqrt{\mathbf{\Sigma}}u_{\sigma(\mu)}\right)\right\} \\ & \times \prod_{k=\mu+1}^{\lambda} \mathbb{1}\left\{f_*\left(z+\sqrt{\mathbf{\Sigma}}u_{\sigma(\mu)}\right) < f_*\left(z+\sqrt{\mathbf{\Sigma}}u_{\sigma(k)}\right)\right\} \\ & \times \eta\left(u_{\sigma(1)}, \dots, u_{\sigma(\mu)}\right) p_U^d(u_1)\dots p_U^d(u_\lambda)\mathrm{d}u_1\dots\mathrm{d}u_\lambda. \end{aligned}$$

However, observe that, for each $k = \mu + 1, \dots, \lambda$, we have

$$\int \mathbb{1}\left\{f_*\left(z + \sqrt{\Sigma}u_{\sigma(\mu)}\right) < f_*\left(z + \sqrt{\Sigma}u_{\sigma(k)}\right)\right\} p_U^d(u_{\sigma(k)})\mathrm{d}u_{\sigma(k)} = 1 - Q_{z,\Sigma}^{f_*}\left(u_{\sigma(\mu)}\right).$$

We deduce the desired result. Since the composition of lower semicontinuous functions is lower semicontinuous and since f is continuous, when R is continuous, the function $((z, p, q, \hat{\Sigma}, r), v) \mapsto p_{z,\hat{\Sigma}/R(\hat{\Sigma})}(v)$ is lower semicontinuous. □

Furthermore, under assumptions detailed in Section 2.2, we verify that **H**1 and **H**2 hold.

Proposition 2.9. Suppose that the objective function f satisfies **F**1-**F**2, that the normalization function R satisfies **R**1-**R**2, and that the stepsize change Γ is such that **Γ**1 hold. Suppose moreover that ρ satisfies ***ρ***1-***ρ***2. Consider the Markov chain $\Theta = \{(z_t, p_t, q_t, \hat{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ defined by (2.38). Define the functions F_Θ and α_Θ via (2.45) and (2.39) respectively. Then, Θ follows (2.51), and **H**1-**H**2 hold.

Proof. By Lemma 2.4, we find that **H**1 holds , with ζ_{V} being the Lebesgue measure on $\mathsf{V} = \mathbb{R}^{d\mu}$. Furthermore, using **R**2, ***ρ***2 and **Γ**1, we deduce, by composition, that **H**2 is satisfied. □

5.2 *Finding steadily attracting states*

In this section and in Section 5.3, we prove that the control model (2.51) satisfies condition **H**3.[7] This is required to apply Theorem 2.2[8] and find that the Markov chain Θ obeying to (2.38) is an irreducible aperiodic T-chain. In this section, we focus on the existence of steadily attracting states. This is formalized in the next proposition.

Proposition 2.10. Suppose that the objective function f satisfies **F**1-**F**2, that the stepsize change satisfies **Γ**1-**Γ**2, that the normalization functions R and ρ satisfy **R**1-**R**2 and ***ρ***1-***ρ***2 respectively, and that the sampling distribution ν_U^d is such that **N**1 holds.
Then $(0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ is a steadily attracting state for the control model (2.51) with the functions F_Θ and α_Θ given by (2.45) and (2.39), respectively.

Proof. Let $\theta_0 \in \mathsf{X}$. By Lemma 2.5, we find v_1 such that $S_{\theta_0}^1(v_1) = (0, p_1, q_1, \hat{\Sigma}_1, r_1)$. If $q_1 \neq 0$, by Lemma 2.7, we set v_2, v_3 such that $S_{\theta_0}^3(v_{1:3}) = (0, p_3, 0, \hat{\Sigma}_3, r_3)$. Using Lemma 2.8 we reach via a $4(d-1)$ steps a state $\theta = (0, \cdot, 0, \mathbf{I}_d, \cdot)$. Using Lemma 2.9, we complete the path v_1 in case $q_1 = 0$ or v_1, v_2, v_3 otherwise into $v_{1:\infty} = (v_1, v_2, \dots) \in \overline{\mathcal{O}_{\theta_0}^\infty}$ such that $\lim_{k\to\infty} S_{\theta_0}^k(v_{1:k}) = (0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$. This implies [52, Corollary 4.5] that $(0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ is a steadily attractive state. □

The proof of Proposition 2.10 relies on Lemmas 2.5 and 2.7 to 2.9 below. First, we state the next proposition, which is useful to provide candidates for the paths between an initial state and the steadily attracting state given by Proposition 2.10.

[7] Or condition **H**1 if we assume no cumulation on the stepsize or the covariance matrix.
[8] Or Theorem 2.3.

Proposition 2.11. In the context of Proposition 2.10, let $\theta_0 = (z_0, p_0, q_0, \hat{\Sigma}_0, r_0) \in \mathsf{X}$, let $k \geqslant 1$ and $v_{1:k} = (v_1, \dots, v_k) \in \mathsf{V}^k$ be such that for $i = 1, \dots, k$, we have $v_i = [v_i^1, \dots, v_i^\mu] \in \mathbb{R}^{d\mu}$ with $v_i^1 = \cdots = v_i^\mu \in \mathbb{R}^d$. Then, $v_{1:k} \in \overline{\mathcal{O}_{\theta_0}^k}$.

Proof. We prove here that $v_1 = [\bar{v}_1, \dots, \bar{v}_1] \in \overline{\mathcal{O}_{\theta_0}^1}$. By Lemma 2.4, it is sufficient to prove that there exists a sequence $\{w_n = [w_n^1, \dots, w_n^\mu] \in \mathbb{R}^{d\mu}\}_{n\in\mathbb{N}}$ which converges to v_1 such that $p_{z_0,\mathbf{\Sigma}_0}(\mathbf{\Sigma}_0^{-1/2} w_n) > 0$ for all $n \in \mathbb{N}$, where $\mathbf{\Sigma}_0 = R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0$ and $p_{z_0,\mathbf{\Sigma}_0}$ is the density defined via (2.52). Moreover, by **N**1 and by definition of $p_{z_0,\mathbf{\Sigma}_0}$, it is sufficient to prove that for every $n \in \mathbb{N}$, $f(z_0 + w_n^1) < \cdots < f(z_0 + w_n^\mu)$. Furthermore, by **F**1, for every $n \in \mathbb{N}$, there exists $z_n^1, \dots, z_n^\mu \in \mathsf{B}(z_0 + \bar{v}_1, 1/n)$ such that $f(z_n^1) < \cdots < f(z_n^\mu)$. We take $w_n^i = z_n^i - z_0$ for $i = 1, \dots, \mu$ and $n \in \mathbb{N}$. Then w_n converges to v_1 and belongs to $\mathcal{O}_{\theta_0}^1$, so that $v_1 \in \overline{\mathcal{O}_{\theta_0}^1}$. Similarly, $v_2 \in \overline{\mathcal{O}_{S_{\theta_0}^1(x_1)}^1}$ for all x_1 and using the continuity of $v \mapsto S_{\theta_0}^1(v)$ in v_1, we deduce that $v_{1:2} \in \overline{\mathcal{O}_{\theta_0}^2} = \overline{\{(x_1, x_2) | p_{\theta_0}^1(x_1) \times p_{S_{\theta_0}^1(x_1)}(x_2) > 0\}}$. Similarly we obtain that $v_{1:k} \in \overline{\mathcal{O}_{\theta_0}^k}$. □

The following lemma is the first step to build a path between an arbitrary initial state $\theta_0 \in \mathsf{X}$ to the steadily attracting state $\theta^* = (0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ given in Proposition 2.10. More precisely, it shows that from $\theta_0 \in \mathsf{X}$, we can reach via a one-step path a state θ_1 such that $z_1 = 0$.

Lemma 2.5. In the context of Proposition 2.10, let $\theta_0 = (z_0, p_0, q_0, \hat{\Sigma}_0, r_0) \in \mathsf{X}$. Then there exists $\theta_1 = (0, p_1, q_1, \hat{\Sigma}_1, r_1) \in \mathsf{X}$ and $v_1 \in \overline{\mathcal{O}_{\theta_0}^1}$ such that $S_{\theta_0}^1(v_1) = \theta_1$. Moreover, we can choose v_1 as a function of z_0 such that v_1 goes to 0 when z_0 tends to 0.

Proof. Let $v_1 = -c_m^{-1} \times [z_0, \dots, z_0] \in (\mathbb{R}^d)^\mu$. It belongs to $\overline{\mathcal{O}_{\theta_0}^1}$ by Proposition 2.11. Set $\theta_1 = (z_1, p_1, q_1, \hat{\Sigma}_1, r_1) = S_{\theta_0}^1(v_1)$. Then, $z_1 = F_z(z_0, p_0, q_0, \hat{\Sigma}_0/R(\hat{\Sigma}_0), r_0; v_1) = r_1^{-1/2}\Gamma(p_1)^{-1} \times (z_0 - c_m \times c_m^{-1}\mathbf{w}_m^\top [z_0, \dots, z_0]) = 0$, see (2.40). We have used in particular that $\sum w_i^m = 1$. □

We make the following observation when the mean z_0 is in 0: by performing one step via $v_1 = (u_1, \dots, u_1)$ for any u_1 in $\mathbb{R}^d$, we can find a zero mean again in two steps by choosing a path $v_{1:2}$ appropriately.

Lemma 2.6. In the context of Proposition 2.10, let $\theta_0 = (0, p_0, q_0, \hat{\Sigma}_0, r_0) \in \mathsf{X}$. Then, given $v_1 = (u_1, \dots, u_1) \in \overline{\mathcal{O}_{\theta_0}^1}$ for some $u_1 \in \mathbb{R}^d$, and by defining $\theta_1 = (z_1, p_1, q_1, \hat{\Sigma}_1, r_1) = S_{\theta_0}^1(v_1)$ and $v_2 = -r_1^{-1/2}\Gamma(p_1)^{-1}v_1$, we have that $v_{1:2} = [v_1, v_2] \in \overline{\mathcal{O}_{\theta_0}^2}$ and $\theta_2 = (z_2, p_2, q_2, \hat{\Sigma}_2, r_2) = S_{\theta_0}^2(v_{1:2})$ satisfies $z_2 = 0$.

Proof. By Proposition 2.11, we have $v_1 \in \overline{\mathcal{O}_{\theta_0}^1}$ and $v_{1:2} = [v_1, v_2] \in \overline{\mathcal{O}_{\theta_0}^2}$. Moreover, we have

$$z_2 = r_2^{-1/2}\Gamma(p_2)^{-1} \times \left(r_1^{-1/2}\Gamma(p_1)^{-1} \times (0 + c_m u_1) - c_m r_1^{-1/2}\Gamma(p_1)^{-1}u_1\right) = 0$$

ending the proof. □

Next, from any initial state $\theta_0 \in \mathsf{X}$ with $z_0 = 0$, we reach via a two-steps path a state $\theta_2 \in \mathsf{X}$ with $z_2 = 0$ and $q_2 = 0$.

Lemma 2.7. In the context of Proposition 2.10, let $\theta_0 = (0, p_0, q_0, \hat{\Sigma}_0, r_0) \in \mathsf{X}$ such that $q_0 \neq 0$. Then, there exist $\theta_2 = (0, p_2, 0, \hat{\Sigma}_2, r_2) \in \mathsf{X}$ and $v_{1:2} \in \overline{\mathcal{O}^2_{\theta_0}}$ such that $S^2_{\theta_0}(v_{1:2}) = \theta_2$. Moreover, we can choose $v_{1:2}$ such that $v_{1:2} \to 0$ when q_0 tends to 0.

Proof.
Let $u_1 \in \mathbb{R}^d$ and set $v_1 = (u_1, \dots, u_1)$. It belongs to $\overline{\mathcal{O}^1_{\theta_0}}$ by Proposition 2.11. Then, define

$$\theta_1 = (z_1, p_1, q_1, \hat{\Sigma}_1, r_1) = F_\Theta\left(\theta_0, \alpha_\Theta\left(\theta_0, v_1\right)\right) = S^1_{\theta_0}(v_1) \ .$$

Then, define $v_2 = -r_1^{-1/2}\Gamma(p_1)^{-1}v_1 \in \overline{\mathcal{O}^1_{\theta_1}}$, and

$$\theta_2 = (z_2, p_2, q_2, \hat{\Sigma}_2, r_2) = F_\Theta\left(\theta_1, \alpha_\Theta\left(\theta_1, v_2\right)\right) = S^2_{\theta_0}(v_{1:2}) \ .$$

Then, by Lemma 2.6, $v_{1:2} = (v_1, v_2) \in \overline{\mathcal{O}^2_{\theta_0}}$ and $z_2 = 0$. Moreover,

$$\begin{aligned} q_2 &= (1-c_c)^2 r_0^{-1/2} r_1^{-1/2} q_0 + (1-c_c) r_1^{-1/2} \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\, u_1 \\ &\quad - r_1^{-1/2}\Gamma(p_1)^{-1}\sqrt{c_c(2-c_c)\mu_{\text{eff}}}\, u_1 \\ &= r_1^{-1/2} \times \left[(1-c_c)^2 (r_0^{-1/2} q_0 + \left(1 - c_c - \Gamma(p_1)^{-1}\right) \times \sqrt{c_c(2-c_c)\mu_{\text{eff}}} \times u_1\right] \ . \end{aligned}$$

Let $\kappa \in \mathbb{R}$, and choose $u_1 = \kappa q_0$. Since $v \mapsto S^2_{\theta_0}(v)$ is continuous, then both r_1 and q_2 depend continuously on κ. Moreover, we have

$$q_2 = r_1^{-1/2} \times \left[(1-c_c)^2 r_0^{-1/2} + \left(1 - c_c - \Gamma(p_1)^{-1}\right) \times \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\kappa\right] \times q_0 \ .$$

But, as $r_1 > 0$, and as $\Gamma(p_1)^{-1} = \Gamma\left((1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}} R(\hat{\Sigma}_0)}\kappa \hat{\Sigma}_0^{-1/2} q_0\right)^{-1}$ is less that $1 - c_c$ when $\kappa \to \pm\infty$ by $\mathbf{\Gamma}2$, then by the intermediate value theorem (since Γ is continuous by $\mathbf{\Gamma}1$), there exists $\kappa \in \mathbb{R}$ such that $q_2 = 0$. With the above choice of $u_1 = \kappa q_0$, $v_1 = (u_1, \dots, u_1)$ and $v_2 = -r_1^{-1/2}\Gamma(p_1)^{-1}v_1 \in \overline{\mathcal{O}^1_{\theta_1}}$, we see that $v_{1:2} \to 0$ when q_0 tends to 0. □

From an initial state θ_0 with $z_0 = q_0 = 0$, we reach via a $4(d-1)$-steps path a state $\theta_{4(d-1)}$ with $z_{4(d-1)} = q_{4(d-1)} = 0$, and $\hat{\Sigma}_{4(d-1)} = \mathbf{I}_d$. This is achieved by applying $(d-1)$ times the following lemma successively to the k-th largest (counted with multiplicity) eigenvalue of $\hat{\Sigma}_0$, for $k = 2, \dots, d$. For the sake of conciseness, the proof of Lemma 2.8 is delayed to Section A.

Lemma 2.8. In the context of Proposition 2.10, let $\theta_0 = (0, p_0, 0, \hat{\Sigma}_0, r_0)$. Consider an orthonormal basis $\mathcal{B}$ of $\mathbb{R}^d$ composed of eigenvectors of $\hat{\Sigma}_0$ such that the matrix $\hat{\Sigma}_0$ writes in the basis $\mathcal{B}$ as

$$[\hat{\Sigma}_0]_{\mathcal{B}} = \operatorname{diag}\left(\lambda_1, \dots, \lambda_d\right),$$

with $\lambda_1 = \lambda_2 = \cdots = \lambda_{k-1} \geqslant \lambda_k \geqslant \cdots \geqslant \lambda_d$ for some $2 \leqslant k \leqslant d$. Then, there exists $\gamma > 0$, such that the matrix $\hat{\Sigma}_4$ defined by

$$[\hat{\Sigma}_4]_{\mathcal{B}} = \gamma \times \operatorname{diag}\left(\lambda_1, \dots, \lambda_{k-1}, \lambda_{k-1}, \lambda_{k+1}, \dots, \lambda_d\right), \tag{2.54}$$

is such that for some $p_4 \in \mathbb{R}^d$ and $r_4 > 0$, and $v_{1:4} \in \overline{\mathcal{O}^4_{\theta_0}}$, we have $S^4_{\theta_0}(v_{1:4}) = \theta_4 = (0, p_4, 0, \hat{\mathbf{\Sigma}}_4, r_4)$.

Finally, as stated in the next lemma, from an initial state $\theta_0 \in \mathsf{X}$ such that $z_0 = q_0 = 0$ and $\hat{\mathbf{\Sigma}}_0 = \mathbf{I}_d$, we can reach any neighborhood of the state $\theta^* = (0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$.

Lemma 2.9. In the context of Proposition 2.10, let $\theta_0 = (0, p_0, 0, \mathbf{I}_d, r_0) \in \mathsf{X}$. Then there exists $v_{1:\infty} \in \overline{\mathcal{O}^\infty_{\theta_0}}$ such that $\lim S^t_{\theta_0}(v_{1:t}) = (0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ when $t \to \infty$.

Proof. Define $v_{1:\infty}$, by $v_t = 0 \in \mathbb{R}^{d\times\mu}$ for all $t \geqslant 1$. By Proposition 2.11, we have $v_{1:\infty} \in \overline{\mathcal{O}^\infty_{\theta_0}}$. Denote

$$\theta_{t+1} = (z_{t+1}, p_{t+1}, q_{t+1}, \hat{\mathbf{\Sigma}}_{t+1}, r_{t+1}) = F_\Theta(\theta_t, \alpha_\Theta(\theta_t, v_{t+1})).$$

Since, $\theta_0 = (0, p_0, 0, \mathbf{I}_d, r_0)$ and $v_t = 0$, by induction, we have $\hat{\mathbf{\Sigma}}_{t+1} = \mathbf{I}_d$, $z_{t+1} = 0$, $q_{t+1} = 0$ $r_{t+1} = R((1 - c_1 - c_\mu)\mathbf{I}_d) = 1 - c_1 - c_\mu$ and $p_{t+1} = (1 - c_\sigma)p_t$. Since $0 \leqslant 1 - c_\sigma < 1$, then $(z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t)$ tends to $(0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ when $t \to \infty$, ending the proof. □

Lastly, as a consequence of Proposition 2.10, we prove that given any normalized covariance matrix $\hat{\mathbf{\Sigma}}^* \in \mathcal{S}^d_{++}$ such that $\rho(\hat{\mathbf{\Sigma}}^*) = 1$, we can find a value for the path $p^* \in \mathbb{R}^d$, for the variable $r^* > 0$, such that the state $\theta^* = (0, p^*, 0, \hat{\mathbf{\Sigma}}^*, r^*) \in \mathsf{X}$ with normalized mean and normalized path for the rank-one update equal to zero is steadily attracting. In Section 5.3, we use these steadily attracting states to prove the controllability condition stated in Proposition 2.12.

Corollary 2.3. Consider the context of Proposition 2.10. Let $\hat{\mathbf{\Sigma}}^* \in \mathcal{S}^d_{++}$ be such that $\rho(\hat{\mathbf{\Sigma}}^*) = 1$. Then, there exist $p^* \in \mathbb{R}^d$ and $r^* > 0$ such that $\theta^* = (0, p^*, 0, \hat{\mathbf{\Sigma}}^*, r^*) \in \mathsf{X}$ is a steadily attracting state.

Proof. By Proposition 2.10, we know that $\theta_0 = (0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ is a steadily attracting state. Hence, in order to prove that a state $\theta^* \in \mathsf{X}$ is steadily attracting, it is sufficient, as explained below, to prove

(i) that there exist $k \in \mathbb{N}$ and $v_{1:k} \in \overline{\mathcal{O}^k_{\theta_0}}$ such that $S^k_{\theta_0}(v_{1:k}) = \theta^*$.

Indeed, assume we have proven (i) and let V be a neighborhood of θ^* and let $\theta \in \mathsf{X}$. Then, by continuity of $w_{1:k} \mapsto S^k_{\theta_0}(w_{1:k})$ around $v_{1:k}$, there exists $v^*_{1:k} \in \mathcal{O}^k_{\theta_0}$ such that $S^k_{\theta_0}(v^*_{1:k}) \in V$. Since $x \mapsto p^k_x(v^*_{1:k})$ is lower semicontinuousand $x \mapsto S^k_x(v^*_{1:k})$ is continuous, then there exists a neighborhood U of θ_0 such that for every $x \in U$, $p^k_x(v^*_{1:k}) > 0$, i.e., $v^*_{1:k} \in \mathcal{O}^k_x$, and $S^k_x(v^*_{1:k}) \in V$. Moreover, since θ_0 is steadily attracting, then there exists $T > 0$ such that for every $t \geqslant T$, there exists $w_{1:t} \in \mathcal{O}^k_\theta$ such that $S^t_\theta(w_{1:t}) \in U$, hence $[w_{1:t}, v^*_{1:k}] \in \mathcal{O}^{t+k}_\theta$ and $S^{t+k}_\theta([w_{1:t}, v^*_{1:k}]) \in V$ and hence θ^* is a steadily attracting state.
Let $\hat{\mathbf{\Sigma}}^* \in \mathcal{S}^d_{++}$ be such that $\rho(\hat{\mathbf{\Sigma}}^*) = 1$. We proceed now as in Lemma 2.8 to prove (i) for a state θ^* that is equal to $(0, p^*, 0, \hat{\mathbf{\Sigma}}^*, r^*)$ for p^* and r^* constructed below. For $i = 1, \ldots, d$, let λ_i be the i-th largest eigenvalue of $\hat{\mathbf{\Sigma}}^*$ (counted with multiplicity), and $(e_1, \ldots, e_d)$ an orthonormal basis of eigenvectors of $\hat{\mathbf{\Sigma}}^*$ such that $\hat{\mathbf{\Sigma}}^* e_i = \lambda_i e_i$. Then, let κ and κ' be real numbers, and by Proposition 2.11, define $v_{1:4} \in \overline{\mathcal{O}^4_{\theta_0}}$ by

$$v_1 = \kappa[e_1, \ldots, e_1] \in \mathbb{R}^{d\mu},\ v_2 = -r_1^{-1/2}\Gamma(p_1)^{-1}v_1,\ v_3 = \kappa'[e_1, \ldots, e_1],\ v_4 = -r_3^{-1/2}\Gamma(p_3)^{-1}v_3,$$

where $\theta_t = (z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t) = S^t_{\theta_0}(v_{1:t})$ for $t = 1, 2, 3, 4$. Then, as in the proof of Lemma 2.8, there exist values of κ and κ' in $\mathbb{R}$ such that $z_4 = q_4 = 0$ and such that there exists $\rho_4 > 0$ with $\hat{\mathbf{\Sigma}}_4$ satisfying $\hat{\mathbf{\Sigma}}_4 e_1 = \rho_4(1 - c_1 - c_\mu)^{-4(d-1)}\lambda_1 e_1$ and $\hat{\mathbf{\Sigma}}_4 e_k = \rho_4(1 - c_1 - c_\mu)^4 e_k$ for $k = 2, \dots, d$.
Then, by repeating these steps with $e_2, \dots, e_d$ instead of e_1 and $\lambda_2, \dots, \lambda_d$ instead of λ_1, then there exist $v_{1:4d} \in \overline{\mathcal{O}^{4d}_{\theta_0}}$ and $\rho_{4d} > 0$ such that $\theta_{4d} = (z_{4d}, p_{4d}, q_{4d}, \hat{\mathbf{\Sigma}}_{4d}, r_{4d}) = S^{4d}_{\theta_0}(v_{1:4d})$ satisfies $z_{4d} = q_{4d} = 0$ and $\hat{\mathbf{\Sigma}}_{4d} e_k = \rho_{4d}\lambda_k e_k$ for $k = 1, \dots, d$. Hence $\hat{\mathbf{\Sigma}}_{4d} = \rho_{4d}\hat{\mathbf{\Sigma}}^*$. But since $\rho(\hat{\mathbf{\Sigma}}^*) = 1$ and $\rho(\hat{\mathbf{\Sigma}}_{4d}) = 1$, then $\rho_{4d} = 1$, i.e. $\hat{\mathbf{\Sigma}}_{4d} = \hat{\mathbf{\Sigma}}^*$ such that we have proven (i) for $\theta^* = (0, p_{4d}, 0, \hat{\mathbf{\Sigma}}^*, r_{4d})$ and in turn that $\theta^* = (0, p_{4d}, 0, \hat{\mathbf{\Sigma}}^*, r_{4d})$ is a steadily attracting state. □

5.3 *Controllability condition*

In the previous section, we prove that the control model (2.51) admits steadily attracting states. In the current section, we prove that a controllability condition, as required to satisfy the assumptions **H**3 or **H**1, is satisfied at a steadily attracting state. By combining Corollary 2.3 and the following Proposition 2.12, we prove **H**3 or **H**1. For a finite-dimensional vectorial space E equipped with a norm $\|\cdot\|$ and an element $h \in \mathsf{E}$, we use the notation $o(h)$, respectively $O(h)$, to be understood as $o(\|h\|)$, respectively $O(\|h\|)$. Besides, it does not depend on the chosen norm, since all norms on a finite-dimensional space induce the same topology.

Proposition 2.12. Suppose that the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$, the normalization functions R and ρ, the stepsize change $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$ and the sampling distribution ν^d_U satisfy **F**1-**F**2, **R**1-**R**3, ***ρ***1-***ρ***2, **Γ**1-**Γ**3 and **N**1, respectively.
Consider the control model (2.51) with the functions F_Θ and α_Θ defined by (2.45) and (2.39) respectively.
Then, there exist a steadily attracting state $\theta_0 \in \mathsf{X}$, $T > 0$ and $v_{1:T} \in \overline{\mathcal{O}^T_{\theta_0}}$ such that $S^T_{\theta_0}$ is differentiable at $v_{1:T}$, and, by denoting $(z_T, p_T, q_T, \hat{\mathbf{\Sigma}}_T, r_T) = S^T_{\theta_0}(v_{1:T})$, we have

(a) if $c_c \neq 1$, $c_\sigma \neq 1$, $1 - c_c \neq (1 - c_\sigma)\sqrt{1 - c_1 - c_\mu}$ and $c_\mu > 0$, then $\mathcal{D}S^T_{\theta_0}(v_{1:T})$ is of maximal rank;

(b) if $c_c = 1$ and $c_\sigma \neq 1$, then, for every $(z, p, \mathbf{\Sigma}) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\})$, there exist $q \in \mathbb{R}^d$ and $r \in \mathbb{R}$ such that $(z, p, q, \mathbf{\Sigma}, r) \in \operatorname{rge} \mathcal{D}S^T_{\theta_0}(v_{1:T})$;

(c) if $c_c \neq 1$, $c_\sigma = 1$ and $c_\mu > 0$, then, for every $(z, q, \mathbf{\Sigma}, r) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\}) \times \mathbb{R}$, there exists $p \in \mathbb{R}^d$ such that $(z, p, q, \mathbf{\Sigma}, r) \in \operatorname{rge} \mathcal{D}S^T_{\theta_0}(v_{1:T})$;

(d) if $c_c = c_\sigma = 1$, then, for every $(z, \mathbf{\Sigma}) \in \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\})$, there exist $(p, q) \in \mathbb{R}^d \times \mathbb{R}^d$ and $r \in \mathbb{R}$, such that $(z, p, q, \mathbf{\Sigma}, r) \in \operatorname{rge} \mathcal{D}S^T_{\theta_0}(v_{1:T})$.

Before proving Proposition 2.12, we state the two following lemmas, which characterize the derivatives of the normalization function ρ and of the transition map $S^1_{\theta_0}$, respectively.

Lemma 2.10. Consider a positively homogeneous function $R\colon \mathcal{S}^d_{++} \to \mathbb{R}_{++}$. Let $\mathbf{A} \in \mathcal{S}^d_{++}$ and $\gamma > 0$ and suppose that R is differentiable at $\mathbf{A}$. Then, R is differentiable at $\gamma\mathbf{A}$ and $\mathcal{D}R(\gamma\mathbf{A}) = \mathcal{D}R(\mathbf{A})$.

Proof. By Taylor expansion, we have, when $\mathbf{H} \in \mathcal{S}^d$ tends to 0, that

$$\begin{aligned} R(\gamma\mathbf{A} + \mathbf{H}) &= \gamma \times R(\mathbf{A} + \gamma^{-1}\mathbf{H}) = \gamma R(\mathbf{A}) + \gamma \mathcal{D}R(\mathbf{A})\gamma^{-1}\mathbf{H} + o(\mathbf{H}) \\ &= R(\gamma\mathbf{A}) + \mathcal{D}R(\mathbf{A})\mathbf{H} + o(\mathbf{H}) \end{aligned}$$

and thus, by Taylor expansion, R is differentiable at $\gamma\mathbf{A}$ and $\mathcal{D}R(\gamma\mathbf{A}) = \mathcal{D}R(\mathbf{A})$. □

Lemma 2.11. Suppose **Γ**1, **R**1, **R**2 and **ρ**2. Let $\theta_0 = (z_0, p_0, q_0, \hat{\Sigma}_0, r_0) \in \mathsf{X}$ and $v_1 \in \overline{\mathcal{O}^1_{\theta_0}}$. Then, if (a) $z_0 = q_0 = 0$ and $v_1 = 0$, or if (b) $p_1 = F_p(p_0, R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0; v_1) \neq 0$, and if moreover R is differentiable in $\hat{\Sigma}_1 = F_{\mathbf{\Sigma}}(q_0, R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0, r_0; v_1)$ (see (2.49) for the definition of $F_{\mathbf{\Sigma}}$), then $v \in \mathsf{V} \mapsto S^1_{\theta_0}(v)$ is differentiable at v_1.

Proof. Suppose (a). Then, for $h_1 = (h_1^1, \dots, h_1^\mu) \in \mathsf{V}$, using the update equations (2.40), (2.41), (2.42), (2.43), (2.44) we have, when $h_1 \to 0$,

$$S^1_{\theta_0}(v_1 + h_1) = \begin{pmatrix} \frac{R((1-c_1-c_\mu)R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0 + o(\|h_1\|))^{-1/2}(0 + c_m \mathbf{w}_m^\top h_1)}{\Gamma((1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}R(\hat{\Sigma}_0)^{1/2}\hat{\Sigma}_0^{-1/2}\mathbf{w}_m^\top h_1)} \\ (1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}R(\hat{\Sigma}_0)^{1/2}\hat{\Sigma}_0^{-1/2}\mathbf{w}_m^\top h_1 \\ 0 + \sqrt{c_c(2-c_v)\mu_{\text{eff}}}\mathbf{w}_m^\top h_1 \\ \frac{(1-c_1-c_\mu)R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0 + o(\|h_1\|)}{\rho((1-c_1-c_\mu)R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0 + o(\|h_1\|))} \\ R((1-c_1-c_\mu)R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0 + o(\|h_1\|)) \end{pmatrix} .$$

However, by **Γ**1, the stepsize change Γ is locally Lispchitz, hence

$$\begin{aligned} \Gamma((1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}R(\hat{\Sigma}_0)^{1/2}\hat{\Sigma}_0^{-1/2}\mathbf{w}_m^\top h_1) &= \Gamma((1-c_\sigma)p_0) + O(\|h_1\|) \\ &= \Gamma((1-c_\sigma)p_0) + o(1) \ . \end{aligned}$$

Moreover, by **ρ**2, ρ is differentiable at $(1-c_1-c_\mu)R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0$. Hence by Taylor expansion

$$\rho((1-c_1-c_\mu)R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0 + o(\|h_1\|)) = \rho((1-c_1-c_\mu)R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0) + o(\|h_1\|) \ .$$

Likewise, by assumption, R is differentiable at $(1-c_1-c_\mu)R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0$. Thus,

$$R((1-c_1-c_\mu)R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0 + o(\|h_1\|)) = R((1-c_1-c_\mu)R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0) + o(\|h_1\|) \ .$$

Therefore,

$$\begin{aligned} S^1_{\theta_0}(v_1 + h_1) = S^1_{\theta_0}(v_1) + \begin{pmatrix} R((1-c_1-c_\mu)R(\hat{\Sigma}_0)^{-1}\hat{\Sigma}_0)^{-1/2}\Gamma((1-c_\sigma)p_0)^{-1}c_m\mathbf{w}_m^\top h_1 \\ \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}R(\hat{\Sigma}_0)^{1/2}\hat{\Sigma}_0^{-1/2}\mathbf{w}_m^\top h_1 \\ \sqrt{c_c(2-c_v)\mu_{\text{eff}}}\mathbf{w}_m^\top h_1 \\ 0 \\ 0 \end{pmatrix} \\ + o(\|h_1\|) \ , \end{aligned}$$

which proves by Taylor expansion that $S^1_{\theta_0}$ is differentiable at $v_1 = 0$.

Now, suppose (b). By **Γ**1, Γ is differentiable at p_1, by **ρ**2, ρ is differentiable on $\mathcal{S}^d_{++}$, and by assumption, R is differentiable at $\hat{\mathbf{\Sigma}}_1 = \mathbf{A}_1/\rho(\mathbf{A}_1)$ where $\mathbf{A}_1 = (1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0 + c_1 q_1 (q_1)^\top + c_\mu \sum_{i=1}^{\mu} w_i^c v_1^i (v_1^i)^\top$. Since R is positively homogeneous, it is also differentiable in any multiple by a scalar of $\hat{\mathbf{\Sigma}}_1$, so in $\mathbf{A}_1$. Thus, by composition $S^1_{\theta_0}$ is differentiable at v_1. □

We prove now Proposition 2.12. The first step of the proof consists in the following lemma which applies to all cases (a)-(d) in Proposition 2.12. It provides a path $v_{1:T} \in \overline{\mathcal{O}^T_{\theta_0}}$ (where θ_0 is the steadily attracting state found in Section 5.2) such that the range of $\mathcal{D}S^T_{\theta_0}$ covers all elements in the tangent space relative to the covariance matrix variable. The proof of Lemma 2.12 is delayed to Section B.

Lemma 2.12. Suppose that the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$, the normalization functions R and ρ, the stepsize change $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$ and the sampling distribution ν^d_U satisfy **F**1-**F**2, **R**1-**R**3, **ρ**1-**ρ**2, **Γ**1-**Γ**3 and **N**1, respectively. Consider the control model (2.51) with the functions F_Θ and α_Θ defined by (2.45) and (2.39) respectively.
Then, there exist a steadily attracting state $\theta_0 \in \mathsf{X}$, $T \in \mathbb{N}$, $v_{1:T} \in \overline{\mathcal{O}^T_{\theta_0}}$, and W a subspace of V^T, such that:

(i) $S^T_{\theta_0}$ is differentiable at $v_{1:T}$;

(ii) for every $h_{\mathbf{\Sigma}} \in \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\})$, there exists $h_z, h_p \in \mathbb{R}^d$, $h_r \in \mathbb{R}$, and $h_{1:T} \in \mathsf{W}$ such that $\mathcal{D}S^T_{\theta_0}(v_{1:T})h_{1:T} = [h_z, h_p, 0, h_{\mathbf{\Sigma}}, h_r]$;

(iii) $z_T = q_T = 0$ and $p_T \neq 0$;

where $\theta_t = (z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t) = S^t_{\theta_0}(v_{1:t})$ for $t = 1, \dots, T$.

The next lemma is the second step of the proof of Proposition 2.12. It deduces from Lemma 2.12 a path in which the transition map is differentiable and is of interest to apply Theorem 2.2 or Theorem 2.3. It applies to all cases (a)-(d). We delay once more the proof of Lemma 2.13 to Section B.

Lemma 2.13. Suppose that the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$, the normalization functions R and ρ, the stepsize change $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$ and the sampling distribution ν^d_U satisfy **F**1-**F**2, **R**1-**R**3, **ρ**1-**ρ**2, **Γ**1-**Γ**3 and **N**1, respectively. Consider the control model (2.51) with the functions F_Θ and α_Θ defined by (2.45) and (2.39) respectively.
Then, there exist $\theta_0 \in \mathsf{X}$ a steadily attracting state, and $p \in \mathbb{R}^d_{\neq 0}$, such that, for every $j \in \mathbb{N}$, there exist $T \in \mathbb{N}$ and $v_{1:T} \in \overline{\mathcal{O}^T_{\theta_0}}$, with $S^T_{\theta_0}$ being differentiable at $v_{1:T}$, and

$$S^T_{\theta_0}(v_{1:T} + h_{1:T}) = S^T_{\theta_0}(v_{1:T}) + \mathbf{C}_j \times L(h_{1:T}) + o(h_{1:T}) \tag{2.55}$$

for every $h_{1:T} \in \mathsf{W}_L$, where W_L is a well-chosen subspace of V^T, $L\colon \mathsf{W}_L \to \mathbb{R}^{s-1} \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ is a surjective linear map, $s = d(d+1)/2$, and $\mathbf{C}_j$ is a matrix of the form:

$$\mathbf{C}_j = \begin{bmatrix} * & \cdots & * & 0 & 0 & L^z \\ * & \cdots & * & & & * \\ * & \cdots & * & & \mathbf{L}^{p,q}_j & * \\ \mathbf{L}^{\mathbf{\Sigma}}_1 & \dots & \mathbf{L}^{\mathbf{\Sigma}}_{s-1} & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.56}$$

with $(\mathbf{L}_1^{\mathbf{\Sigma}}, \dots, \mathbf{L}_{s-1}^{\mathbf{\Sigma}})$ being a basis of $\ker \mathcal{D}\rho(\hat{\mathbf{\Sigma}}_T)$ (with $\theta_t = (z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t) = S_{\theta_0}^t(v_{1:t})$ for $t = 0, \dots, T$), $L_z \in \mathbb{R}_{\neq 0}$, and

$$\mathbf{L}_j^{p,q} = \begin{bmatrix} (1-c_\sigma)^3 c_{j+1}^p R(\hat{\mathbf{\Sigma}}_T)^{1/2} \hat{\mathbf{\Sigma}}_T^{-1/2} & (1-c_\sigma)^1 c_{j+3}^p R(\hat{\mathbf{\Sigma}}_T)^{1/2} \hat{\mathbf{\Sigma}}_T^{-1/2} \\ (1-c_c)^3 (1-c_1-c_\mu)^{-3/2} d_{j+1}^p \mathbf{I}_d & (1-c_c)^1 (1-c_1-c_\mu)^{-1/2} d_{j+3}^p \mathbf{I}_d \end{bmatrix}, \tag{2.57}$$

where $c_k^p := (1 - c_\sigma - (1-c_1-c_\mu)^{-1/2} \Gamma((1-c_\sigma)^k p))^{-1} \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}$ and $d_k^p := (1 - c_1 - c_\mu)^{-1/2} \left[1 - c_c - \Gamma((1-c_\sigma)^k p)^{-1}\right] \sqrt{c_c(2-c_c)\mu_{\text{eff}}}$. The symbol $*$ in (2.56) represents the elements of the matrix $\mathbf{C}_j$ that we do not give explicitly (their values do not change the rank of $\mathbf{C}_j$).

Next, in order to deduce the case (a) in Proposition 2.12 from Lemma 2.13, we first show in the next lemma that the matrix $\mathbf{L}_j^{p,q}$ defined via (2.57) is invertible when the integer j is sufficiently large.

Lemma 2.14. In the context of Lemma 2.13, there exists $j \in \mathbb{N}$ such that, if $c_c \neq 1$, $c_\sigma \neq 1$, $1 - c_c \neq (1-c_\sigma)\sqrt{1-c_1-c_\mu}$, then the matrix $\mathbf{L}_j^{p,q}$ defined via (2.57) is invertible.

Proof. We have, since $c_\sigma \neq 1$, $c_c \neq 1$:

$$\begin{bmatrix} (1-c_\sigma)^{-1} R(\hat{\mathbf{\Sigma}}_T)^{-1/2} \hat{\mathbf{\Sigma}}_T^{1/2} & 0 \\ 0 & (1-c_c)^{-1}(1-c_1-c_\mu)^{1/2} \mathbf{I}_d \end{bmatrix} \times \mathbf{L}_j^{p,q} = \begin{bmatrix} (1-c_\sigma)^2 c_{j+1}^p \mathbf{I}_d & c_{j+3}^p \mathbf{I}_d \\ (1-c_c)^2 (1-c_1-c_\mu)^{-1} d_{j+1}^p \mathbf{I}_d & d_{j+3}^p \mathbf{I}_d \end{bmatrix}.$$

Therefore, it is sufficient to find some $j \in \mathbb{N}$ such that the RHS in the above equation is invertible, i.e., such that the matrix

$$\mathbf{A}_j = \begin{bmatrix} (1-c_\sigma)^2 \left[1 - c_\sigma - (1-c_1-c_\mu)^{-1/2} \Gamma_{j+1}^{-1}\right] & 1 - c_\sigma - (1-c_1-c_\mu)^{-1/2} \Gamma_{j+3}^{-1} \\ (1-c_c)^2 (1-c_1-c_\mu)^{-1} \left[1 - c_c - \Gamma_{j+1}^{-1}\right] & 1 - c_c - \Gamma_{j+3}^{-1} \end{bmatrix},$$

where $\Gamma_k = \Gamma((1-c_\sigma)^k p)$ for $k = j+2, j+4$, is full rank. Moreover, when $j \to \infty$, by continuity of Γ (by $\mathbf{\Gamma}1$), we have that Γ_{j+1} and Γ_{j+3} tend to $\Gamma(0)$. Hence,

$$\begin{aligned} &\lim_{j\to\infty} \det \mathbf{A}_j \\ &= \begin{vmatrix} (1-c_\sigma)^2 (1 - c_\sigma - (1-c_1-c_\mu)^{-1/2} \Gamma(0)^{-1}) & 1 - c_\sigma - (1-c_1-c_\mu)^{-1/2} \Gamma(0)^{-1} \\ (1-c_c)^2 (1-c_1-c_\mu)^{-1} (1 - c_c - \Gamma(0)^{-1}) & 1 - c_c - \Gamma(0)^{-1} \end{vmatrix} \\ &= (1 - c_\sigma - (1-c_1-c_\mu)^{-1/2} \Gamma(0)^{-1}) \times (1 - c_c - \Gamma(0)^{-1}) \times \begin{vmatrix} (1-c_\sigma)^2 & 1 \\ (1-c_c)^2 (1-c_1-c_\mu)^{-1} & 1 \end{vmatrix}, \end{aligned}$$

where $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ denotes the determinant of the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. However, $\Gamma(0)^{-1} > 1$ (by $\mathbf{\Gamma}3$) and $(1-c_1-c_\mu)^{-1} > 1$. Hence, there exists $j \in \mathbb{N}$, such that, if $1 - c_c \neq (1-c_\sigma)\sqrt{1-c_1-c_\mu}$, then $\det \mathbf{L}_j^{p,q} \neq 0$. □

We can now end the proof of Proposition 2.12. Depending on the case (a)-(d), the end of the proof goes differently. We present here the proofs of cases (b) and (d), and we delay those of (a) and (d) to Section B.3.

Proof of Proposition 2.12(d). Suppose that $c_c = c_\sigma = 1$. Apply Lemma 2.13, we have then that the matrix $\mathbf{L}_j^{p,q}$ defined via (2.57) is the zero matrix. Then, there exist a steadily attracting state $\theta_0 \in \mathsf{X}$, $T > 0$ and $v_{1:T} \in \overline{\mathcal{O}_{\theta_0}^T}$ such that we have that $\operatorname{rge} \mathcal{D}S_{\theta_0}^T(v_{1:T}) \supset \mathbb{R}^d \times \{0\} \times \{0\} \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\}) \times \{0\}$ and thus by taking $p = q = 0$ and $r = 0$, we have, for every $(z, \hat{\mathbf{\Sigma}}) \in \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\})$, $(z, p, q, \hat{\mathbf{\Sigma}}, r) \in \operatorname{rge} \mathcal{D}S_{\theta_0}^T(v_{1:T})$. □

Proof of Proposition 2.12(b). Suppose that $c_c = 1$ and $c_\sigma \neq 1$. By Lemma 2.13, there exist a steadily attracting state $\theta_0 \in \mathsf{X}$, $T > 0$ and $v_{1:T} \in \overline{\mathcal{O}_{\theta_0}^T}$ such that the matrix $\mathbf{L}_j^{p,q}$ defined via (2.57) satisfies:

$$\mathbf{L}_j^{p,q} = \begin{bmatrix} (1-c_\sigma)^3 c_j^p \mathbf{\Sigma}_T^{-1/2} & (1-c_\sigma) c_{j+2}^p \mathbf{\Sigma}_T^{-1/2} \\ 0 & 0 \end{bmatrix} , \tag{2.58}$$

with $\operatorname{rank} \mathbf{\Sigma}_T^{-1/2} = d$, and $c_j^p \neq 0$, $c_{j+2}^p \neq 0$. Thus, $\operatorname{rge} \mathbf{L}_j^{p,q} = \mathbb{R}^d \times \{0\}$ and $\operatorname{rge} \mathcal{D}S_{\theta_0}^T(v_{1:T}) \supset \mathbb{R}^d \times \mathbb{R}^d \times \{0\} \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\}) \times \{0\}$, and thus by taking $q = 0$ and $r = 0$, we have, for every $(z, p, \hat{\mathbf{\Sigma}}) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\})$, $(z, p, q, \hat{\mathbf{\Sigma}}, r) \in \operatorname{rge} \mathcal{D}S_{\theta_0}^T(v_{1:T})$. □

5.4 *Proof of Theorem 2.1*

In Sections 5.1 to 5.3, we have proven all required conditions to apply Theorem 2.2 or Theorem 2.3 to the Markov chain Θ defined in (2.38). The conclusion is summarized in the next theorem.

Theorem 2.5. Suppose the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$, the normalization functions R and ρ, the stepsize change $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$ and the sampling distribution ν_U^d satisfy **F**1-**F**2, **R**1-**R**3, ***ρ***1-***ρ***2, **Γ**1-**Γ**3 and **N**1, respectively.
Let $\Theta = \{(z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t)\}_{t\geqslant 1}$ be the normalized Markov chain associated to CMA-ES defined via (2.38) and P its transition kernel. Then,

(i) if $c_c, c_\sigma \in (0,1)$ are such that $1 - c_c \neq (1-c_\sigma)\sqrt{1-c_1-c_\mu}$, and if $c_\mu > 0$, then P is an irreducible aperiodic T-kernel, such that compact sets of $\mathsf{X} = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \rho^{-1}(\{1\}) \times \mathbb{R}_{++}$ are small;

(ii) if $c_c \in (0,1)$, $c_\sigma = 1$ and $c_\mu > 0$, then the normalized chain $\{(z_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t)\}_{t\geqslant 1}$ is a time-homogeneous Markov chain with an irreducible aperiodic T-kernel, such that compact sets of $\mathsf{X}_2 = \mathbb{R}^d \times \mathbb{R}^d \times \rho^{-1}(\{1\}) \times \mathbb{R}_{++}$ are small;

(iii) if $c_\sigma \in (0,1)$ and $c_c = 1$, then the normalized chain $\{(z_t, p_t, \hat{\mathbf{\Sigma}}_t)\}_{t\geqslant 1}$ is a time-homogeneous Markov chain with an irreducible aperiodic T-kernel, such that compact sets of $\mathsf{X}_3 = \mathbb{R}^d \times \mathbb{R}^d \times \rho^{-1}(\{1\})$ are small;

(iv) if $c_c = c_\sigma = 1$, then the normalized chain $\{(z_t, \hat{\mathbf{\Sigma}}_t)\}_{t\geqslant 1}$ is a time-homogeneous Markov chain with an irreducible aperiodic T-kernel, such that compact sets of $\mathsf{X}_4 = \mathbb{R}^d \times \rho^{-1}(\{1\})$ are small.

Proof. By Proposition 2.8, the Markov chain Θ follows the control model (2.51), and by Proposition 2.9, **H**1 and **H**2 hold.
Suppose first that $c_c, c_\sigma \neq 1$ and $1 - c_c \neq (1 - c_\sigma)\sqrt{1 - c_1 - c_\mu}$. By Proposition 2.12, there exist a steadily attracting state $\theta^* \in \mathsf{X}$, $T \geqslant 1$ $v_{1:T} \in \overline{\mathcal{O}^T_{\theta^*}}$ such that $\mathcal{D}S^T_{\theta^*}(v_{1:T})$ exists and is of maximal rank. Hence **H**3 holds, and we deduce then (i) by applying Theorem 2.2.
Now suppose that $c_c \neq 1$ and $c_\sigma = 1$, resp. $c_\sigma \neq 1$ and $c_c = 1$, and $c_c = c_\sigma = 1$. Then, by Corollary 2.2, $\Theta^q := \{(z_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t)\}_{t\geqslant 1}$, resp. $\Theta^p := \{(z_t, p_t, \hat{\mathbf{\Sigma}}_t)\}_{t\geqslant 1}$, and $\Theta^r := \{(z_t, \hat{\mathbf{\Sigma}}_t)\}_{t\geqslant 1}$, defines a time-homogeneous Markov chain. Moreover, since θ^* is a steadily attracting state for Θ, then, by Proposition 2.12, resp. Θ^p, Θ^q and Θ^r, follows a control model which satisfies **H**1. Thus, by Theorem 2.3, we obtain (ii), (iii) and (iv). □

Our main result Theorem 2.1, stated in Section 2, is a consequence of Theorem 2.5 and of Theorem 2.4. Indeed, consider $\rho = \det(\cdot)^{1/d}$. It is a normalization function that satisfies $\boldsymbol{\rho}$1-$\boldsymbol{\rho}$2. By Proposition 2.7, the associated Markov chain Θ following (2.38) with this normalization function is a transformation of the chain Φ defined via (2.17) by the homeomorphism ξ defined in (2.36).

By Theorem 2.5, Θ is an irreducible aperiodic T-chain. By Theorem 2.4, so is Φ. Therefore [110, Theorem 6.2.5], compact sets are small sets.

6 Conclusion and perspectives

This paper expands a methodology to analyze irreducibility and other stability properties of complex Markov chains when they are expressed as nonsmooth state-space models. We apply the methodology in the context of optimization to the CMA-ES [70]. We prove irreducibility, aperiodicity and topological properties of a stochastic process obtained by normalizing the Markov chain that represents the state of CMA-ES when optimizing scaling-invariant functions. This is an important milestone to prove the linear convergence of CMA-ES.

Our stability analysis encompasses more general processes than the one underlying CMA-ES by considering an abstract stepsize change function. Compared to previous work [141], we relax the assumption on the stepsize change from $\mathcal{C}^1$ to locally Lipschitz. This now allows to analyze the default stepsize change of CMA-ES. We also consider an abstract sampling distribution ν_U which includes multivariate normal distributions as used in CMA-ES.

We summarize the assumptions to prove stability of CMA-ES:

- The objective function is scaling-invariant. This is inherent to our methodology because we define a time-homogeneous Markov chain based upon the normalization of the state variables of CMA-ES.

- The objective function has Lebesgue negligible level sets. This is needed to obtain a lower semicontinuous density for the distribution of the ranked candidate solutions. This is a main assumption to deduce irreducibility from the analysis of an underlying control model.

- The normalization function $R(\cdot)$ is positively homogeneous and continuously Lipschitz, but $R(\cdot)$ may be nonsmooth. This includes natural normalizations, e.g., by the determinant (which is smooth) or an eigenvalue (which is nonsmooth). Positive homogeneity is needed for building the normalized Markov chain and thus (too) inherent to the Markov chain methodology. Lipschitz continuity yields a locally Lipschitz function F for the nonsmooth model (2.25) and allows to connect irreducibility to the analysis of an underlying control model [52].

- The hyperparameter setting assumptions cover all practically relevant algorithm variants (with/without cumulation, with rank-one and rank-mu updates) except when $c_1 + c_\mu = 1$, or when $c_c < 1$ and either $c_\mu = 0$ or $1 - c_c = (1 - c_\sigma)\sqrt{1 - c_1 - c_\mu}$. Without cumulation ($c_c = 1$), the rank-one update is sufficient to prove irreducibility and aperiodicity. However, we need the rank-mu update for our proof when cumulation is used ($c_c < 1$). None of the above cases is important in practice.

Limitations and perspectives We believe that some of the above assumptions can be relaxed with further work, specifically, and based on empirical observations, the assumptions that

- the hyperparameters have to be chosen suitably (in particular $0 < c_\mu < 1$),
- the objective function f has Lebesgue negligible level sets, and
- the sampling distribution is positive and continuous on the entire search space (which is not the case for a distribution on the unit sphere).

In order to conclude—with the approach pursued in this paper—the linear convergence of CMA-ES and its learning of the inverse Hessian, it still remains to be proven that the normalized Markov chain converges geometrically fast to a stationary distribution and satisfies a Law of Large Numbers. This proof could be achieved by finding a potential function for which a geometric drift condition holds [110].

A Proofs in Section 5.2

A.1 *Proof of Lemma 2.8*

Proof of Lemma 2.8. Let e_k be the k-th vector of the basis $\mathcal{B}$. Let κ be positive and κ' be real. Consider the sequence $\{\theta_t\}_{t=0,1,2,3,4}$ defined by

$$\theta_{t+1} = (z_{t+1}, p_{t+1}, q_{t+1}, \hat{\mathbf{\Sigma}}_{t+1}, r_{t+1}) = F_\Theta\left(\theta_t, \alpha_\Theta\left(\theta_t, v_{t+1}\right)\right)$$

with $v_1 = [\kappa e_k]_{i=1,\dots,\mu}$, $v_2 = -r_1^{-1/2}\Gamma(p_1)^{-1}v_1$, $v_3 = [\kappa' e_k]_{i=1,\dots,\mu}$, and $v_4 = -r_3^{-1/2}\Gamma(p_3)^{-1}v_3$. By Proposition 2.11, we have $v_{1:4} \in \overline{\mathcal{O}_{\theta_0}^4}$ and by Lemma 2.6 we obtain $z_4 = z_2 = 0$. Moreover,

$$q_2 = \kappa \times \sqrt{c_c(2 - c_c)\mu_{\text{eff}}}\, r_1^{-1/2}\left[1 - c_c - \Gamma(p_1)^{-1}\right] e_k.$$

Let $\eta \in \mathbb{R}$, and set $\kappa' = \eta \times \left(\kappa\sqrt{c_c(2 - c_c)\mu_{\text{eff}}}\, r_1^{-1/2}\left[1 - c_c - \Gamma(p_1)^{-1}\right]\right) = \eta q_2$. Then, similarly to the proof of Lemma 2.7, we have, since $v_3 = \eta[q_2, \dots, q_2]$:

$$q_4 = r_3^{-1/2} \times \left((1 - c_c)^2 r_2^{-1/2} + (1 - c_c - \Gamma(p_3)^{-1})\sqrt{c_c(2 - c_c)\mu_{\text{eff}}}\,\eta\right) \times q_2 \tag{2.59}$$

where

$$\begin{aligned} p_3 &= (1 - c_c)p_2 + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}} R(\hat{\mathbf{\Sigma}}_2)}\,\hat{\mathbf{\Sigma}}_2^{-1/2}\mathbf{w}_m^\top v_3 \\ &= (1 - c_\sigma)p_2 + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}} R(\hat{\mathbf{\Sigma}}_2)}\,\eta\hat{\mathbf{\Sigma}}_2^{-1/2} q_2 \end{aligned}$$

and thus

$$\Gamma(p_3)^{-1} = \Gamma\left((1 - c_\sigma)p_2 + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}} R(\hat{\mathbf{\Sigma}}_2)}\,\eta\hat{\mathbf{\Sigma}}_2^{-1/2} q_2\right)^{-1}.$$

We apply the intermediate value theorem to the function

$$\zeta : \eta \mapsto \left[1 - c_c - \Gamma\left((1-c_\sigma)p_2 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}} R(\hat{\mathbf{\Sigma}}_2)}\eta \hat{\mathbf{\Sigma}}_2^{-1/2} q_2\right)^{-1}\right] \times \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\eta \ ,$$

which is such that $q_4 = r_3^{-1/2} \times \left((1-c_c)^2 r_2^{-1/2} + \zeta(\eta)\right) \times q_2$. Since Γ is continuous by **R**2 and such that when η goes to $\pm\infty$, $\left[1 - c_c - \Gamma\left((1-c_\sigma)p_2 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}} R(\hat{\mathbf{\Sigma}}_2)}\eta \hat{\mathbf{\Sigma}}_2^{-1/2} q_2\right)^{-1}\right]$ is strictly positive by **Γ**2, we find that ζ is continuous and $\zeta(\eta)$ tends to $+\infty$ when η to $+\infty$, and to $-\infty$ when η to $-\infty$. Hence we find $\eta_\kappa \in \mathbb{R}$ (which depends continuously on κ) such that when $\eta = \eta_\kappa$, we have

$$q_4 = 0.$$

For the covariance matrix $\hat{\mathbf{\Sigma}}_1$, we have

$$\hat{\mathbf{\Sigma}}_1 = \frac{(1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_1 c_c(2-c_c)\mu_{\text{eff}}\kappa^2 e_k e_k^\top + c_\mu \kappa^2 e_k e_k^\top}{\rho((1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_1 c_c(2-c_c)\mu_{\text{eff}}\kappa^2 e_k e_k^\top + c_\mu \kappa^2 e_k e_k^\top)} \ ,$$

where $\mathbf{\Sigma}_0 = R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0$. Let $\rho_1 = \rho((1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_1 c_c(2-c_c)\mu_{\text{eff}}\kappa^2 e_k e_k^\top + c_\mu \kappa^2 e_k e_k^\top)$ and $\omega_1(\kappa) = c_1 c_c(2-c_c)\mu_{\text{eff}}\kappa^2 + c_\mu\kappa^2$ such that

$$\begin{aligned}\hat{\mathbf{\Sigma}}_1 &= \rho_1^{-1}\left[(1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_1 c_c(2-c_c)\mu_{\text{eff}}\kappa^2 e_k e_k^\top + c_\mu\kappa^2 e_k e_k^\top\right] \\ &=: \rho_1^{-1}\left[(1-c_1-c_\mu)\mathbf{\Sigma}_0 + \omega_1(\kappa) e_k e_k^\top\right].\end{aligned}$$

The map ω_1 is continuous, with $\omega_1(0) = 0$ and $\omega_1(\kappa) \to \infty$ when $\kappa \to \infty$. Similarly, setting

$$\rho_2 = \rho\left(R(\hat{\mathbf{\Sigma}}_1)^{-1}\hat{\mathbf{\Sigma}}_1 + \left(c_1 c_c(2-c_c)\mu_{\text{eff}}\kappa^2\left(1 - c_c - \Gamma(p_1)^{-1}\right) + c_\mu\kappa^2\Gamma(p_1)^{-2}\right)e_k e_k^\top\right)$$

we get

$$\begin{aligned}\hat{\mathbf{\Sigma}}_2 = \rho_2^{-1}\Big[&R(\hat{\mathbf{\Sigma}}_1)^{-1}(1-c_1-c_\mu)^2\mathbf{\Sigma}_0 \\ &+ \left(R(\hat{\mathbf{\Sigma}}_1)^{-1}(1-c_1-c_\mu)\omega_1(\kappa) + c_1 c_c(2-c_c)\mu_{\text{eff}}\kappa^2\left(1-c_c-\Gamma(p_1)^{-1}\right) + c_\mu\kappa^2\Gamma(p_1)^{-2}\right)e_k e_k^\top\Big] \\ &=: \rho_2^{-1}\left[R(\hat{\mathbf{\Sigma}}_1)^{-1}(1-c_1-c_\mu)^2\mathbf{\Sigma}_0 + \omega_2(\kappa)e_k e_k^\top\right],\end{aligned}$$

with ω_2 continuous since R and Γ are continuous by **R**2 and **Γ**1, $\omega_2(0) = 0$ and $\omega_2(\kappa) \to \infty$ when $\kappa \to \infty$.
Likewise, for the next two steps, we find $\rho_4 > 0$, ω_4 continuous such that $\omega_4(0) = 0$, and $\omega_4(\kappa) \to \infty$ when $\kappa \to \infty$, and

$$\hat{\mathbf{\Sigma}}_4 = \rho_4^{-1}\left[R(\hat{\mathbf{\Sigma}}_1)^{-1}R(\hat{\mathbf{\Sigma}}_2)^{-1}R(\hat{\mathbf{\Sigma}}_3)^{-1}(1-c_1-c_\mu)^4\mathbf{\Sigma}_0 + \omega_4(\kappa)e_k e_k^\top\right].$$

Then, by the intermediate value theorem, there exists $\kappa > 0$ such that

$$\omega_4(\kappa) = R(\hat{\mathbf{\Sigma}}_0)^{-1}R(\hat{\mathbf{\Sigma}}_1)^{-1}R(\hat{\mathbf{\Sigma}}_2)^{-1}R(\hat{\mathbf{\Sigma}}_3)^{-1}(1-c_1-c_\mu)^4(\lambda_{k-1} - \lambda_k) > 0 \ .$$

Therefore,

$$\begin{aligned}[\hat{\mathbf{\Sigma}}_4]_{\mathcal{B}} = \rho_4^{-1}&R(\hat{\mathbf{\Sigma}}_0)^{-1}R(\hat{\mathbf{\Sigma}}_1)^{-1}R(\hat{\mathbf{\Sigma}}_2)^{-1}R(\hat{\mathbf{\Sigma}}_3)^{-1}(1-c_1-c_\mu)^4 \\ &\times \operatorname{diag}(\lambda_1, \ldots, \lambda_{k-1}, \lambda_{k-1}, \lambda_{k+1}, \ldots, \lambda_d) \ .\end{aligned}$$

Setting $\gamma = \rho_4^{-1}R(\hat{\mathbf{\Sigma}}_0)^{-1}R(\hat{\mathbf{\Sigma}}_1)^{-1}R(\hat{\mathbf{\Sigma}}_2)^{-1}R(\hat{\mathbf{\Sigma}}_3)^{-1}(1-c_1-c_\mu)^4$, we have proven that we can reach θ_4 with the matrix $\hat{\mathbf{\Sigma}}_4$ defined in (2.54). □

B Proofs in Section 5.3

B.1 Proof of Lemma 2.12

Proof of Lemma 2.12. By **R**3, there exists $\mathbf{C}_0 \in \mathcal{S}^d_{++}$, such that R is differentiable on a neighborhood of $\mathbf{C}_0$. Since R is positively homogeneous by **R**1, then by Lemma 2.10 R is also differentiable on a neighborhood of $\hat{\mathbf{\Sigma}}_0 \coloneqq \rho(\mathbf{C}_0)^{-1}\mathbf{C}_0$. Then, by Corollary 2.3, there exists $\theta_0 = (z_0, p_0, q_0, \hat{\mathbf{\Sigma}}_0, r_0)$ with $z_0 = q_0 = 0$ which is a steadily attracting state.
Let $T \in \mathbb{N}$, $v_{1:T} \in \overline{\mathcal{O}^T_{\theta_0}}$ and $h_{1:T} \in \mathsf{V}^T$. We denote for $t \in \{1, \dots, T\}$:

$$\theta_t = (z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t) = S^t_{\theta_0}(v_{1:t}) \quad \text{and} \quad \theta^h_t = (z^h_t, p^h_t, q^h_t, \hat{\mathbf{\Sigma}}^h_t, r^h_t) = S^t_{\theta_0}(v_{1:t} + h_{1:t}) \ .$$

We have that, if $v_{1:T} = 0$, then since $q_0 = 0$, $q_t = 0$ for $t = 1, \dots, T$ and using (2.43) we find $\hat{\mathbf{\Sigma}}_t = \hat{\mathbf{\Sigma}}_0$. Since $v \mapsto S^t_{\theta_0}(v)$ is continuous, and since R is differentiable in a neighborhood of $\hat{\mathbf{\Sigma}}_0$, then there exists $M_V > 0$ such that, if $\|v_{1:T}\|^2 \leqslant M_V$, then R is differentiable at $\hat{\mathbf{\Sigma}}_t$. Hence we impose that $\|v_t\|^2 \leqslant M_V/T$ for all $t \in \{1, \dots, T\}$.
Define, for $t = 0, \dots, T$, $b_t = r_0 \times \cdots \times r_t \times R(\hat{\mathbf{\Sigma}}_t)^{-1}$ and $b^h_t = r^h_0 \times \dots r^h_t \times R(\hat{\mathbf{\Sigma}}^h_t)^{-1}$, and let $\mathbf{B}_t = b_t\hat{\mathbf{\Sigma}}_t$ and likewise $\mathbf{B}^h_t = b^h_t\hat{\mathbf{\Sigma}}^h_t$. Therefore by positive homogeneity of ρ, $\rho(\mathbf{B}_t) = \rho(b_t\hat{\mathbf{\Sigma}}_t) = b_t\rho(\hat{\mathbf{\Sigma}}_t) = b_t$ since $\rho(\hat{\mathbf{\Sigma}}_t) = 1$. Similarly $\rho(\mathbf{B}^h_t) = b^h_t$ and thus

$$\hat{\mathbf{\Sigma}}_t = \frac{\mathbf{B}_t}{\rho(\mathbf{B}_t)} \quad \text{and} \quad \hat{\mathbf{\Sigma}}^h_t = \frac{\mathbf{B}^h_t}{\rho(\mathbf{B}^h_t)} \ . \tag{2.60}$$

Moreover, define $\tilde{q}_t = \sqrt{\tilde{r}_{t-1}}q_t$ and $\tilde{q}^h_t = \sqrt{\tilde{r}^h_{t-1}}q^h_t$ as well as $\tilde{v}_{t+1} = \sqrt{\tilde{r}_t}v_{t+1}$ and $\tilde{h}_{t+1} = \sqrt{\tilde{r}^h_t}h_{t+1}$ where $\tilde{r}_t = r_0 \times \cdots \times r_t$ and $\tilde{r}^h_t = r^h_0 \times \cdots \times r^h_t$. Hence, by applying (2.42):

$$\begin{aligned}\tilde{q}_{t+1} &= \sqrt{\tilde{r}_t}r_t^{-1/2}(1-c_c)q_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\sqrt{\tilde{r}_t}\mathbf{w}_m^\top v_{t+1} \\ &= (1-c_c)\sqrt{\tilde{r}_{t-1}}q_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top \tilde{v}_{t+1} = (1-c_c)\tilde{q}_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top \tilde{v}_{t+1}\end{aligned}$$

and likewise

$$\tilde{q}^h_{t+1} = (1-c_c)\tilde{q}^h_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top \left((\tilde{r}^h_t/\tilde{r}_t)^{1/2}\tilde{v}_{t+1} + \tilde{h}_{t+1}\right) \ . \tag{2.61}$$

Denote $\mathbf{A}_{t+1}$ such that $\hat{\mathbf{\Sigma}}_{t+1} = \mathbf{A}_{t+1}/\rho(\mathbf{A}_{t+1})$ in (2.43). (Alternatively the matrix $\tilde{\mathbf{\Sigma}}_{t+1}$ in (2.19) equals $\mathbf{A}_{t+1}$). Then by positive homogeneity of R, $R(\hat{\mathbf{\Sigma}}_{t+1}) = R(\mathbf{A}_{t+1})/\rho(\mathbf{A}_{t+1})$ such that

$$\frac{\mathbf{A}_{t+1}}{R(\mathbf{A}_{t+1})} = \frac{\mathbf{A}_{t+1}}{\rho(\mathbf{A}_{t+1})R(\hat{\mathbf{\Sigma}}_{t+1})} = \frac{\hat{\mathbf{\Sigma}}_{t+1}}{R(\hat{\mathbf{\Sigma}}_{t+1})} \tag{2.62}$$

Then, using the previous equation and (2.43):

$$\mathbf{B}_{t+1} = b_{t+1}\hat{\mathbf{\Sigma}}_{t+1} = r_0 \times \cdots \times r_{t+1} \times R(\hat{\mathbf{\Sigma}}_{t+1})^{-1}\hat{\mathbf{\Sigma}}_{t+1} = \tilde{r}_{t+1}R(\hat{\mathbf{\Sigma}}_{t+1})^{-1}\hat{\mathbf{\Sigma}}_{t+1} \tag{2.63}$$

$$= \tilde{r}_{t+1}R(\mathbf{A}_{t+1})^{-1}\mathbf{A}_{t+1} \tag{2.64}$$

$$= \underbrace{\tilde{r}_{t+1} \times r_{t+1}^{-1}}_{\tilde{r}_t} \times \left((1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_t)^{-1}\hat{\mathbf{\Sigma}}_t + c_1 q_{t+1}q_{t+1}^\top + c_\mu \sum_{i=1}^{\mu} w_i^c v_{t+1}^i (v_{t+1}^i)^\top\right) \tag{2.65}$$

$$= (1-c_1-c_\mu)\tilde{r}_t R(\hat{\mathbf{\Sigma}}_t)^{-1}\hat{\mathbf{\Sigma}}_t + c_1\tilde{r}_t q_{t+1}q_{t+1}^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \tilde{r}_t v_{t+1}^i (v_{t+1}^i)^\top \tag{2.66}$$

$$= (1-c_1-c_\mu)\mathbf{B}_t + c_1\tilde{q}_{t+1}\tilde{q}_{t+1}^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \tilde{v}_{t+1}^i (\tilde{v}_{t+1}^i)^\top \ . \tag{2.67}$$

Likewise,

$$\mathbf{B}_{t+1}^h = (1-c_1-c_\mu)\mathbf{B}_t^h + c_1\tilde{q}_{t+1}^h(\tilde{q}_{t+1}^h)^\top + c_\mu\tilde{r}_t^h \sum_{i=1}^{\mu} w_i^c (v_{t+1}^i + h_{t+1}^i)(v_{t+1}^i + h_{t+1}^i)^\top \ . \tag{2.68}$$

Let $s = d(d+1)/2$ be the dimension of the set of symmetric matrices $\mathcal{S}^d$ as a real vector space. Let $\psi_1 \in \mathbb{R}^d$ be a nonzero vector and define then $\psi_2, \dots, \psi_s$ nonzero vectors of $\mathbb{R}^d$, such that $(\psi_1\psi_1^\top, \dots, \psi_s\psi_s^\top)$ forms a basis of $\mathcal{S}^d$. Scaling down the length of ψ_k does not change that we have a basis of $\mathcal{S}^d$ and thus we impose that $\|\psi_k\| \leqslant \varepsilon$, where ε is a positive constant that we precise in the next paragraph. Set $T = 2s(s-1)+4$ and set $v_{1:T}$ as below.
For $t \in \{0, \dots, s-1\}$, we set

$$v_{2t+1} = \tilde{r}_{2t}^{-1/2}[\psi_{t+1}, \dots, \psi_{t+1}] \in \mathbb{R}^{d\mu} \text{ and } v_{2t+2} = -(1-c_c)\tilde{r}_{2t+1}^{-1/2}[\psi_{t+1}, \dots, \psi_{t+1}] \in \mathbb{R}^{d\mu} \tag{2.69}$$

such that

$$\tilde{v}_{2t+1} = [\psi_{t+1}, \dots, \psi_{t+1}] \text{ and } \tilde{v}_{2t+2} = -(1-c_c)[\psi_{t+1}, \dots, \psi_{t+1}]. \tag{2.70}$$

Moreover, we choose $\varepsilon > 0$ small enough so that $\|v_k\|^2 \leqslant M_V/T$ for all $k = 1, \dots, 2s$. By definition of M_V earlier in the proof, we have that R is differentiable in $\hat{\mathbf{\Sigma}}_t$ for $t = 0, \dots, T$. If moreover $p_t \neq 0$ for $t = 1, \dots, T$, then by Lemma 2.11, $v \to S_{\theta_0}^T(v)$ is differentiable in $v_{1:T}$. Besides, by Proposition 2.11, we have $v_{1:2s} \in \overline{\mathcal{O}_{\theta_0}^{2s}}$.
Observe now that there exists $\psi_1 \in \mathbb{R}^d$ such that $\|\psi_1\| \leqslant \varepsilon$ and p_1, p_2 are nonzero. Indeed, $p_1 = (1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}R(\hat{\mathbf{\Sigma}}_0)^{1/2}\hat{\mathbf{\Sigma}}_0^{-1/2}\sum_{i=1}^{\mu} w_i^m v_1^i$ and using $\mathbf{B}_0 = r_0 R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0$ we find $p_1 = (1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}r_0^{1/2}\mathbf{B}_0^{-1/2}\sum_{i=1}^{\mu} w_i^m \tilde{r}_0^{-1/2}\psi_1 = (1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbf{B}_0^{-1/2}\psi_1$ and

$$p_2 = (1-c_\sigma)p_1 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}R(\hat{\mathbf{\Sigma}}_1)^{1/2}\hat{\mathbf{\Sigma}}_1^{-1/2}\sum_{i=1}^{\mu} w_i^m v_2^i \tag{2.71}$$

$$= (1-c_\sigma)^2 p_0 + (1-c_\sigma)\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbf{B}_0^{-1/2}\psi_1 \tag{2.72}$$

$$- \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\tilde{r}_1^{1/2}\mathbf{B}_1^{-1/2}\sum w_i^m(1-c_c)\tilde{r}_1^{-1/2}\psi_1 \tag{2.73}$$

$$= (1-c_\sigma)^2 p_0 + [(1-c_\sigma)\mathbf{B}_0^{-1/2} - (1-c_c)\mathbf{B}_1^{-1/2}]\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\psi_1 \ . \tag{2.74}$$

Since $v_2 = -(1-c_c)\tilde{r}_1^{-1/2}[\psi_1,\dots,\psi_1]$ with $\psi_1 \neq 0$, given that according to (2.67), $\mathbf{B}_1 = \alpha_1\mathbf{B}_0 + \alpha_2\psi_1\psi_1^\top + \alpha_3 q_1 q_1^\top$, for $\alpha_1,\alpha_2,\alpha_3$ some nonnegative constants and $\alpha_2+\alpha_3>0$ since $c_1+c_\mu>0$, and $\psi_1, q_1 \neq 0$ (see below), we have $(1-c_\sigma)\mathbf{B}_0^{-1/2} \neq (1-c_c)\mathbf{B}_1^{-1/2}$. Moreover, up to scaling $\psi_2,\dots,\psi_s$ sufficiently smaller than ψ_1, we can ensure that $p_t \neq 0$ for $t \in \{3,\dots,2s\}$. Then, by Lemma 2.11, and by composition since $S_{\theta_0}^{t+1}(v_{1:t+1}) = S^1_{S^t_{\theta_0}(v_{1:t})}(v_{t+1})$ we find by induction that $S^{2s}_{\theta_0}$ is differentiable at $v_{1:2s}$. Then, by induction, since $\tilde{q}_{t+1} = (1-c_c)\tilde{q}_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top \tilde{v}_{t+1}$ with $\tilde{v}_{2t+1} = [\psi_{t+1},\dots,\psi_{t+1}]$ and $\tilde{v}_{2t+2} = -(1-c_c)[\psi_{t+1},\dots,\psi_{t+1}]$, we find that, for every $t \in \{0,\dots,s-1\}$, we have:

$$\tilde{q}_{2t+1} = \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\psi_{t+1} \text{ and } q_{2t+2} = 0. \tag{2.75}$$

For $t=0,\dots,s-1$, let $\kappa_t^1 \in \mathbb{R}$ be arbitrary (we fix the value of κ_t^1 later in the proof). We set, given an arbitrary real number $\varepsilon_1 \in \mathbb{R}$, for $t=0,\dots,s-1$:

$$h_{2t+1} = [(\tilde{r}_{2t}^h)^{-1/2} - \tilde{r}_{2t}^{-1/2} + (\tilde{r}_{2t}^h)^{-1/2}\kappa_t^1\varepsilon_1] \times [\psi_{t+1},\dots,\psi_{t+1}] \in \mathbb{R}^{d\mu}$$

which implies

$$\tilde{h}_{2t+1} = [1 - (\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2} + \kappa_t^1\varepsilon_1] \times [\psi_{t+1},\dots,\psi_{t+1}] \tag{2.76}$$

and

$$h_{2t+2} = -(1-c_c)[(\tilde{r}_{2t+1}^h)^{-1/2} - \tilde{r}_{2t+1}^{-1/2} + (\tilde{r}_{2t+1}^h)^{-1/2}\kappa_t^1\varepsilon_1] \times [\psi_{t+1},\dots,\psi_{t+1}] \in \mathbb{R}^{d\mu} \ ,$$

so that, by induction, starting from (2.61) we have:

$$\begin{aligned}
\tilde{q}_{2t+1}^h &= (1-c_c)\tilde{q}_{2t}^h + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top((\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2}\tilde{v}_{2t+1} + \tilde{h}_{2t+1}) && (2.77)\\
&= (1-c_c)\times 0 && (2.78)\\
&\quad + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top\left((\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2} + \left(1 - (\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2} + \kappa_t^1\varepsilon_1\right)\right)[\psi_{t+1},\dots,\psi_{t+1}] && (2.79)\\
&= \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\left(1+\kappa_t^1\varepsilon_1\right)\psi_{t+1} && (2.80)
\end{aligned}$$

and

$$\begin{aligned}
\tilde{q}_{2t+2}^h &= (1-c_c)\tilde{q}_{2t+1}^h + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top((\tilde{r}_{2t+1}^h/\tilde{r}_{2t+1})^{1/2}\tilde{v}_{2t+2} + \tilde{h}_{2t+2})\\
&= (1-c_c)\sqrt{c_c(2-c_c)\mu_{\text{eff}}}\left(1+\kappa_t^1\varepsilon_1\right)\psi_{t+1}\\
&\quad - (1-c_c)\sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top\left(\sqrt{\frac{\tilde{r}_{2t+1}^h}{\tilde{r}_{2t+1}}} + \left(1 - \sqrt{\frac{\tilde{r}_{2t+1}^h}{\tilde{r}_{2t+1}}} + \kappa_t^1\varepsilon_1\right)\right)[\psi_{t+1},\dots,\psi_{t+1}]\\
&= 0
\end{aligned}$$

Note that, for $i=1,\dots,\mu$:

$$(\tilde{r}_{2t}^h)^{1/2}\times(v_{2t+1}^i + h_{2t+1}^i) = \left((\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2} + \left(1 - (\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2} + \kappa_t^1\varepsilon_1\right)\right)\psi_{t+1} = \left(1+\kappa_t^1\varepsilon_1\right)\psi_{t+1} \ .$$

Then, using (2.68), we obtain for $t \in \{0, \dots, s-1\}$, when $\varepsilon_1 \to 0$:

$$\mathbf{B}^h_{2t+1} = (1-c_1-c_\mu)\mathbf{B}^h_{2t} + c_1c_c(2-c_c)\mu_{\text{eff}}(1+\kappa^1_t\varepsilon_1)^2\psi_{t+1}\psi_{t+1}^\top \tag{2.81}$$

$$+ c_\mu \sum_{i=1}^{\mu} w_i^c \left(1+\kappa^1_t\varepsilon_1\right)^2 \psi_{t+1}\psi_{t+1}^\top \tag{2.82}$$

$$= (1-c_1-c_\mu)\mathbf{B}^h_{2t} + \left[c_1c_c(2-c_c)\mu_{\text{eff}} + c_\mu\right] \times \left(1+2\kappa^1_t\varepsilon_1\right)\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1) \tag{2.83}$$

From (2.67), we have that

$$\begin{aligned}
\mathbf{B}_{2t+1} &= (1-c_1-c_\mu)\mathbf{B}_{2t} + c_1\tilde{q}_{2t+1}\tilde{q}_{2t+1}^\top + c_\mu\sum_{i=1}^{\mu} w_i^c \tilde{v}^i_{2t+1}(\tilde{v}^i_{2t+1})^\top \\
&= (1-c_1-c_\mu)\mathbf{B}_{2t} + c_1c_c(2-c_c)\mu_{\text{eff}}\psi_{t+1}\psi_{t+1}^\top + c_\mu\sum_{i=1}^{\mu} w_i^c \psi_{t+1}\psi_{t+1}^\top \\
&= (1-c_1-c_\mu)\mathbf{B}_{2t} + \left[c_1c_c(2-c_c)\mu_{\text{eff}} + c_\mu\right]\psi_{t+1}\psi_{t+1}^\top
\end{aligned}$$

that we use in (2.83) to obtain

$$\begin{aligned}
\mathbf{B}^h_{2t+1} &= \mathbf{B}_{2t+1} + (1-c_1-c_\mu)\left(\mathbf{B}^h_{2t} - \mathbf{B}_{2t}\right) + \underbrace{\left[c_1c_c(2-c_c)\mu_{\text{eff}} + c_\mu\right] \times 2}_{:=c_b}\kappa^1_t\varepsilon_1\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1) \\
&= \mathbf{B}_{2t+1} + (1-c_1-c_\mu)\left(\mathbf{B}^h_{2t} - \mathbf{B}_{2t}\right) + c_b\kappa^1_t\varepsilon_1\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1) \ . \qquad (2.84)
\end{aligned}$$

Moreover, for $i = 1, \dots, \mu$:

$$\begin{aligned}
(\tilde{r}^h_{2t+1})^{1/2}(v^i_{2t+2} + h^i_{2t+2}) &= -(1-c_c)\left((\tilde{r}^h_{2t+1}/\tilde{r}_{2t+1})^{1/2} + \left(1 - (\tilde{r}^h_{2t+1}/\tilde{r}_{2t+1})^{1/2} + \kappa^1_t\varepsilon_1\right)\right)\psi_{t+1} \\
&= -(1-c_c)\left(1+\kappa^1_t\varepsilon_1\right)\psi_{t+1} \ .
\end{aligned}$$

Thus, we obtain, by (2.68) and (2.84):

$$\begin{aligned}
\mathbf{B}^h_{2t+2} &= (1-c_1-c_\mu)\mathbf{B}^h_{2t+1} + c_\mu\sum_{i=1}^{\mu} w_i^c(1-c_c)^2(1+\kappa^1_t\varepsilon_1)^2\psi_{t+1}\psi_{t+1}^\top \\
&= (1-c_1-c_\mu)\mathbf{B}_{2t+1} + (1-c_1-c_\mu)^2\left(\mathbf{B}^h_{2t} - \mathbf{B}_{2t}\right) + (1-c_1-c_\mu)c_b\kappa^1_t\varepsilon_1\psi_{t+1}\psi_{t+1}^\top \\
&\quad + c_\mu(1-c_c)^2(1+\kappa^1_t\varepsilon_1)^2\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1) \\
&= (1-c_1-c_\mu)\mathbf{B}_{2t+1} + c_\mu(1-c_c)^2\psi_{t+1}\psi_{t+1}^\top + (1-c_1-c_\mu)^2\left(\mathbf{B}^h_{2t} - \mathbf{B}_{2t}\right) \\
&\quad + d_b\kappa^1_t\varepsilon_1\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1)
\end{aligned}$$

with

$$d_b = (1-c_1-c_\mu)c_b + 2c_\mu(1-c_c)^2 = 2(1-c_1-c_\mu)\left(c_1c_c(2-c_c)\mu_{\text{eff}} + c_\mu\right) + 2c_\mu(1-c_c)^2 \ .$$

Yet, by (2.67) since by (2.75) $q_{2t+2} = 0$ and by (2.70) $\tilde{v}^i_{2t+2} = -(1-c_c)\psi_{t+1}$:

$$\mathbf{B}_{2t+2} = (1-c_1-c_\mu)\mathbf{B}_{2t+1} + c_\mu(1-c_c)^2\psi_{t+1}\psi_{t+1}^\top \ .$$

Therefore,

$$\mathbf{B}^h_{2t+2} - \mathbf{B}_{2t+2} = (1-c_1-c_\mu)^2\left(\mathbf{B}^h_{2t} - \mathbf{B}_{2t}\right) + d_b\kappa^1_t\varepsilon_1\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1) \ .$$

Then, by induction, we get,

$$\mathbf{B}_{2s}^h = \mathbf{B}_{2s} + \varepsilon_1 \sum_{t=1}^{s} (1-c_1-c_\mu)^{2s-2t} d_b \kappa_{t-1}^1 \psi_t \psi_t^\top + (1-c_1-c_\mu)^{2s}(\mathbf{B}_0^h - \mathbf{B}_0) + o(\varepsilon_1) \ ,$$

with $\mathbf{B}_0^h - \mathbf{B}_0 = 0$ by definition. By induction on $k \in \{1, \dots, s-2\}$, we set for $t \in \{0, \dots, s-1\}$:

$$v_{2t+2ks+1} = \tilde{r}_{2t+2ks}^{-1/2}[\psi_{t+1}, \dots, \psi_{t+1}] \in \mathbb{R}^{d\mu} \ ,$$

and

$$v_{2t+2ks+2} = -(1-c_c)\tilde{r}_{2t+2ks+1}^{-1/2}[\psi_{t+1}, \dots, \psi_{t+1}] \in \mathbb{R}^{d\mu} \ .$$

We also set, given arbitrary real numbers $\varepsilon_k \in \mathbb{R}$ and for some $\kappa_t^k \in \mathbb{R}$ for $t \in \{0, \dots, s-1\}$:

$$h_{2t+2ks+1} = [(\tilde{r}_{2t+2ks}^h)^{-1/2} - \tilde{r}_{2t+2ks}^{-1/2} + (\tilde{r}_{2t+2ks}^h)^{-1/2}\kappa_t^s \varepsilon_1] \times [\psi_{t+1}, \dots, \psi_{t+1}] \in \mathbb{R}^{d\mu}$$

and

$$h_{2t+2ks+2} = -(1-c_c)[(\tilde{r}_{2t+2ks+1}^h)^{-1/2} - \tilde{r}_{2t+2ks+1}^{-1/2} + (\tilde{r}_{2t+2ks+1}^h)^{-1/2}\kappa_t^s \varepsilon_1] \times [\psi_{t+1}, \dots, \psi_{t+1}].$$

Then, similarly to above, we obtain $q_{2(k+1)s} = 0$ and

$$\mathbf{B}_{2(k+1)s}^h = \mathbf{B}_{2(k+1)s} + \varepsilon_k \sum_{t=1}^{s} (1-c_1-c_\mu)^{2s-2t} d_b \kappa_{t-1}^k \psi_t \psi_t^\top + (1-c_1-c_\mu)^{2s}(\mathbf{B}_{2ks}^h - \mathbf{B}_{2ks}) + o(\varepsilon_k) \ .$$

Thus, by induction, we get $q_{2s(s-1)}^h = 0$ and

$$\mathbf{B}_{2s(s-1)}^h = \mathbf{B}_{2s(s-1)} + \sum_{k=1}^{s-1} \varepsilon_k (1-c_1-c_\mu)^{2s(s-1)-2ks} \sum_{t=1}^{s} (1-c_1-c_\mu)^{2s-2t} d_b \kappa_{t-1}^k \psi_t \psi_t^\top + o(\varepsilon_{1:s-1})$$

Note moreover that we can assume again that $p_t \neq 0$, up to choosing again the ψ_k, $k \geqslant 2$, sufficiently smaller than ψ_1.

By Lemmas 2.5 and 2.7, for any $v_{2s(s-1)+1} \in \overline{\mathcal{O}^1_{\theta_{2s(s-1)}}}$, there exists $v_{2s(s-1)+2:2s(s-1)+4} \in \overline{\mathcal{O}^3_{\theta_{2s(s-1)+1}}}$ such that $z_{2s(s-1)+4} = q_{2s(s-1)+4} = 0$. Moreover, when $v_{2s(s-1)+1} \to 0$, then we have that $z_{2s(s-1)+1}$ and $q_{2s(s-1)+1}$ tend to 0 and thus we can impose that $v_{2s(s-1)+2:2s(s-1)+4} \to 0$ as well. In particular, we can choose $v_{2s(s-1)+1}$ small enough such that $p_t \neq 0$ for $t = 2s(s-1)+1, \dots, 2s(s-1)+4$. Hence, by Lemma 2.11, $S_{\theta_0}^{2s(s-1)+4}$ is differentiable at $v_{2s(s-1)+4}$ (we have that R is differentiable at $\hat{\Sigma}_t$ for all $t = 1, \dots, T$ by imposing $v_{1:T}$ small enough, see the beginning of the proof).

Consider then $(\mathbf{S}_1, \dots, \mathbf{S}_{s-1})$ a basis of $\ker \mathcal{D}\rho(\mathbf{B}_{2s(s-1)+4})$. For $k = 1, \dots, s-1$, we can choose then the $\kappa_t^k \in \mathbb{R}$, $t = 0, \dots, s-1$ so that we have

$$(1-c_1-c_\mu)^{2s(s-1)-2ks} \sum_{t=1}^{s} (1-c_1-c_\mu)^{2s-2t} d_b \kappa_{t-1}^k \psi_t \psi_t^\top = \mathbf{S}_k \ . \tag{2.85}$$

This is possible since $(\psi_1\psi_1^\top, \dots, \psi_s\psi_s^\top)$ is a basis of $\mathcal{S}^d$, and by the intermediate value theorem applied to the LHS of (2.85), for $k = 1, \dots, s$. Set $T = 2s(s-1)+4$. Then, we have, when $\varepsilon_{1:s-1} \to 0$,

$$\mathbf{B}_T^h = \mathbf{B}_T + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1}) \ . \tag{2.86}$$

Therefore, since $\mathbf{S}_k \in \ker \mathcal{D}\rho(\mathbf{B}_T)$ for $k = 1, \dots, s$, we have

$$\rho\left(\mathbf{B}_T + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1})\right) = \rho(\mathbf{B}_T) + \sum_{k=1}^{s-1} \varepsilon_k \underbrace{\mathcal{D}\rho(\mathbf{B}_T)\mathbf{S}_k}_{=0} + o(\varepsilon_{1:s-1}) = \rho(\mathbf{B}_T) + o(\varepsilon_{1:s-1})$$

and using (2.60) and (2.86)

$$\begin{aligned}\hat{\mathbf{\Sigma}}_T^h &= \frac{\mathbf{B}_T + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1})}{\rho\left(\mathbf{B}_T + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1})\right)} = \frac{\mathbf{B}_T + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k}{\rho\left(\mathbf{B}_T\right)} + o(\varepsilon_{1:s-1}) \\ &= \hat{\mathbf{\Sigma}}_T + \sum_{k=1}^{s-1} \varepsilon_k \rho\left(\mathbf{B}_T\right)^{-1} \mathbf{S}_k + o(\varepsilon_{1:s-1}).\end{aligned}$$

However, $(\mathbf{S}_1, \dots, \mathbf{S}_{s-1})$ is a basis of $\ker \mathcal{D}\rho(\mathbf{B}_T)$, and by Lemma 2.10, $\ker \mathcal{D}\rho(\mathbf{B}_T) = \ker \mathcal{D}\rho(\hat{\mathbf{\Sigma}}_T) = \mathrm{T}_{\hat{\mathbf{\Sigma}}_T} \rho^{-1}(\{1\})$. Thus we have shown that for every $h_{\mathbf{\Sigma}} \in T_{\hat{\mathbf{\Sigma}}_T} \rho^{-1}(\{1\})$ for which we can find ε_k such that $h_{\mathbf{\Sigma}} = \sum_{k=1}^{s-1} \varepsilon_k \rho\left(\mathbf{B}_T\right)^{-1} \mathbf{S}_k$, there exist $h_z, h_p \in \mathbb{R}^d$, $h_r \in \mathbb{R}$ and $h_{1:T} \in (\mathbb{R}^{d\mu})^T$ such that $\mathcal{D}S_{\theta_0}^T(v_{1:T}) h_{1:T} = [h_z, h_p, 0, h_{\mathbf{\Sigma}}, h_t]$ which is the statement (ii) of the lemma. The statements (i) and (iii) have been proven earlier in the proof. □

B.2 *Proof of Lemma 2.13*

Proof of Lemma 2.13. Let $\theta_0 \in \mathsf{X}$ be a steadily attracting state satisfying Lemma 2.12. Then, there exists $T_0 > 0$ and $v_{1:T0} \in \overline{\mathcal{O}_{\theta_0}^T}$ such that conditions (i), (ii), (iii) of Lemma 2.12 are satisfied. Let $T > T_0$, $v_{1:T} \in \overline{\mathcal{O}_{\theta_0}^T}$ and $h_{1:T} \in \mathsf{V}^T$. We denote for every $t \in \{1, \dots, T\}$

$$\theta_t = (z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t) \coloneqq S_{\theta_0}^t(v_{1:t}) \quad \text{and} \quad \theta_t^h = (z_t^h, p_t^h, q_t^h, \hat{\mathbf{\Sigma}}_t^h, r_t^h) \coloneqq S_{\theta_0}^t(v_{1:t} + h_{1:t}).$$

Let $s = d(d+1)/2$ be the dimension of $\mathcal{S}^d$. Then, $\ker \mathcal{D}\rho(\hat{\mathbf{\Sigma}}_{T_0}) = \mathrm{T}_{\hat{\mathbf{\Sigma}}_{T_0}} \rho^{-1}(\{1\})$ is a vector space of dimension $s-1$. Let $(\mathbf{S}_1, \dots, \mathbf{S}_{s-1})$ be a basis of $\ker \mathcal{D}\rho(\hat{\mathbf{\Sigma}}_{T_0})$. Then for $k = 1, \dots, s-1$, by condition (ii) in Lemma 2.12, there exists $\xi_{1:T_0}^k \in \mathsf{V}^{T_0}$ such that $\mathcal{D}S_{\theta_0}^{T_0}(v_{1:T_0}) \xi_{1:T_0}^k = [h_z, h_p, 0, \mathbf{S}_k, h_r]$ (for some h_z, h_p, h_r). If $h_{1:T_0} = \varepsilon_k \xi_{1:T_0}^k \in \mathsf{V}^{T_0}$ for $\varepsilon_k \in \mathbb{R}$, we have by Taylor expansion and linearity of the differential:

$$\hat{\mathbf{\Sigma}}_{T_0}^h = \hat{\mathbf{\Sigma}}_{T_0} + \varepsilon_k \mathbf{S}_k + o(\varepsilon_k) \ .$$

Set then $h_{1:T_0} = \sum_{k=1}^{s-1} \varepsilon_k \xi_{1:T_0}^k$, so that, by linearity of the differential:

$$\hat{\mathbf{\Sigma}}_{T_0}^h = \hat{\mathbf{\Sigma}}_{T_0} + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1}) \ . \tag{2.87}$$

Moreover, by conditions (ii) and (iii) in Lemma 2.12, we get

$$z_{T_0} = q_{T_0} = 0 \quad \text{and} \quad p \coloneqq p_{T_0} \neq 0 \ , \tag{2.88}$$

and by Taylor expansion since $S_{\theta_0}^{T_0}(v_{1:T_0} + h_{1:T_0}) = S_{\theta_0}^{T_0}(v_{1:T_0}) + \sum_{k=1}^{s-1} \varepsilon_k \mathcal{D}S_{\theta_0}^{T_0}(v_{1:T_0}) \xi_{1:T_0}^k + o(\varepsilon_{1:s-1})$

$$z_{T_0}^h = 0 + O(\varepsilon_{1:s-1}) \quad \text{and} \quad q_{T_0}^h = 0 + o(\varepsilon_{1:s-1}) \quad \text{and} \quad p_{T_0}^h = p + O(\varepsilon_{1:s-1}) \ . \tag{2.89}$$

Let $j \in \mathbb{N}$ and set $T = T_0 + j + 5$ and

$$v_{T_0+1:T_0+j+5} = 0 \tag{2.90}$$

Then $v_{T_0+1:T_0+j+5} = 0 \in \overline{\mathcal{O}^{j+5}_{\theta_{T_0}}}$ by Proposition 2.11. Since $z_{T_0} = q_{T_0} = 0$, then we obtain by applying the update equations (2.40) and (2.42), that $z_{T_0:T} = q_{T_0:T} = 0$. By condition (i) in Lemma 2.12, $S^{T_0}_{\theta_0}$ is differentiable at $v_{1:T_0}$. Moreover, for $t = T_0, \dots, T-1$, we have $z_t = q_t = 0$ and $v_{t+1} = 0$, hence by Lemma 2.11 (case a) $S^1_{\theta_t}$ is differentiable at v_{t+1}. By chain rule, $S^T_{\theta_0}$ is differentiable at $v_{1:T}$.
We set $h_{T_0+1:T_0+j} = 0$, then since $\theta^h_{T_0+j} = S^j_{\theta^h_{T_0}}(v_{T_0+1:T_0+j})$ by applying (2.40), (2.41) and (2.42) with $v = 0$ and using (2.89), we have

$$z^h_{T_0+j} = 0 + O(\varepsilon_{1:s-1}) \quad \text{and} \quad q^h_{T_0+j} = 0 + o(\varepsilon_{1:s-1}) \quad \text{and} \quad p^h_{T_0+j} = (1-c_\sigma)^j p + O(\varepsilon_{1:s-1}) \ . \tag{2.91}$$

Moreover, set

$$\begin{cases} h_{T_0+j+1} = (H_1, \dots, H_1) \in \mathbb{R}^{d\mu} & \text{for some } H_1 \in \mathbb{R}^d \\ h_{T_0+j+2} = -(1-c_1-c_\mu)^{-1/2}\Gamma(p_{T_0+j+1})^{-1} h_{T_0+j+1} & \\ h_{T_0+j+3} = (H_3, \dots, H_3) \in \mathbb{R}^{d\mu} & \text{for some } H_3 \in \mathbb{R}^d \\ h_{T_0+j+4} = -(1-c_1-c_\mu)^{-1/2}\Gamma(p_{T_0+j+3})^{-1} h_{T_0+j+3} & \\ h_{T_0+j+5} = (H_5, \dots, H_5) \in \mathbb{R}^{d\mu} & \text{for some } H_5 \in \mathbb{R}^d \end{cases}$$

Note that $p_{T_0+k} = (1-c_\sigma)^k p$ for $k = 1, \dots, j+5$ by (2.41), and by Taylor expansion:

$$p^h_{T_0+j+1} = (1-c_\sigma)^{j+1} p + O(\varepsilon_{1:s-1}, H_1) \ .$$

Yet, Γ is locally Lipschitz by **Γ**1, thus $\Gamma(p^h_{T_0+j+1}) = \Gamma((1-c_\sigma)^{j+1}p) + O(\varepsilon_{1:s-1}, H_1)$. Moreover, since by **R**1, we have $r_{T_0+k} = R((1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_{T_0+k})^{-1}\hat{\mathbf{\Sigma}}_{T_0+k}) = 1-c_1-c_\mu$ and since R is locally Lipschitz by **R**2, we have

$$\begin{aligned} r^h_{T_0+k} &= R((1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_{T_0+k})^{-1}\hat{\mathbf{\Sigma}}_{T_0+k} + O(h_{1:T_0+k})) = 1-c_1-c_\mu+O(h_{1:T_0+k}) \\ &= 1-c_1-c_\mu+O(\varepsilon_{1:s-1}) + O(h_{T_0+1:k}) \ . \end{aligned} \tag{2.92}$$

When $H_1, H_3, H_5, \varepsilon_{1:s-1} \to 0$, since

$$\theta^h_{T_0+j+1} = S^1_{\theta^h_{T_0+j}}(0 + h_{T_0+j+1}) \tag{2.93}$$

by applying (2.40) we find

$$z^h_{T_0+j+1} = \frac{z^h_{T_0+j} + c_m H_1}{\sqrt{r^h_{T_0+j+1}}\Gamma(p^h_{T_0+j+1})}$$

and thus using (2.91) and (2.92):

$$z^h_{T_0+j+1} = c_m(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)^{-1}H_1 + O(\varepsilon_{1:s-1}) + o(H_1)$$

so,

$$\begin{aligned} z^h_{T_0+j+2} &= \frac{z^h_{T_0+j+1} - c_m(1-c_1-c_\mu)^{-1/2}\Gamma(p_{T_0+j+1})^{-1}H_1}{\sqrt{r^h_{T_0+j+2}}\Gamma(p^h_{T_0+j+2})} \\ &= O(\varepsilon_{1:s-1}) + o(H_1) \ . \end{aligned}$$

Likewise,

$$z^h_{T_0+j+4} = O(\varepsilon_{1:s-1}) + o(H_1, H_3),$$

so that, in the end, since R is locally Lipschitz by **R**2 and Γ is locally Lipschitz by **Γ**1, then

$$z^h_T = z^h_{T_0+j+5} = O(\varepsilon_{1:s-1}) + o(H_1, H_3) + c_m r_T^{-1/2}\Gamma(p_T)^{-1}H_5 + o(H_5).$$

Furthermore using (2.93) and (2.91),

$$\begin{aligned} q^h_{T_0+j+1} &= (1-c_c)(r^h_{T_0+j})^{-1/2} q^h_{T_0+j} + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}(0 + H_1) \\ &= \sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_1 + o(\varepsilon_{1:s-1}) \end{aligned}$$

and

$$\begin{aligned} q^h_{T_0+j+2} &= (1-c_c)(r^h_{T_0+j+1})^{-1/2} q^h_{T_0+j+1} \\ &\quad + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}(0 - (1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)^{-1}H_1) \\ &= (1-c_c)(1-c_1-c_\mu + O(\varepsilon_{1:s-1}, H_1))^{-1/2}\sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_1 + o(\varepsilon_{1:s-1}) \\ &\quad + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)^{-1}H_1 \\ &= \underbrace{(1-c_1-c_\mu)^{-1/2}\left[1-c_c-\Gamma((1-c_\sigma)^{j+1}p)^{-1}\right]\sqrt{c_c(2-c_c)\mu_{\text{eff}}}}_{=:d^p_{j+1}} H_1 + o(\varepsilon_{1:s-1}, H_1) \ . \end{aligned}$$

Likewise,

$$\begin{aligned} q^h_{T_0+j+3} &= (r_{T_0+j+2})^{-1/2}(1-c_c)q^h_{T_0+j+2} + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_3 \\ &= (1-c_1-c_\mu)^{-1/2}(1-c_c)d^p_{j+1}H_1 + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_3 + o(\varepsilon_{1:s-1}, H_1) \end{aligned}$$

and

$$\begin{aligned} q^h_{T_0+j+4} &= (r_{T_0+j+3})^{-1/2}(1-c_c)q^h_{T_0+j+3} - \sqrt{c_c(2-c_c)\mu_{\text{eff}}}(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+3})^{-1}H_3 \\ &= (1-c_1-c_\mu)^{-1}(1-c_c)^2 d^p_{j+1}H_1 + (1-c_1-c_\mu)^{-1/2}(1-c_c)\sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_3 \\ &\quad - \sqrt{c_c(2-c_c)\mu_{\text{eff}}}(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+3})^{-1}H_3 + o(\varepsilon_{1:s-1}, H_1, H_3) \\ &= (1-c_1-c_\mu)^{-1}(1-c_c)^2 d^p_{j+1}H_1 + d^p_{j+3}H_3 + o(\varepsilon_{1:s-1}, H_1, H_3) \ , \end{aligned}$$

where $d^p_k := (1-c_1-c_\mu)^{-1/2}\left[1-c_c-\Gamma((1-c_\sigma)^k p)^{-1}\right]\sqrt{c_c(2-c_c)\mu_{\text{eff}}}$ for $k = j+1, j+3$. Then,

$$\begin{aligned} q^h_T = q^h_{T_0+j+5} &= (r^h_{T_0+j+4})^{-1/2}(1-c_c)q^h_{T_0+j+4} + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_5 \\ &= (1-c_1-c_\mu)^{-3/2}(1-c_c)^3 d^p_{j+1}H_1 \\ &\quad + (1-c_1-c_\mu)^{-1/2}(1-c_c)d^p_{j+3}H_3 + O(H_5) + o(\varepsilon_{1:s-1}, H_1, H_3) \ . \end{aligned}$$

For $t = 0, \dots, T$, we denote $\mathbf{\Sigma}_t = R(\hat{\mathbf{\Sigma}}_t)^{-1}\hat{\mathbf{\Sigma}}_t$ and $\mathbf{\Sigma}^h_t = R(\hat{\mathbf{\Sigma}}^h_t)^{-1}\hat{\mathbf{\Sigma}}^h_t$. For $t = T_0, \dots, T-1$, given the choice of $v_{T_0+1:T_0+j+5} = 0$ in (2.90), we have then $\hat{\mathbf{\Sigma}}_{t+1} = (1-c_1-c_\mu)\mathbf{\Sigma}_t/\rho((1-c_1-c_\mu)\mathbf{\Sigma}_t)$ and by **R**1

$$\mathbf{\Sigma}_{t+1} = \frac{\hat{\mathbf{\Sigma}}_{t+1}}{R(\hat{\mathbf{\Sigma}}_{t+1})} = \frac{(1-c_1-c_\mu)\mathbf{\Sigma}_t}{R((1-c_1-c_\mu)\mathbf{\Sigma}_t)} = \mathbf{\Sigma}_t \ .$$

Thus, $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_T$ for $t = T_0, \dots, T$. Moreover, we have for $k = 0, \dots, j$:

$$\hat{\boldsymbol{\Sigma}}^h_{T_0+k+1} = \frac{(1-c_1-c_\mu)\boldsymbol{\Sigma}^h_{T_0+k} + c_1 q^h_{T_0+k+1}(q^h_{T_0+k+1})^\top + c_\mu \sum_{i=1}^\mu w_i^c h^i_{T_0+k+1}(h^i_{T_0+k+1})^\top}{\rho((1-c_1-c_\mu)\boldsymbol{\Sigma}^h_{T_0+k} + c_1 q^h_{T_0+k+1}(q^h_{T_0+k+1})^\top + c_\mu \sum_{i=1}^\mu w_i^c h^i_{T_0+k+1}(h^i_{T_0+k+1})^\top)}$$

Since ρ is $\mathcal{C}^1$ by $\boldsymbol{\rho}2$, hence locally Lipschitz, and $q^h_{T_0+k+1}(q^h_{T_0+k+1})^\top = 0 + O(\|h_{1:T_0+k}\|^2) = o(h_{1:T_0+k})$, we have then:

$$\hat{\boldsymbol{\Sigma}}^h_{T_0+k+1} = \frac{(1-c_1-c_\mu)\boldsymbol{\Sigma}^h_{T_0+k}}{\rho((1-c_1-c_\mu)\boldsymbol{\Sigma}^h_{T_0+k})} + o(h_{1:T_0+k}) = \frac{\boldsymbol{\Sigma}^h_{T_0+k}}{\rho(\boldsymbol{\Sigma}^h_{T_0+k})} + o(h_{1:T_0+k}) = \hat{\boldsymbol{\Sigma}}^h_{T_0+k} + o(h_{1:T_0+k}) \ ,$$

where we have used $\boldsymbol{\rho}1$ to simplify the above equation. Therefore, we obtain by induction and using (2.87) that

$$\hat{\boldsymbol{\Sigma}}^h_{T_0+k} = \hat{\boldsymbol{\Sigma}}_{T_0} + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1}, H_1, H_3, H_5)$$

and thus, using $\boldsymbol{\Sigma}_{T_0+k} = \boldsymbol{\Sigma}_T$, and since R is locally Lipschitz by $\mathbf{R}2$:

$$\begin{aligned}\boldsymbol{\Sigma}^h_{T_0+k} &= \frac{\hat{\boldsymbol{\Sigma}}^h_{T_0+k}}{R(\hat{\boldsymbol{\Sigma}}^h_{T_0+k})} = \frac{\hat{\boldsymbol{\Sigma}}_{T_0} + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1}, H_1, H_3, H_5)}{R(\hat{\boldsymbol{\Sigma}}_{T_0} + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1}, H_1, H_3, H_5))} \\ &= \underbrace{\frac{\hat{\boldsymbol{\Sigma}}_{T_0}}{R(\hat{\boldsymbol{\Sigma}}_{T_0})}}_{=\boldsymbol{\Sigma}_{T_0}} + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5) = \boldsymbol{\Sigma}_T + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5) \ . \quad (2.94)\end{aligned}$$

It follows that:

$$\begin{aligned}p^h_{T_0+j+1} &= (1-c_\sigma)p^h_{T_0+j} + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(\boldsymbol{\Sigma}^h_{T_0+j})^{-1/2} \times (0 + H_1) \\ &= (1-c_\sigma)^{j+1}p + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\boldsymbol{\Sigma}_T^{-1/2}H_1 + O(\varepsilon_{1:s-1}) + o(H_1)\end{aligned}$$

and

$$\begin{aligned}p^h_{T_0+j+2} &= (1-c_\sigma)p^h_{T_0+j+1} \\ &\quad + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(\boldsymbol{\Sigma}^h_{T_0+j+1})^{-1/2} \times (0 - (1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)^{-1}H_1) \\ &= (1-c_\sigma)^{j+1}p + (1-c_\sigma)\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\boldsymbol{\Sigma}_T^{-1/2}H_1 \\ &\quad - \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)^{-1}\boldsymbol{\Sigma}_T^{-1/2}H_1 \\ &\quad + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5) \\ &= (1-c_\sigma)^{j+2}p + c^p_{j+1}\boldsymbol{\Sigma}_T^{-1/2}H_1 + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H5) \ .\end{aligned}$$

Likewise,

$$\begin{aligned}p^h_{T_0+j+3} &= (1-c_\sigma)p^h_{T_0+j+2} + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(\boldsymbol{\Sigma}^h_{T_0+j+2})^{-1/2} \times (0 + H_3) \\ &= (1-c_\sigma)^{j+3}p + (1-c_\sigma)c^p_{j+1}\boldsymbol{\Sigma}_T^{-1/2}H_1 \\ &\quad + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\boldsymbol{\Sigma}_T^{-1/2}H_3 + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5)\end{aligned}$$

and

$$
\begin{aligned}
p^h_{T_0+j+4} &= (1-c_\sigma)p^h_{T_0+j+3} \\
&\quad + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(\mathbf{\Sigma}^h_{T_0+j+3})^{-1/2} \times (0 - (1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+3}p)^{-1}H_3) \\
&= (1-c_\sigma)^{j+4}p + (1-c_\sigma)^2 c^p_{j+1}\mathbf{\Sigma}_T^{-1/2}H_1 + (1-c_\sigma)\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbf{\Sigma}_T^{-1/2}H_3 \\
&\quad - \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+3}p)^{-1}\mathbf{\Sigma}_T^{-1/2}H_3 \\
&\quad + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5) \\
&= (1-c_\sigma)^{j+4}p + (1-c_\sigma)^2 c^p_{j+1}\mathbf{\Sigma}_T^{-1/2}H_1 + c^p_{j+3}\mathbf{\Sigma}_T^{-1/2}H_3 + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5)
\end{aligned}
$$

where $c^p_k \coloneqq (1-c_\sigma-(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^k p))^{-1}\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}$ for $k = j+1, j+3$. Finally,

$$
\begin{aligned}
p^h_{T_0+j+5} &= (1-c_\sigma)p^h_{T_0+j+4} + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(\mathbf{\Sigma}^h_{T_0+j+4})^{-1/2} \times (0 + H_5) \\
&= (1-c_\sigma)^{j+5}p + (1-c_\sigma)^3 c^p_{j+1}\mathbf{\Sigma}_T^{-1/2}H_1 \\
&\quad + (1-c_\sigma)c^p_{j+3}\mathbf{\Sigma}_T^{-1/2}H_3 + O(\varepsilon_{1:s-1}) + o(H_1, H_3) + O(H_5) \ .
\end{aligned}
$$

By **R**2, we have

$$
\begin{aligned}
r^h_T &= R\left((1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}^h_{T-1})^{-1}\hat{\mathbf{\Sigma}}^h_{T-1} + o(H_1, H_3, H_5)\right) \\
&= (1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}^h_{T-1})^{-1}R(\hat{\mathbf{\Sigma}}^h_{T-1}) + o(H_1, H_3, H_5) \\
&= 1-c_1-c_\mu + o(H_1, H_3, H_5) = r_T + o(H_1, H_3, H_5) \ ,
\end{aligned}
$$

where we have used **R**1 to simplify the first line into the second line in the above equation. All in all, when $H_1, H_3, H_5, \varepsilon_{1:s-1} \to 0$,

$$
\begin{aligned}
\theta^h_T &= \theta_T + \begin{bmatrix} O(\varepsilon_{1:s-1}) \\ O(\varepsilon_{1:s-1}) \\ 0 \\ \sum_{t=1}^{s-1}\varepsilon_t \mathbf{S}_t \\ 0 \end{bmatrix} + o(\varepsilon_{1:s-1}, H_1, H_3, H_5) \\
&\quad + \begin{bmatrix} (1-c_1-c_\mu)^{-1/2}\Gamma(p_T)^{-1}c_m H_5 \\ (1-c_\sigma) \times \left[(1-c_\sigma)^2 c^p_{j+1}\mathbf{\Sigma}_T^{-1/2}H_1 + c^p_{j+3}\mathbf{\Sigma}_T^{-1/2}H_3\right] + O(H_5) \\ (1-c_c)(1-c_1-c_\mu)^{-1/2} \times \left[(1-c_c)^2(1-c_1-c_\mu)^{-1}d^p_{j+1}H_1 + d^p_{j+3}H_3\right] + O(H_5) \\ 0 \\ 0 \end{bmatrix} \ .
\end{aligned}
$$

We identify the Taylor expansion of $S^T_{\theta_0}(v_{1:T} + h_{1:T})$ in (2.55), with $L_z = (1-c_1-c_\mu)^{-1/2}\Gamma(p_T)^{-1}c_m$ and $\mathbf{L}^{\mathbf{\Sigma}}_k = \mathbf{S}_k$ for $k = 1, \ldots, s-1$ and

$$
\begin{aligned}
\mathsf{W}_L = \text{span}(\xi^1_{1:T_0}, \ldots, \xi^{s-1}_{1:T_0}) \times \{0\}^j \times \begin{pmatrix} 1 \\ -(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p) \end{pmatrix}^\top \mathbb{R}^{d\mu} \\
\times \begin{pmatrix} 1 \\ -(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+3}p) \end{pmatrix}^\top \mathbb{R}^{d\mu} \times \mathbb{R}^{d\mu}
\end{aligned}
$$

and $L\colon \mathsf{W}_L \to \mathbb{R}^{s-1} \times (\mathbb{R}^d)^3$ maps a vector $h_{1:T} \in \mathsf{W}_L$ to a vector $(\varepsilon_{1:s-1}, H_1, H_3, H_5) \in \mathbb{R}^{s-1} \times (\mathbb{R}^d)^3$ such that

$$h_{1:s-1} = \sum_{k=1}^{s-1} \varepsilon_k \xi^k_{1:T_0}; h_{s+j:s+j+1} = \begin{pmatrix} H_1 \\ -(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)H_1 \end{pmatrix}$$

$$h_{s+j+2:s+j+3} = \begin{pmatrix} H_3 \\ -(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)H_3 \end{pmatrix}; h_{s+j+4} = H_5 \ .$$

Since the scalars $\varepsilon_1, \dots, \varepsilon_{s-1} \in \mathbb{R}$ and the vectors $H_1, H_3, H_5 \in \mathbb{R}^d$ above can be chosen arbitrary and independently of each other, the linear application $L\colon \mathsf{W}_L \to \mathbb{R}^{s-1} \times (\mathbb{R}^d)^3$ is surjective. □

B.3 *Proof of Proposition 2.12*

Proof of Proposition 2.12(a) and (c). Suppose that either $c_c \neq 1$, $c_\sigma \neq 1$ and $1 - c_c \neq (1-c_\sigma)\sqrt{1-c_1-c_\mu}$, or that $c_c \neq 1$, $c_\sigma = 1$. Assume moreover that $c_\mu > 0$. Apply then Lemmas 2.13 and 2.14 to get that there exist $T \in \mathbb{N}$ and $v_{1:T} \in \overline{\mathcal{O}^T_{\theta_0}}$ with

(a) in the case $c_\sigma \neq 1$, $\operatorname{rge} \mathcal{D}S^T_{\theta_0}(v_{1:T}) \supset \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathrm{T}_{\mathbf{\Sigma}_T}\rho^{-1}(\{1\}) \times \{0\}$;

(c) in the case $c_\sigma = 1$, $\operatorname{rge} \mathcal{D}S^T_{\theta_0}(v_{1:T}) \supset \mathbb{R}^d \times \{0\} \times \mathbb{R}^d \times \mathrm{T}_{\mathbf{\Sigma}_T}\rho^{-1}(\{1\}) \times \{0\}$.

In both cases, consider arbitrary $h_z, h_p, h_q \in \mathbb{R}^d$, $\mathbf{H}_\mathbf{\Sigma} \in T_{\mathbf{\Sigma}_T}\rho^{-1}(\{1\})$, with $h_p = 0$ if $c_\sigma = 1$, so that there exists $h_{1:T} \in \mathsf{V}^T$ satisfying $\mathcal{D}S^T_{\theta_0}(v_{1:T})h_{1:T} = (h_z, h_p, h_q, \mathbf{H}_\mathbf{\Sigma}, 0)^\top$. By Taylor expansion, we have then

$$S^T_{\theta_0}(v_{1:T} + h_{1:T}) = \begin{bmatrix} z_T + h_z \\ p_T + h_p \\ q_T + h_q \\ \hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma} \\ r_T \end{bmatrix} + o(h_{1:T}) \ .$$

Since R is positive and positively homogeneous, it is not constant around $\hat{\mathbf{\Sigma}}_T$. Besides, R is differentiable at $\hat{\mathbf{\Sigma}}_T$ and thus there exists $w \in \mathbb{R}^d$ such that $\mathcal{D}R(\hat{\mathbf{\Sigma}}_T)(ww^\top) \neq 0$. Consider the nonconstant smooth function

$$G_w\colon s \in \mathbb{R} \mapsto F_\mathbf{\Sigma}(q_t, R(\hat{\mathbf{\Sigma}}_t)^{-1}\hat{\mathbf{\Sigma}}_t, r_t; s[w, \dots, w]) \ ,$$

see (2.49). Since R is locally Lipschitz on $\mathcal{S}^d_{++}$, then R is locally Lipschitz on the submanifold $\operatorname{rge} G_w$, which is nontrivial since G_w is nonconstant. Then, by Rademacher's theorem [52, Corollary B.5], R is differentiable at $G_w(s)$ for almost every $s \in \mathbb{R}$. Moreover, we know that $\hat{\mathbf{\Sigma}}_T = G_w(0)$ and that $\mathcal{D}R(\hat{\mathbf{\Sigma}}_T)(ww^\top) \neq 0$. Thus, by upper semicontinuity of Clarke's Jacobian [52, Proposition B.9], there exists a sufficiently small $s > 0$ such that $\mathcal{D}R(G_w(s))(ww^\top) \neq 0$.
Then, we can find $\epsilon = sw \in \mathbb{R}^d$ a nonzero vector small enough so that, if $v_{T+1} = [\epsilon, \dots, \epsilon] \in \overline{\mathcal{O}^1_{\theta_T}}$, then R is differentiable at $\hat{\mathbf{\Sigma}}_{T+1}$, and $\mathcal{D}R(\hat{\mathbf{\Sigma}}_{T+1})(\epsilon\epsilon^\top) \neq 0$. Moreover, up to taking $s > 0$

smaller, we can assume that Γ is differentiable at $p_{T+1} \neq 0$ by $\mathbf{\Gamma}1$. Hence by composition and by Lemma 2.11, $S_{\theta_0}^{T+1}$ is differentiable at $v_{1:T+1}$. Indeed, by chain rule [30, Corollary 2.6.6], we have

$$\mathcal{D}S_{\theta_0}^{T+1}(v_{1:T+1})h_{1:T+1} = \mathcal{D}F(S_{\theta_0}^{T}(v_{1:T}), v_{T+1})\left(\mathcal{D}S_{\theta_0}^{T}(v_{1:T})(h_{1:T}), h_{T+1}\right)$$

Let $h_{T+1} = [h, \dots, h] \in \mathbb{R}^{d\mu}$ for some arbitrary $h \in \mathbb{R}^d$. Then,

$$\begin{aligned}
&S_{\theta_0}^{T+1}(v_{1:T+1} + h_{1:T+1}) \\
&= \begin{bmatrix} F_z(z_T + h_z, p_T + h_p, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}}), r_T; [\epsilon + h, \dots, \epsilon + h]) \\ F_p(p_T + h_p, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}}); [\epsilon + h, \dots, \epsilon + h]) \\ F_q(q_T + h_q, r_T; [\epsilon + h, \dots, \epsilon + h]) \\ F_{\mathbf{\Sigma}}(q_T + h_q, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}}), r_T; [\epsilon + h, \dots, \epsilon + h]) \\ F_r(q_T + h_q, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}}), r_T; [\epsilon + h, \dots, \epsilon + h]) \end{bmatrix} + o(h_{1:T+1})
\end{aligned}$$

see (2.46)-(2.50). Moreover, we have

$$\begin{aligned}
&F_z(z_T + h_z, p_T + h_p, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}}), r_T; [\epsilon + h, \dots, \epsilon + h]) \\
&\quad = \frac{z_T + h_z + c_m(h + \epsilon)}{r_{T+1}^{1/2}\Gamma(p_{T+1}) + O(h, h_p, h_q, \mathbf{H}_{\mathbf{\Sigma}})} + o(h_{1:T+1}) \\
&\quad = F_z(z_T, p_T, R(\hat{\mathbf{\Sigma}}_T)^{-1}(\hat{\mathbf{\Sigma}}_T), r_T; [\epsilon, \dots, \epsilon]) + \frac{h_z}{r_{T+1}^{1/2}\Gamma(p_{T+1})} + O(h, h_p, h_q, \mathbf{H}_{\mathbf{\Sigma}}) + o(h_{1:T+1}) \ ,
\end{aligned}$$

$$\begin{aligned}
&F_p(p_T + h_p, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}}); [\epsilon + h, \dots, \epsilon + h]) \\
&\quad = (1 - c_\sigma)(p_T + h_p) + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}}}R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{1/2}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{-1/2}(\epsilon + h) + o(h_{1:T+1}) \\
&\quad = F_p(p_T, R(\hat{\mathbf{\Sigma}}_T)^{-1}(\hat{\mathbf{\Sigma}}_T); [\epsilon, \dots, \epsilon]) + (1 - c_\sigma)h_p + O(h, \mathbf{H}_{\mathbf{\Sigma}}) + o(h_{1:T+1}) \ ,
\end{aligned}$$

$$\begin{aligned}
&F_q(q_T + h_q, r_T; [\epsilon + h, \dots, \epsilon + h]) \\
&\quad = r_T^{-1/2}(1 - c_c)(q_T + h_q) + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}}(\epsilon + h) + o(h_{1:T+1}) \\
&\quad = F_q(q_T, r_T; [\epsilon, \dots, \epsilon]) + (1 - c_c)r_T^{-1/2}h_q + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}}h + o(h_{1:T+1}) \\
&\quad = q_{T+1} + h_q^+ + o(h_{1:T+1}) \ ,
\end{aligned}$$

where $h_q^+ = (1 - c_c)r_T^{-1/2}h_q + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}}h$,

$$\begin{aligned}
&F_r(q_T + h_q, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}}), r_T; [\epsilon + h, \dots, \epsilon + h]) \\
&\quad = R\left((1 - c_1 - c_\mu)\frac{\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}}}{R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})} + c_1[q_{T+1} + h_q^+][q_{T+1} + h_q^+]^\top + c_\mu[\epsilon + h][\epsilon + h]^\top\right) \\
&\qquad + o(h_{1:T+1}) \\
&\quad = R(\tilde{\mathbf{\Sigma}}_{T+1}) + \mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})[(1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{-1}\mathbf{H}_{\mathbf{\Sigma}}] \\
&\qquad + \mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})[c_1[q_{T+1}(h_q^+)^\top + h_q^+ q_{T+1}^\top]] + \mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})[c_\mu[\epsilon h^\top + h\epsilon^\top]] + o(h_{1:T+1}) \\
&\quad = F_r(q_T, R(\hat{\mathbf{\Sigma}}_T)^{-1}(\hat{\mathbf{\Sigma}}_T), r_T; [\epsilon, \dots, \epsilon]) + (1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}}_T)^{-1}\mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})\mathbf{H}_{\mathbf{\Sigma}} \\
&\qquad + c_1\mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})[q_{T+1}(h_q^+)^\top + h_q^+ q_{T+1}^\top] + c_\mu\mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})[\epsilon h^\top + h\epsilon^\top] + o(h_{1:T+1}) \\
&\quad = F_r(q_T, R(\hat{\mathbf{\Sigma}}_T)^{-1}(\hat{\mathbf{\Sigma}}_T), r_T; [\epsilon, \dots, \epsilon]) + \mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})\tilde{\mathbf{H}}_{\mathbf{\Sigma}}^+ + o(h_{1:T+1}) \ ,
\end{aligned}$$

where

$$\tilde{\mathbf{H}}^+_{\mathbf{\Sigma}} = (1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_T)^{-1}\mathbf{H}_{\mathbf{\Sigma}} + c_1[q_{T+1}(h_q^+)^\top + h_q^+ q_{T+1}^\top] + c_\mu[\epsilon h^\top + h\epsilon^\top]$$

and

$$\tilde{\mathbf{\Sigma}}_{T+1} = (1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_T)^{-1}(\hat{\mathbf{\Sigma}}_T) + c_1 q_{T+1}q_{T+1}^\top + c_\mu \epsilon\epsilon^\top \ .$$

Let $(x,y) \in \mathbb{R}^d \times \mathbb{R}$. Since $c_\mu > 0$ and $\mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})(\epsilon\epsilon^\top) \neq 0$ (since as seen above $\mathcal{D}R(\hat{\mathbf{\Sigma}}_{T+1})(\epsilon\epsilon^\top) \neq 0$ and $\hat{\mathbf{\Sigma}}_{T+1}$ is proportional to $\tilde{\mathbf{\Sigma}}_{T+1}$, see Lemma 2.10), there exists $h = l\varepsilon$, where

$$l = \frac{y - c_1\mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})[q_{T+1}x^\top + xq_{T+1}^\top] - (1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_T)\mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})\mathbf{H}_{\mathbf{\Sigma}}}{2c_\mu \mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})(\epsilon\epsilon^\top)} \ ,$$

and $h_q = (1-c_c)^{-1}r_T^{1/2}(x - \sqrt{c_c(2-c_c)\mu_{\text{eff}}}l\epsilon)$ such that $h_q^+ = x$ and

$$\begin{aligned}&\mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})\tilde{\mathbf{H}}^+_{\mathbf{\Sigma}} \\ &= (1-c_1-c_\mu)DR(\tilde{\mathbf{\Sigma}}_{T+1})R(\hat{\mathbf{\Sigma}}_T)^{-1}\mathbf{H}_{\mathbf{\Sigma}} + c_1 DR(\tilde{\mathbf{\Sigma}}_{T+1})[q_{T+1}x^\top + xq_{T+1}^\top] + 2lc_\mu[\epsilon\epsilon^\top] = y\end{aligned}$$

Therefore, the linear map $(h_q, h) \mapsto (h_q^+, \mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})\tilde{\mathbf{H}}^+_{\mathbf{\Sigma}})$ valued in $\mathbb{R}^d \times \mathbb{R}$ is surjective. Besides,

$$\hat{\mathbf{\Sigma}}^h_{T+1} \coloneqq F_{\mathbf{\Sigma}}(q_T + h_q, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}}), r_T; [\epsilon + h, \dots, \epsilon + h]) = \frac{\tilde{\mathbf{\Sigma}}^h_{T+1}}{\rho(\tilde{\mathbf{\Sigma}}^h_{T+1})} \ ,$$

where

$$\begin{aligned}\tilde{\mathbf{\Sigma}}^h_{T+1} = \tilde{\mathbf{\Sigma}}_{T+1} &+ (1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_T)^{-1}\mathbf{H}_{\mathbf{\Sigma}} + c_1[q_{T+1}(h_q^+)^\top + h_q^+ q_{T+1}^\top] \\ &+ c_\mu[\epsilon h^\top + h\epsilon^\top] + o(h_{1:T+1}) = \tilde{\mathbf{\Sigma}}_{T+1} + \tilde{\mathbf{H}}^+_{\mathbf{\Sigma}} + o(h_{1:T+1}) \ .\end{aligned}$$

Therefore, by using the Taylor expansion $\rho(\tilde{\mathbf{\Sigma}}^h_{T+1})^{-1} = \rho(\tilde{\mathbf{\Sigma}}_{T+1})^{-1} - \mathcal{D}\rho(\tilde{\mathbf{\Sigma}}_{T+1})\tilde{\mathbf{H}}^+_{\mathbf{\Sigma}} + o(h_{1:T+1})$ since ρ is positive and continuously differentiable by $\boldsymbol{\rho}2$, we get

$$\begin{aligned}&F_{\mathbf{\Sigma}}(q_T + h_q, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_{\mathbf{\Sigma}}), r_T; [\epsilon + h, \dots, \epsilon + h]) \\ &= F_{\mathbf{\Sigma}}(q_T, R(\hat{\mathbf{\Sigma}}_T)^{-1}(\hat{\mathbf{\Sigma}}_T), r_T; [\epsilon, \dots, \epsilon]) + \rho(\tilde{\mathbf{\Sigma}}_{T+1})^{-1}\tilde{\mathbf{H}}^+_{\mathbf{\Sigma}} - (\mathcal{D}\rho(\tilde{\mathbf{\Sigma}}_{T+1})\tilde{\mathbf{H}}^+_{\mathbf{\Sigma}})\tilde{\mathbf{\Sigma}}_{T+1} + o(h_{1:T+1}) \ .\end{aligned}$$

All in all, by Taylor expansion,

$$\mathcal{D}S^{T+1}_{\theta_0}(v_{1:T+1})h_{1:T+1} = \begin{bmatrix} r_{T+1}^{-1/2}\Gamma(p_{T+1})^{-1}h_z + O(h, h_p, h_q, \mathbf{H}_{\mathbf{\Sigma}}) \\ (1-c_\sigma)h_p + O(h, \mathbf{H}_{\mathbf{\Sigma}}) \\ h_q^+ \\ \rho(\tilde{\mathbf{\Sigma}}_{T+1})^{-1}\tilde{\mathbf{H}}^+_{\mathbf{\Sigma}} - (\mathcal{D}\rho(\tilde{\mathbf{\Sigma}}_{T+1})\tilde{\mathbf{H}}^+_{\mathbf{\Sigma}})\tilde{\mathbf{\Sigma}}_{T+1} \\ \mathcal{D}R(\tilde{\mathbf{\Sigma}}_{T+1})\tilde{\mathbf{H}}^+_{\mathbf{\Sigma}} \end{bmatrix} \ ,$$

which proves that every element in $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_{T+1}}\rho^{-1}(\{1\}) \times \mathbb{R}$ is reached by the linear map $\mathcal{D}S^{T+1}_{\theta_0}(v_{1:T+1})$ when $c_\sigma \neq 1$ so $\mathcal{D}S^{T+1}_{\theta_0}(v_{1:T+1})$ is surjective, hence of maximal rank, which proves the statement (a) (with $T+1$ instead of T). When $c_\sigma = 1$, the statement (c) follows as well as there exists $p = 0 \in \mathbb{R}^d$ such that for every $(z, q, \mathbf{\Sigma}, r) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_{T+1}}\rho^{-1}(\{1\}) \times \mathbb{R}$, $(z, p, q, \mathbf{\Sigma}, r)$ belongs to the range of $\mathcal{D}S^{T+1}_{\theta_0}(v_{1:T+1})$. □

Chapter 3

Asymptotic estimations of a perturbed symmetric eigenproblem

Comments on Chapter 3: This chapter is constituted of the paper "Asymptotic estimations of a perturbed symmetric eigenproblem" (Armand Gissler, Anne Auger, Nikolaus Hansen) 2023 in the journal *Applied mathematics letters* in 2023 [51]. The main result of this chapter is Theorem 3.1. Given a symmetric positive definite matrix $\mathbf{B}$, it provides an upper bound on the projection of the eigenvectors of a specific perturbation $\mathbf{A}^{(m)}$ of $\mathbf{B}$ on the eigenvectors of $\mathbf{B}$. The perturbed system $\mathbf{A}^{(m)}$ is given by (P_m) below, and is useful in our context when $\mathbf{B}$ is the covariance matrix $\mathbf{C}_t$ of CMA-ES at iteration t and $\mathbf{A}^{(m)}$ equals the covariance matrix $\mathbf{C}_{t+1}$ at iteration $t+1$. This result is used later in Chapter 4 to prove Proposition , which provides an upper bound of the expected maximum eigenvalue of the updated covariance matrix in function of the maximum eigenvalue of the initial covariance matrix, when it is highly ill-conditioned. This is a major step in the proof for a state-dependent drift condition which yields to the geometric ergodicity of the normalized chain underlying CMA-ES introduced in Chapter 2.

Abstract

We study ill-conditioned positive definite matrices that are disturbed by the sum of m rank-one matrices of a specific form. We provide estimates for the eigenvalues and eigenvectors. When the condition number of the initial matrix tends to infinity, we bound the values of the coordinates of the eigenvectors of the perturbed matrix. Equivalently, in the coordinate system where the initial matrix is diagonal, we bound the rate of convergence of coordinates that tend to zero.

1 Introduction 108
2 Bounds on the eigenvalues of (P_m) 109
3 Estimating the eigenvectors of (P_m) 110
3.1 Rank-one perturbation 110
3.2 Sum of m rank-one matrices pertubation 112
3.3 Thightness 113

1 Introduction

Given a $d \times d$ symmetric matrix with known eigenvectors and eigenvalues denoted $\mathbf{B}$, and a rank-one matrix $vv^\top$ where $v \in \mathbb{R}^d$, eigenvalues and eigenvectors of matrices of the form

$$\mathbf{A} = \mathbf{B} + vv^\top \tag{P_1}$$

have been widely studied, notably in the context of perturbation theory. For instance, the eigenvalues of (P_1) can be estimated and a formula for the eigenvectors is known [57, 27, 84]. Specifically, if $\mathbf{B}$ is a diagonal matrix $\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ with distinct eigenvalues and v has only nonzero entries, then the component j of the unit eigenvector associated to eigenvalue ν_i of the updated matrix $\mathbf{A}$ satisfies the so-called Bunch-Nielsen-Sorensen formula

$$C_i \times \frac{[v]_j}{\lambda_j - \nu_i} \quad \text{for } i, j = 1, \ldots, d \tag{3.1}$$

where C_i is a nonzero normalization constant and $[\,.\,]_j$ denotes the j-th coordinate, a notation we will continue to use in the sequel. Several results have been established for additive perturbations of rank 1 [38, 18] and of higher rank [138, 105]. Symmetric and nonsymmetric perturbation eigenvalue problems have been studied [20] as well as perturbation results for invariant subspaces [94]. In this paper, we provide relative perturbation bounds for the eigenvectors of positive definite matrices. In contrast to previous relative perturbation results for eigenvalues [85] and invariant subspaces [143, 86], the bounds in our result depend on the eigenvalues of the initial matrix $\mathbf{B}$ rather than the norm of the perturbation, see Eq. (3.2) below.

Specifically, we consider the perturbation with a sum of m rank-one matrices of the form

$$\mathbf{A}^{(m)} = \mathbf{B} + \sqrt{\mathbf{B}} \sum_{i=1}^{m} [v^{(i)}][v^{(i)}]^\top \sqrt{\mathbf{B}} \tag{P_m}$$

with $\mathbf{B} = \mathbf{P} \operatorname{diag}(\lambda_1, \ldots, \lambda_d)\mathbf{P}^\top$ where $\mathbf{P}$ is an orthogonal matrix, $\lambda_1 \geqslant \cdots \geqslant \lambda_d > 0$ are the eigenvalues of $\mathbf{B}$, and $v^{(1)}, \ldots, v^{(m)} \in \mathbb{R}^d$. The square root $\sqrt{\mathbf{B}} := \mathbf{P} \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d})\mathbf{P}^\top$ is defined as the unique symmetric positive definite matrix such that $\sqrt{\mathbf{B}} \times \sqrt{\mathbf{B}} = \mathbf{B}$, see e.g. [79, Theorem 7.2.6]. Matrices of the form (P_m) are used in various applications in different domains. For instance, low rank updates of covariance matrices are used in stochastic optimization [97, 73], system identification [102, p. 369], and adaptive Markov Chain Monte Carlo methods [61]. Our motivation is to study the eigenvectors of $\mathbf{A}$ in (P_m), denoted as $e_i^{(m)}$ in the sequel, *when the matrix* $\mathbf{B}$ *is highly ill-conditioned*. When $d = 2$ and $m = 1$ we can compute the eigenvectors of $\mathbf{A}^{(1)}$ explicitly. As an example, consider $\mathbf{B} = \operatorname{diag}(\lambda_1, 1)$ where $\lambda_1 > 1$ and $v^{(1)} = [1, 1]^\top$. Then, the unit eigenvector associated to the largest eigenvalue of $\mathbf{A}^{(1)}$ obeys $\sqrt{1+s^2} \times e_1^{(1)} = [1, s]^\top$ with $s = \lambda_1^{1/2} \times (1 - \lambda_1^{-1} - \sqrt{1 - \lambda_1^{-1} + \lambda_1^{-2}}) = -\lambda_1^{-1/2}/2 + \mathcal{O}(\lambda_1^{-3/2})$ and hence $[e_1^{(1)}]_2 = \lambda_1^{-1/2} + o(\lambda_1^{-1/2})$ when $\lambda_1 \to \infty$. Hence, the (second) coordinate of the (first) unit eigenvector of $\mathbf{A}^{(1)}$ vanishes like $1/\sqrt{\lambda_1}$ when $\lambda_1 \to \infty$. In this paper, we generalize this result to the case where $d \geqslant 2$ and $m \geqslant 1$, as summarized in the following theorem which directly follows from Theorem 3.2 below.

Theorem 3.1. If $e_i^{(m)}$ is a unit eigenvector corresponding to the i-th largest eigenvalue (counted with multiplicity) of $\mathbf{A}^{(m)}$ in (P_m) and $e_j^{(0)}$ is a unit eigenvector corresponding to the j-th largest

eigenvalue λ_j of $\mathbf{B}$, then

$$\left|\left\langle e_i^{(m)}, e_j^{(0)}\right\rangle\right| \leqslant C_m \times \sqrt{\frac{\min\{\lambda_i, \lambda_j\}}{\max\{\lambda_i, \lambda_j\}}} \tag{3.2}$$

where $C_m > 0$ is a constant which depends polynomially on d and $\max_{k=1,\dots,m} \|v^{(k)}\|$.

When $\mathbf{B}$ is diagonal, $e_j^{(0)}$ is the j-th canonical unit vector. Hence $|\langle e_i^{(m)}, e_j^{(0)}\rangle| = [e_i^{(m)}]_j$ and the theorem implies in particular that the j-th coordinate of $e_i^{(m)}$ converges to zero at least as fast as $\sqrt{\min\{\lambda_i, \lambda_j\}/\max\{\lambda_i, \lambda_j\}}$ when the latter tends to 0 (which is tight in the above example when $d = 2$ and $m = 1$), thereby limiting the change of the angle between these eigenvectors. Considerations on the angle between eigenspaces have been made previously [36], however matrices on the form of (P_m) have not been studied in this context. In the remainder, we always choose w.l.o.g. the coordinate system where the matrix $\mathbf{B}$ of (P_m) is diagonal and has decreasingly ordered diagonal values.

This inequality is crucial to study the stability of a Markov chain underlying the CMA-ES algorithm [76, 73]. Proofs of linear convergence for Evolutionary Strategies (ES) rely on a drift condition [110, Theorem 17.0.1] to prove the ergodicity of an underlying Markov chain, see e.g. [17, 141]. To apply this approach to CMA-ES, a potential function is defined on the state-space of this Markov chain and its expected decrease is proven outside a compact set. The state space includes a covariance matrix, updated as

$$\mathbf{C}_{t+1} = (1-c)\mathbf{C}_t + c\sqrt{\mathbf{C}_t}\sum_{i=1}^{m} w_i U_i U_i^\top \sqrt{\mathbf{C}_t} \ , \tag{3.3}$$

where $c \in [0, 1]$, $w_1, \dots, w_m$ are positive weights that sum to 1, and the vectors U_i, $i = 1, \dots, m$, are Gaussian vectors ranked according to a fitness function [73, Eq. (11)]. Hence, (P_m) encompasses the update of this covariance matrix. Eq. (3.2) is needed to bound the expected condition number of the updated covariance matrix, since it controls the influence of small eigenvalues on the growth of the largest eigenvalues.

This paper is organized as follows. In Section 2, we study the eigenvalues of (P_m). In Section 3, we provide bounds for the coordinates of the eigenvectors using Eq. (3.1), and provide an empirical result suggesting that these bounds are tight.

2 Bounds on the eigenvalues of (P_m)

The Bunch-Nielsen-Sorensen formula (3.1) which we will use in Section 3 requires the eigenvalues of the *updated* matrix. Thus, we first derive bounds on the (decreasingly ordered) eigenvalues

$$\lambda_i(\mathbf{A}^{(m)}) = \max_{V\subset\mathbb{R}^d, \dim V = i} \min_{v\in V, v\neq 0} \frac{v^\top \mathbf{A}^{(m)} v}{v^\top v} = \min_{V\subset\mathbb{R}^d, \dim V = d-i+1} \max_{v\in V, v\neq 0} \frac{v^\top \mathbf{A}^{(m)} v}{v^\top v} \quad \text{for } i = 1, \dots, d \tag{3.4}$$

where the equalities ensue from the min-max principle and from Gersgorin's circle theorem [79, Theorems 4.2.6 and 6.1.1].

Theorem 2.7 in [85] and Theorem 2.1 in [135, p. 175] provide an estimation for the eigenvalues of (P_m):

$$\lambda_i \leqslant \nu_i \leqslant \lambda_i \times \left(1 + md \times \max_{k=1,\dots,m} \|v^{(k)}\|_\infty^2\right) \quad \text{for } i \in \{1, \dots, d\} \ . \tag{3.5}$$

The next lemma provides a slightly tighter upper bound on these eigenvalues after a single rank-one pertubation ($m = 1$) and is used in Proposition 3.1.

Lemma 3.1. Let $\mathbf{D} = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ be a diagonal matrix with $\lambda_1 > \cdots > \lambda_d > 0$. Let $v \in \mathbb{R}^d_{\neq 0}$ be a vector with only nonzero entries. Let $\mathbf{A} = \mathbf{D} + \sqrt{\mathbf{D}} v v^\top \sqrt{\mathbf{D}}$ and $\nu_1 \geqslant \nu_2 \geqslant \cdots \geqslant \nu_d$ denote the eigenvalues of $\mathbf{A}$. Then,

$$\nu_i \leqslant \lambda_i \times (1 + (d - i + 1)\|v\|_\infty |[v]_{j_i}|) \quad \text{for all } i \in \{1, \dots, d\} \tag{3.6}$$

where $j_i \in \mathrm{Arg\,max}_{j=i,\dots,d} \left\{ |[v]_j| \;\middle|\; \lambda_j \geqslant \lambda_i \times \left(1 - \sqrt{\tfrac{\lambda_j}{\lambda_i}}(d - i + 1)\|v\|_\infty |[v]_j|\right) \right\}$.

Proof. Fix $i \in \{1, \dots, d\}$ and remark that, by Eq. (3.4), we have $\nu_i \leqslant \max_{v \in \bar{V}_i, \|v\|=1} v^\top \mathbf{A} v = \lambda_1 \left([\mathbf{A}]_{i:d,i:d}\right)$, where $\bar{V}_i = \mathrm{Vect}(e_i, \dots, e_d)$ with e_i being the i^{th} vector of the standard basis of $\mathbb{R}^d$, and with $[\mathbf{A}]_{i:d,i:d}$ denoting the submatrix of $\mathbf{A}$ from rows and columns with indices between i and d included. But, by [79, Theorem 6.1.1], we also have that

$$\lambda_1 \left([\mathbf{A}]_{i:d,i:d}\right) \leqslant \max_{j=i,\dots,d} \left(\sum_{k=i}^{d} |[\mathbf{A}]_{j,k}| \right) =: \max_{j \geqslant i} B_j.$$

Since $\mathbf{A} = \mathbf{D} + \sqrt{\mathbf{D}} v v^\top \sqrt{\mathbf{D}}$, then $|[\mathbf{A}]_{j,k}| \leqslant \sqrt{\lambda_j \lambda_k}(\mathbb{1}\{j = k\} + \|v\|_\infty |[v]_j|)$. If $j \geqslant i$ is such that $\lambda_j \geqslant \lambda_i \times (1 - \sqrt{\lambda_j/\lambda_i}(d - i + 1)\|v\|_\infty |[v]_j|)$, then by definition of j_i we have then $|[v]_j| \leqslant |[v]_{j_i}|$, yielding to $B_j \leqslant \lambda_i \times (1 + (d - i + 1)\|v\|_\infty |[v]_{j_i}|)$. Any other $j \geqslant i$ satisfies $\lambda_j < \lambda_i - \sqrt{\lambda_j \lambda_i}(d - i + 1)\|v\|_\infty |[v]_j|$, hence by sum $B_j \leqslant \lambda_i$. All in all, $\max_{j \geqslant i} B_j \leqslant \lambda_i \times (1 + (d - i + 1)\|v\|_\infty |[v]_{j_i}|)$, proving Eq. (3.6). □

3 Estimating the eigenvectors of (P_m)

We use the bounds from Lemma 3.1 and Eq. (3.5) to estimate the eigenvectors of (P_m) by applying Eq. (3.1), for $m = 1$ in the next section and $m \geq 1$ in Section 3.2.

3.1 Rank-one perturbation

In Proposition 3.1, we obtain bounds on the coordinates of the eigenvectors of (P_m) ($\mathbf{B}$ is assumed to be diagonal) when $m = 1$, which comes as a consequence of Eq. (3.1).

Proposition 3.1. Let $\mathbf{D} = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ be a diagonal matrix with $\lambda_1 \geqslant \cdots \geqslant \lambda_d > 0$. Let $v \in \mathbb{R}^d$ and $V := \max\{d^{-1/2}, \|v\|_\infty\}$. Consider the matrix $\mathbf{A} = \mathbf{D} + \sqrt{D} v v^\top \sqrt{\mathbf{D}}$ and $\nu_1 \geqslant \cdots \geqslant \nu_d$ its eigenvalues and $(e_1^{(1)}, \dots, e_d^{(1)})$ a corresponding orthonormal basis of eigenvectors. Then,

$$\left|[e_i^{(1)}]_j\right| \leqslant 5 d^2 V^4 \sqrt{\frac{\min\{\lambda_i, \lambda_j\}}{\max\{\lambda_i, \lambda_j\}}} \quad \text{for all } i, j \in \{1, \dots, d\} \tag{3.7}$$

Proof. We prove first that, if $\max\{\lambda_i,\lambda_j\} > (1+dV^2)\times\min\{\lambda_i,\lambda_j\}$, then

$$\left|[e_i^{(1)}]_j\right| \leqslant (d-i+1)V^2 \times \frac{\inf_{\rho\in(0,1)}\psi(\rho,(d-i+1)V^2)}{1-(1+(d-i+1)V^2)\frac{\min\{\lambda_i,\lambda_j\}}{\max\{\lambda_i,\lambda_j\}}}\sqrt{\frac{\min\{\lambda_i,\lambda_j\}}{\max\{\lambda_i,\lambda_j\}}} \tag{3.8}$$

with $\psi(\rho,W)=\max\{2(1-\rho)^{-1/2},2\rho^{-1}W\}$, from which we deduce Eq. (3.7).
First suppose that the eigenvalues λ_i of $\mathbf{D}$ are distinct, and that all entries of v are nonzero. Then, by Eq. (3.1), we have for $i,j\in\{1,\dots,d\}$ that $[e_i^{(1)}]_j = C_i\frac{[\sqrt{\mathbf{D}}v]_j}{\lambda_j-\nu_i} = C_i\frac{\sqrt{\lambda_i}[v]_j}{\lambda_j-\nu_i}$, where $C_i\in\mathbb{R}$ is chosen such that $\|e_i^{(1)}\|=1$, hence

$$|C_i| = \left\|\left(\frac{\sqrt{\lambda_j}[v]_j}{\lambda_j-\nu_i}\right)_{j=1,\dots,d}\right\|^{-1} = \left(\sum_{j=1}^d\left|\frac{\sqrt{\lambda_j}[v]_j}{\lambda_j-\nu_i}\right|^2\right)^{-1/2} \leqslant \min_{1\leqslant j\leqslant d}\frac{|\lambda_j-\nu_i|}{\sqrt{\lambda_j}|[v]_j|}. \tag{3.9}$$

Combining [135, Theorem 2.1, p. 175] with Eq. (3.6), we have $\lambda_i<\nu_i\leqslant\lambda_i\times(1+(d-i+1)V|[v]_{j_i}|)$, where j_i is defined in Lemma 3.1. By definition of j_i we have $0\leqslant\lambda_i-\lambda_{j_i}\leqslant\lambda_i(d-i+1)V|[v]_{j_i}|$. By sum, we obtain $0<\nu_i-\lambda_{j_i}\leqslant 2\lambda_i(d-i+1)V|[v]_{j_i}|$. We apply this to Eq. (3.9) to get

$$|C_i| \leqslant \frac{\nu_i-\lambda_{j_i}}{\sqrt{\lambda_{j_i}}|[v]_{j_i}|} \leqslant \frac{2\lambda_i(d-i+1)V|[v]_{j_i}|}{\sqrt{\lambda_{j_i}}|[v]_{j_i}|} = \frac{\lambda_i}{\sqrt{\lambda_{j_i}}}2(d-i+1)V\ . \tag{3.10}$$

Let $\rho\in(0,1)$. If $(d-i+1)V^2\sqrt{\lambda_{j_i}}\leqslant\rho\sqrt{\lambda_i}$, then by definition of j_i, $\lambda_{j_i}\geqslant\lambda_i\times(1-\rho)$, and by Eq. (3.10), $|C_i|\leqslant(1-\rho)^{-1/2}\times 2(d-i+1)V\sqrt{\lambda_i}$. Otherwise, $|C_i|\leqslant\rho^{-1}\times 2(d-i+1)^2V^3\sqrt{\lambda_i}$. All in all, for $\rho\in(0,1)$,

$$|C_i| \leqslant (d-i+1)V\sqrt{\lambda_i}\times\max\left\{2(1-\rho)^{-1/2},2\rho^{-1}(d-i+1)V^2\right\} =: C_\rho\sqrt{\lambda_i}\ . \tag{3.11}$$

Then, $|[e_i^{(1)}]_j| = |C_i|\sqrt{\lambda_j}|[v]_j|/|\lambda_j-\nu_i|\leqslant\sqrt{\lambda_i\lambda_j}/|\lambda_j-\nu_i|\times\inf_{\rho\in(0,1)}C_\rho$. By Eq. (3.6), when $\lambda_j<\lambda_i$, $|[e_i^{(1)}]_j|\leqslant\min_{\rho\in[0,1]}C_\rho\times(1-\lambda_j/\lambda_i)^{-1}\sqrt{\lambda_j/\lambda_i}$. By Eq. (3.5), when $\lambda_j>(1+dV^2)\lambda_i$, $|[e_i^{(1)}]_j|\leqslant\min_{\rho\in[0,1]}C_\rho\times(1-(1+dV^2)\lambda_i/\lambda_j)^{-1}\sqrt{\lambda_i/\lambda_j}$.
If the eigenvalues of $\mathbf{D}$ are not distinct or not all entries of v are nonzero, we consider a sequence of diagonal matrices $\{\mathbf{D}_k=\mathrm{diag}(\lambda_1^k,\dots,\lambda_d^k)\}_{k\in\mathbb{N}}$ such that the diagonal elements $\lambda_1^k>\lambda_2^k>\cdots>\lambda_d^k>0$ are distinct and $\mathbf{D}_k\to\mathbf{D}$ when $k\to\infty$, and a sequence of vectors $\{v_k\in\mathbb{R}^d_{\neq 0}\}_{k\in\mathbb{N}}$ with only nonzero entries where $\|v_k\|_\infty\leqslant N$ and $v_k\to v$ when $k\to\infty$. Denote then $\mathbf{A}_k=\mathbf{D}_k+\sqrt{\mathbf{D}_k}v_kv_k^\top\sqrt{\mathbf{D}_k}$ and $\nu_1^k\geqslant\cdots\geqslant\nu_d^k$ its eigenvalues. Note that $\mathbf{A}_k\to\mathbf{A}$ when $k\to\infty$, so by continuity of the eigenvalues, $\nu_i^k\to\nu_i$ when $k\to\infty$. Furthermore, we just proved that if $e_1^k,\dots,e_d^k$ are unit eigenvectors of A_k corresponding respectively to the eigenvalues $\nu_1^k,\dots,\nu_d^k$, then $e_1^k,\dots,e_d^k$ and $\lambda_1^k,\dots,\lambda_d^k$ satisfy Eq. (3.8). Moreover, the vectors e_i^k all belong to the unit sphere of $\mathbb{R}^d$, so up to considering a subsequence of $\{\mathbf{A}_k\}_{k\in\mathbb{N}}$, we can assume w.l.o.g. that each e_i^k tends to a vector $e_i^{(1)}\in\mathbb{R}^d$ when $k\to\infty$. As $(e_1^k,\dots,e_d^k)$ is an orthonormal system of $\mathbb{R}^d$, so is its limit $(e_1^{(1)},\dots,e_d^{(1)})$ and $e_i^{(1)}$ is an eigenvector of A corresponding to the eigenvalue ν_i. Therefore, Eq. (3.8) holds by taking the limit $k\to\infty$ in the equation satisfied by $e_1^k,\dots,e_d^k$ and $\lambda_1^k,\dots,\lambda_d^k$.

To obtain Eq. (3.7), note that when $\rho = 1/2$, we have $2(1-\rho)^{-1/2} \leqslant 4 \leqslant 4dV^2$, and $2\rho^{-1}(d-i+1)V^2 \leqslant 4dV^2$, and when $\max\{\lambda_i,\lambda_j\}/\min\{\lambda_i,\lambda_j\} > 1+4dV^2$, by Eq. (3.8), then,

$$(1-(1+dV^2)\min\{\lambda_i,\lambda_j\}/\max\{\lambda_i,\lambda_j\})^{-1} \leqslant (1+4dV^2)/(4dV^2) \leqslant 5/4,$$

as $V \geqslant d^{-1/2}$, and thus Eq. (3.7) holds.
If otherwise $\max\{\lambda_i,\lambda_j\} \leqslant (1+4dV^2)\min\{\lambda_i,\lambda_j\}$, as $\max\{\lambda_i,\lambda_j\}/\min\{\lambda_i,\lambda_j\} \geqslant 1$, we find $|[e_i]_j| \leqslant 1 \leqslant (1+4dV^2)\sqrt{\max\{\lambda_i,\lambda_j\}/\min\{\lambda_i,\lambda_j\}}$. Since $1 \leqslant dV^2$, then $(1+4dV^2) \leqslant 5dV^2$ and Eq. (3.7) holds. □

3.2 *Sum of m rank-one matrices pertubation*

Our final Theorem 3.2 generalizes Proposition 3.1 to any value $m \geqslant 1$ and is obtained by induction using Eq. (3.8). Theorem 3.2 implies in particular Theorem 3.1 via the spectral theorem.

Theorem 3.2. Let $\mathbf{D} = \operatorname{diag}(\lambda_1,\dots,\lambda_d)$ be a diagonal matrix with $\lambda_1 \geqslant \cdots \geqslant \lambda_d > 0$. Let $V \geqslant 1/\sqrt{d}$ and consider a sequence of vectors $v^{(i)} \in \mathbb{R}^d$ such that $\|v^{(i)}\|_\infty \leqslant V$ for all $i \in \mathbb{N}$. For $m \in \mathbb{N}$, let $\mathbf{A}^{(m)} = \mathbf{D} + \sqrt{\mathbf{D}}\sum_{i=1}^m [v^{(i)}][v^{(i)}]^\top\sqrt{\mathbf{D}}$ and $\nu_1^{(m)} \geqslant \cdots \geqslant \nu_d^{(m)}$ the eigenvalues of $\mathbf{A}^{(m)}$ and $(e_1^{(m)},\dots,e_d^{(m)})$ a corresponding orthonormal system of eigenvectors. Then

$$|[e_i^{(m)}]_j| \leqslant C_m\sqrt{\frac{\min\{\lambda_i,\lambda_j\}}{\max\{\lambda_i,\lambda_j\}}} \quad \text{for all } i,j \in \{1,\dots,d\} \text{ and } m \in \mathbb{N} \tag{3.12}$$

with $C_0 = 1$ and $C_{m+1} = 5d^7V^4C_m^5\sqrt{1+dmV^2}$.

Proof. For $a,b > 0$, denote $\alpha(a,b) = \sqrt{\min\{a,b\}/\max\{a,b\}}$. Let $m \in \mathbb{N}$ and assume that Eq. (3.12) holds which is true if $m = 0$ since $C_0 = 1$. Observe now that $\mathbf{A}^{(m+1)} = \mathbf{A}^{(m)} + \sqrt{\mathbf{D}}[v^{(m+1)}][v^{(m+1)}]^\top\sqrt{\mathbf{D}}$. In the system of coordinates $\mathbf{B}^{(m)} := (e_1^{(m)},\dots,e_d^{(m)})$, $\mathbf{A}^{(m)}$ writes as $\mathbf{D}^{(m)} := \operatorname{diag}(\nu_1^{(m)},\dots,\nu_d^{(m)})$. Since $\lambda_1,\dots,\lambda_d > 0$, and as $\mathbf{A}^{(m)} \succeq \mathbf{D}$, then $\nu_i^{(m)} \geqslant \lambda_i > 0$ for $i \in \{1,\dots,d\}$, and

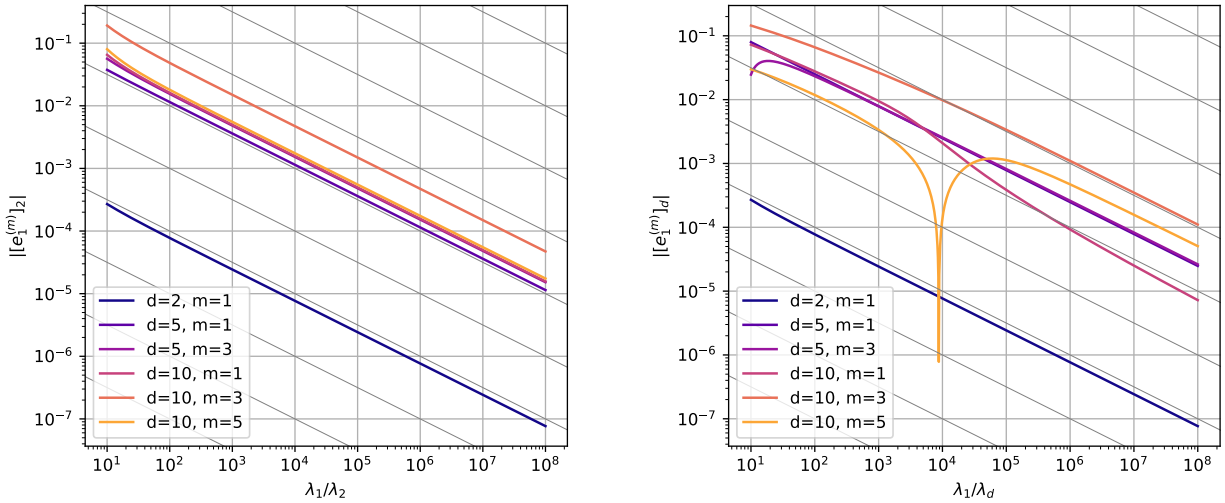
$$\begin{aligned}\left\langle\sqrt{\mathbf{D}}v^{(m+1)}, e_i^{(m)}\right\rangle &= \sum_{j=1}^d [e_i^{(m)}]_j\sqrt{\lambda_j}[v^{(m+1)}]_j = \sqrt{\mathbf{D}_{ii}^{(m)}} \times \sum_{j=1}^d \sqrt{\lambda_j/\nu_i^{(m)}}[e_i^{(m)}]_j[v^{(m+1)}]_j \\ &=: \left[\sqrt{\mathbf{D}^{(m)}}w^{(m+1)}\right]_i.\end{aligned}$$

Hence $\left[\mathbf{A}^{(m+1)}\right]_{\mathcal{B}^{(m)}} = \mathbf{D}^{(m)} + \sqrt{\mathbf{D}^{(m)}}[w^{(m+1)}]\,[w^{(m+1)}]^\top\sqrt{\mathbf{D}^{(\mathbf{m})}}$ with

$$\|w^{(m+1)}\|_\infty \leqslant \sum_{j=1}^d \sqrt{\lambda_j/\nu_i^{(m)}}|[e_i^{(m)}]_j| \times V \leqslant \sum_{j=1}^d \sqrt{\lambda_j/\lambda_i}|[e_i^{(m)}]_j| \times V \leqslant dV \times C_m.$$

We apply Proposition 3.1 to $\left[\mathbf{A}^{(m+1)}\right]_{\mathcal{B}^{(m)}}$ so that, for $i,k \in \{1,\dots,d\}$,

$$|\langle e_i^{(m+1)}, e_k^{(m)}\rangle| = |[[e_i^{(m+1)}]_{\mathcal{B}^{(m)}}]_k| \leqslant 5d^6V^4C_m^4\alpha(\nu_i^{(m)},\nu_k^{(m)}).$$



(a) $j = 2$ (b) $j = d$

Figure 3.1: Value of $|[e_1^{(m)}]_j|$ as a function of λ_1/λ_j where $e_1^{(m)}$ is an eigenvector associated to the largest eigenvalue of $\mathbf{A}^{(m)}$ from (P_m) for different dimensions and values of m as given in the legend. The $v^{(i)}$ are independent standard Gaussian vectors (with the same realization for all values of λ_1) and the eigenvalues of the diagonal matrix $\mathbf{B}$ are chosen uniformly on a log scale between $\lambda_d = 1$ and λ_1. The value $|[e_1^{(m)}]_j|$ behaves consistent with $\Theta(\sqrt{\lambda_j/\lambda_1})$.

By Eq. (3.5), $|\langle e_i^{(m+1)}, e_k^{(m)}\rangle| \leqslant (5d^6V^4C_m^4)(1 + dmV^2)^{1/2}\ \alpha(\lambda_i, \lambda_k)$. Since $|[e_k^{(m)}]_j| \leqslant C_m\alpha(\lambda_i, \lambda_k)$, then,

$$|[e_i^{(m+1)}]_j| \leqslant \sum_{k=1}^{d} |[e_k^{(m)}]_j| \times |\langle e_i^{(m+1)}, e_k^{(m)}\rangle| \leqslant dC_m \times 5d^6V^4C_m^4(1 + dmV^2)^{1/2}\alpha(\lambda_i, \lambda_j).$$

This proves by induction that Eq. (3.12) holds for all $m \in \mathbb{N}$. □

3.3 Thightness

Figure 3.1 shows numerical computations of coordinates of the first eigenvector of $\mathbf{A}^{(m)}$ in dimension 2, 5, 10. The coordinates seem to obey $\Theta(\min\{\lambda_i, \lambda_j\}/\max\{\lambda_i, \lambda_j\})$ in all cases which suggests that this rate in our upper bounds is tight. However we do not expect the constant C_m given in Theorem 3.2 to be tight.

Acknowledgments

The authors would like to thank Stéphane Gaubert for his constructive remarks and feedback on a prior version of the manuscript as well as the anonymous referee for their valuable review and comments.

Chapter 4

Geometric ergodicity of Markov chains underlying CMA-ES

Comments on Chapter 4: This chapter contains the working paper "Geometric ergodicity of Markov chains underlying CMA-ES" (Armand Gissler, Anne Auger, Nikolaus Hansen) that we aim to submit soon for publication. This constitutes the main difficulties of this thesis : establishing a drift condition for Markov chains associated to CMA-ES is an active open question for several years. Our methodology is to normalize the states of CMA-ES, and proving that the resulting Markov chain is geometrically ergodic and converges to a stationary probability distribution. This is the direct sequel of Chapter 2, where we find normalized chains that are irreducible and aperiodic T-chains. The irreducibility and aperiodicity are conditions required to use drift criteria and deduce the ergodicity, and the T-chain property implies that compact sets are small, which facilitates our work since we only have to find a drift for initial conditions outside a small set. However, we need supplementary assumptions, in particular we only include ellipsoidal objective functions. The consequences of the geometric ergodicity are the linear convergence of CMA-ES and the learning of the inverse Hessian, which are stated and proven in Chapter 5.

Abstract

We prove the geometric ergodicity of Markov chains defined via the normalization of the states of the stochastic derivative-free optimization algorithm CMA-ES on ellipsoidal functions—defined as composites of strictly increasing functions with convex-quadratic functions—under appropriate choice of the hyperparameters of the algorithm. Those results constitute a major milestone to establish the global convergence together with the learning of second order information of the algorithm. It relies on the extension of a state-dependent drift criterion for ergodicity which enables the analysis of projected Markov chains when we know the updates of a redundant chain.

Keywords: Markov chains, geometric ergodicity, drift, CMA-ES, linear convergence.

1 Introduction . . . 116
1.1 Notations . . . *118*
2 Geometric ergodicity of a normalized chain underlying CMA-ES . . . 119
2.1 The CMA-ES algorithm and first assumptions . . . *119*
2.2 Objective function assumptions . . . *122*
2.3 Definition of a normalized chain underlying the CMA-ES algorithm . . . *123*
2.4 Main results . . . *125*
3 Proof of the main results . . . 129
3.1 Methodology to prove the geometric ergodicity of the normalized Markov chain . . . *129*
3.2 Proof of Theorem 4.3 . . . *132*
3.3 Preliminary definitions and results for the proof of Proposition 4.4 . . . *133*
3.4 Bounding the expected largest eigenvalue of the (normalized) covariance matrix . . . *147*
3.5 Bounding the expected (normalized) mean . . . *156*
3.6 Proof of Proposition 4.4 . . . *166*
4 Discussion . . . 178
A Technical results . . . 179
B Proof of Proposition 4.1 . . . 179
C Proofs in Section 3.3 . . . 181
C.1 Proof of Theorem 4.6 . . . *181*
C.2 Proof of Corollary 4.2 . . . *181*
C.3 Proof of Lemma 4.2 . . . *182*
C.4 Proof of Proposition 4.6 . . . *183*
C.5 Proof of Proposition 4.7 . . . *184*
C.6 Proof of Proposition 4.8 . . . *185*
C.7 Proof of Proposition 4.9 . . . *186*
C.8 Proof of Lemma 4.3 . . . *188*

1 Introduction

The analysis of the convergence of stochastic processes is central in many (applied) mathematical problems ranging from population dynamics [45] to biology [33]. In optimization, the convergence reveals whether an algorithm will eventually solve the optimization problem. Proving additionally convergence rates hints towards how fast the problem will be solved. In this context, we focus on discrete-time stochastic processes with continuous state space and investigate the geometric ergodicity of a Markov chain—convergence at a geometric rate of its t-step transition kernel towards a stationary distribution—underlying the covariance matrix adaptation evolution strategy (CMA-ES) [76, 73, 63]. The CMA-ES algorithm is a well-known and widely used[1] derivative-free stochastic optimization algorithm, loosely inspired from evolutionary biology [68] and often employed to optimize difficult black-box functions that are typically nonconvex, nondifferentiable, ill-conditioned, multimodal. Finding a convergence proof of CMA-ES has been challenging researchers for over twenty years. When using a Markov chain methodology to prove linear convergence [17, 23], a method which has been applied to analyze simpler algorithms [141, 14, 16], a key step is the study of geometric ergodicity of a normalized Markov chain which underlies the algorithm. As the icing on the cake, this methodology would also allow to prove that CMA-ES learns legitimate second order information.

[1] Two Python implementations of CMA-ES [67, 117] receive 1.3 million monthly downloads from PyPI and overall more than 70 million downloads (as of November 2024).

In a nutshell, CMA-ES is an adaptive stochastic algorithm to optimize a numerical function $f : \mathbb{R}^d \to \mathbb{R}$ referred to as objective function. It samples candidate solutions from a multivariate normal distribution parametrized by a mean and a covariance matrix. The adaptation of those parameters—realized after the sampling and using the feedback given by the objective function f on how good the sampled candidate solutions are—is crucial to a fast convergence. On wide classes of functions that include nonconvex functions, the mean is empirically observed to converge geometrically to the optimum. The covariance matrix that encodes a metric around the mean converges to zero while learning the topography of the function to be optimized. On composite of a strictly increasing function with a convex-quadratic function, the covariance matrix is observed to be proportional in average to the inverse Hessian of the convex-quadratic function. Since both the mean and the covariance matrix converge to the optimum and zero respectively, one needs to normalize them to observe a stationary process. The Markov chain methodology then consists in analyzing the stability of a normalized Markov chain which exists when the objective function is scaling-invariant [17]. We retrieve the linear convergence of the algorithm by applying limit theorems to the normalized chain and the mathematical study therefore consists in investigating stability properties for a Law of Large Numbers to be used. It can typically be implied from two main properties: the irreducibility of the Markov chain (that has been investigated in a prior work [53]) and its geometric ergodicity (that implies the existence of an invariant probability measure and Harris-recurrence) which is the focus of this paper.

We prove the geometric ergodicity using a multistep or state-dependent Foster-Lyapunov drift condition. The multistep aspect is needed to include the so-called cumulation mechanisms of the algorithms. Such multistep conditions have already been successfully applied in the analysis of Markov chains [147, 32]. Alternative techniques to drift conditions exist to prove geometric ergodicity. For instance, the time of first return to a small set [41, Theorems 11.4.1 and 11.4.2] can be investigated as was recently done to analyze a sampling algorithm [11]. They might, however, be even more complex to use when applied to CMA-ES.

In the case of stochastic optimization algorithms belonging to the same class than CMA-ES but where only the stepsize and the mean are adapted (instead of a full covariance matrix), proving the geometric ergodicity by means of a Foster-Lyapunov drift condition is relatively intuitive and simple [141]. It comes from the property of the algorithm that the stepsize increases when optimizing a linear function which corresponds to the situation where the normalized Markov chain (on scaling-invariant functions with a unique optimum) is outside a compact set of the state space (i.e., far away from the optimum). The stepsize increase corresponds to an expected inverse stepsize strictly smaller than 1 which gives a geometric drift condition. The situation is much more complex when we want to analyze an algorithm with a full covariance matrix. In this case, the Markov chain to be studied lives indeed on a manifold and many more subcases need to be considered to prove a drift condition outside a compact set of the state space. Each of them involves a different mechanism of the algorithm to be controlled to succeed in proving a drift condition outside a compact set.

We perform the study of geometric ergodicity on the class of ellipsoidal objective functions, that can be defined equivalently as quasiconvex functions with ellipsoidal level sets or as increasing transformations of convex-quadratic functions. This class of functions is the primary one to consider to demonstrate linear convergence and learning of second order information by the covariance matrix. It is indeed the one used to demonstrate empirically the learning of inverse Hessian property. It is included in the class of scaling-invariant objective functions needed to define a normalized Markov chain [53, Proposition 2.2] as well as on the class where we can prove the irreducibility of the Markov chain [53, Theorem 3.1]. Since establishing its geometric ergodicity is a much more complex task we

restrict ourselves to ellipsoidal functions. This simplifies the analysis as we can use explicit forms of what we define as selection functions and easily measure the impact of each coordinate on the ranking of the candidate solutions, as performed by CMA-ES at each iteration.

The paper is organized as follows. In Section 2, we provide a mathematical framework for the CMA-ES algorithm as well as the different assumptions used for the mathematical study. We additionally recall the definition of normalized Markov chains underlying CMA-ES when the objective function is scaling-invariant and remind that they are irreducible and aperiodic under additional assumptions on the objective function, the stepsize change and the hyperparameters used to update the algorithm states. We finally state our two main results: (i) that a multistep geometric drift condition holds (Proposition 4.4) and thus that (ii) the Markov chains are geometrically ergodic (Theorem 4.3). Section 3 is devoted to the proofs of our main results. For the sake of conciseness, several proofs are presented in the appendices.

1.1 Notations

Given $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space the set Ω called outcome space is equipped with a σ-field denoted $\mathcal{A}$ and called event set. Given a random variable valued in a measurable set $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ where $\mathcal{B}(\mathsf{X})$ is a σ-field of X (which is a measurable function X from Ω to X), a realization of the random variable X is denoted $X(\omega)$ for an outcome $\omega \in \Omega$. Given $\mathsf{A} \in \mathcal{B}(\mathsf{X})$, we define the event $X \in \mathsf{A}$ as $X^{-1}(\mathsf{A}) \in \mathcal{A}$. The distribution or law of a random variable X is the probability measure ν_X on $\mathcal{B}(\mathsf{X})$ satisfying for $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ that $\nu_X(\mathsf{A}) = \mathbb{P}[X \in \mathsf{A}]$.

We denote the set of nonnegative integers by $\mathbb{N}$, the set of positive integers by $\mathbb{N}^*$, the set of real numbers by $\mathbb{R}$, the set of nonnegative real numbers by $\mathbb{R}_+$, the set of positive real numbers by $\mathbb{R}_{++}$. For any $n \in \mathbb{N}^*$, the set of permutations of $\{1, \dots, n\}$, i.e., the set of bijective functions between $\{1, \dots, n\}$ and itself, is denoted $\mathfrak{S}_n$. Unless precised otherwise, $\|\cdot\|$ will denote the Euclidean norm.

For a vector $u \in \mathbb{R}^d$ with $d \in \mathbb{N}^*$ and $k = 1, \dots, d$, we denote $[u]_k$ the k-th coordinate of u in the canonical basis of $\mathbb{R}^d$. We denote by $\mathcal{S}^d$, $\mathcal{S}^d_+$ and $\mathcal{S}^d_{++}$, the sets of symmetric, positive semidefinite and symmetric positive definite matrices of size $d \times d$, respectively. Given $\mathbf{A} \in \mathcal{S}^d_{++}$, there exist $a_1, \dots, a_d > 0$ and an orthogonal matrix $\mathbf{P} \in \mathbb{R}^{d \times d}$, i.e., $\mathbf{P}^\top = \mathbf{P}^{-1}$, such that $\mathbf{A} = \mathbf{P} \times \mathrm{diag}(a_1, \dots, a_d) \times \mathbf{P}^\top$ where $\mathrm{diag}(a_1, \dots, a_d)$ is the diagonal matrix of size $d \times d$, with diagonal elements $a_1, \dots, a_d$. We then denote the unique (positive definite) *square root* of $\mathbf{A}$ as $\sqrt{\mathbf{A}} \coloneqq \mathbf{P} \times \mathrm{diag}(\sqrt{a_1}, \dots, \sqrt{a_d}) \times \mathbf{P}^\top$ and hence $\sqrt{\mathbf{A}} \times \sqrt{\mathbf{A}} = \mathbf{A}$.

Given $\mathbf{A} \in \mathcal{S}^d_{++}$, we denote $\lambda_k(\mathbf{A})$ the k-th largest eigenvalue (counted with multiplicity) of $\mathbf{A}$, and $e_k(\mathbf{A})$ an eigenvector associated to the eigenvalue $\lambda_k(\mathbf{A})$, such that $\|e_k(\mathbf{A})\| = 1$, and $e_k(\mathbf{A})^\top e_j(\mathbf{A}) = 0$ dor $j \neq k$ in $\{1, \dots, d\}$. We use the notation $\lambda_{\max} = \lambda_1$ and $\lambda_{\min} = \lambda_d$. We denote by $\det(\mathbf{A})$ the determinant of the matrix $\mathbf{A}$.

Given a distribution ν on the measured space X and a distribution μ on the measured space Y, the joint distribution of ν and μ on $\mathsf{X} \times \mathsf{Y}$ is denoted $\nu \otimes \mu$. Likewise, the joint distribution of ν with itself k times on X^k is denoted $\nu^{\otimes k}$. When a distribution on $\mathbb{R}^d$ has a density with respect to the Lebesgue measure, we use the same notation for the distribution than for the density function, unless it is precised otherwise.

Once the normalized process (4.17) is introduced, we use throughout the paper the notation $\mathbb{E}_t[\cdot]$ for $t \in \mathbb{N}$ as the conditional expectation $\mathbb{E}[\,\cdot \mid \mathcal{F}_t]$ where $\mathcal{F}_t$ is the filtration induced by the states of (4.17) until iteration t included.

For any nonnegative number a, the condition 'a *is sufficiently small*' means that there exists an $\varepsilon > 0$ such that all values $a \leqslant \varepsilon$ satisfy the condition. Moreover, for a positive number b, we say that 'a *is sufficiently smaller than* b' if the quotient a/b is sufficiently small.

2 Geometric ergodicity of a normalized chain underlying CMA-ES

We present in this section the mathematical description of CMA-ES and outline the main assumptions related to the algorithm, including those concerning the hyperparameters, the sampling distribution, and the stepsize updates. We also specify the assumptions on the class of functions considered in this article. Next, we introduce the associated normalized Markov chains, detailing the assumptions on the normalization function and previous results on the irreducibility of these chains. Last, we state our main results, which provide sufficient conditions for the geometric ergodicity of this Markov chain.

2.1 The CMA-ES algorithm and first assumptions

The CMA-ES algorithm aims to minimize an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ by iteratively sampling new candidate solutions. It can be described by the iteration sequence of its states, namely $\{(m_t, \sigma_t, \mathbf{C}_t, p_t^\sigma, p_t^c)\}_{t\in\mathbb{N}}$ where the mean $m_t \in \mathbb{R}^d$ gives an estimate of the optimum x^* of f, $\mathbf{C}_t \in \mathcal{S}_{++}^d$ is a positive definite matrix called covariance matrix of the algorithm, the stepsize $\sigma_t > 0$ allows to quickly change the scale of the distribution of new candidate solutions allowing for linear convergence at a close-to-optimal rate, and the evolution paths $p_t^\sigma \in \mathbb{R}^d$ and $p_t^c \in \mathbb{R}^d$ track the direction taken by the mean during the previous iterations in order to update the stepsize and the covariance matrix, respectively. The stepsize increases when $\|p_t^\sigma\|$ is larger than its expected value under neutral selection (or no selection step). Likewise, p_t^c indicates a direction in which the covariance matrix should increase its variance.

We consider $\lambda \geqslant 2$ independent and identically distributed (i.i.d.) random vectors $U_{t+1}^i \in \mathbb{R}^d$ following a probability distribution ν_U^d on $\mathcal{B}(\mathbb{R}^d)$ (usually in practice ν_U^d is the standard multivariate normal distribution) where $i = 1, \dots \lambda$. The parameter λ is referred to as the population size. It is an integer greater than or equal to 2, and $t \in \mathbb{N}$ denotes the current iteration. Besides, we assume that the process $\mathbf{U} = \{(U_{t+1}^1, \dots, U_{t+1}^\lambda)\}_{t\in\mathbb{N}}$ forms an i.i.d. sequence of random vectors independent of the initial algorithm state.

We consider the following assumptions for the sampling distribution:

N1. *ν_U^d has a positive continuous density $p_U^d(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}^d$,*

N2. *$\nu_U^d(\mathbf{R}\mathsf{A}) = \nu_U^d(\mathsf{A})$ for every $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ and orthogonal matrix $\mathbf{R}$,*

N3. *given $u \sim \nu_U^d$, the coordinates $[u]_k$ of u are i.i.d. with marginal distribution ν_U^1,*

N4. *given $u \sim \nu_U^d$, the coordinates $[u]_k$ admit n-th order moments for all $n \in \mathbb{N}$ and $\mathbb{E}[u]_k^2 = 1$ for $k = 1, \dots, d$.*

Assumption **N3** allows us to define for every $k \geqslant 1$ the distribution $\nu_U^k = (\nu_U^1)^{\otimes k}$ on $\mathcal{B}(\mathbb{R}^k)$. A random variable $u \sim \nu_U^d$ satisfies **N2** and **N3** and $\mathbb{E}[u]_k^2 = 1$ for $k = 1, \dots, d$ if and only if ν_U^d is the standard multivariate normal distribution, a property often referred to as the Herschel-Maxwell theorem [114, Theorem 3.2]. As last assumption, we consider that ν_U^d is a standard multivariate normal distribution:

N5. *ν_U^d is the standard multivariate normal distribution $\mathcal{N}(0, \mathbf{I}_d)$.*

We now describe the update from one iteration to the next. Let $t \in \mathbb{N}$. Given a state $\theta_t = (m_t, \sigma_t, \mathbf{C}_t, p_t^\sigma, p_t^c) \in \mathbb{R}^d \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \times \mathbb{R}^d \times \mathbb{R}^d$, we define $\lambda \geqslant 2$ candidate solutions

$$x_{t+1}^i = m_t + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^i \quad \text{for } i = 1, \dots, \lambda \ . \tag{4.1}$$

If U_{t+1}^i satisfies **N**5, as in CMA-ES, then, conditionally to θ_t, x_{t+1}^i is a random variable that follows a multivariate normal distribution with mean vector m_t and covariance matrix $\sigma_t^2 \mathbf{C}_t$. If U_{t+1}^i satisfies **N**2 and **N**4, then m_t is the mean vector of the candidate solutions.

We define the permutation $s_{t+1} \in \mathfrak{S}_\lambda$ that extracts the ranking of candidate solutions on f such that

$$f\left(x_{t+1}^{s_{t+1}(1)}\right) \leqslant f\left(x_{t+1}^{s_{t+1}(2)}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{s_{t+1}(\lambda)}\right) \quad . \tag{4.2}$$

Any tie-break that does not break the desired invariance properties is eligible and under mild assumptions (such as f has ν_U^d-negligible level sets), s_{t+1} is almost surely unique. Yet, for the sake of completeness, we impose $s_{t+1}^{-1}(i) < s_{t+1}^{-1}(j)$ when $i < j$ and $f(x_{t+1}^i) = f(x_{t+1}^j)$. The permutation determined with (4.2) is the only component in the algorithm that directly depends on the objective function f and is used in most of the update equations below. Consequently, the algorithm is invariant to composing f to the left by a strictly increasing transformation. Given the permutation s_{t+1} giving the ranking of candidate solutions on f, we update the mean vector by

$$m_{t+1} = m_t + c_m \sigma_t \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \quad , \tag{4.3}$$

where $1 \leqslant \mu \leqslant \lambda$, $c_m > 0$ is the learning rate for the mean, and the recombination weights, $\mathbf{w}_m \in \mathbb{R}^\mu$, obey $w_1^m \geqslant \ldots \geqslant w_\mu^m > 0$ to favor better candidate solutions, and $\sum_{i=1}^\mu w_i^m = 1$. Preferably, the weights also obey $1/5 \leqslant \mu_{\text{eff}}/\lambda < 1/2$, where

$$\mu_{\text{eff}} := \frac{1}{\sum_{i=1}^\mu (w_i^m)^2} = \frac{1}{\|\mathbf{w}_m\|_2^2} \tag{4.4}$$

is the effective selection mass. The learning rate is usually set to $c_m = 1$ which simplifies (4.3) to a mean update which is equal to a weighted recombination of the μ best candidate solutions:

$$m_{t+1} = \sum_{i=1}^{\mu} w_i^m x_{t+1}^{s_{t+1}(i)} \quad . \tag{4.5}$$

We update the evolution paths as

$$p_{t+1}^c = (1 - c_c) p_t^c + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}} \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \quad , \tag{4.6}$$

and

$$p_{t+1}^\sigma = (1 - c_\sigma) p_t^\sigma + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}}} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \tag{4.7}$$

where $c_\sigma \in (0, 1]$ and $c_c \in (0, 1]$ can be interpreted as decay parameters. The update for the covariance matrix then reads

$$\mathbf{C}_{t+1} = (1 - c_1 - c_\mu \sum_{i=1}^{\lambda} w_i^c)\mathbf{C}_t + \underbrace{c_1 p_{t+1}^c {p_{t+1}^c}^\top}_{\text{rank-one update}} + \underbrace{c_\mu \sum_{i=1}^{\lambda} w_i^c \left(\sqrt{\mathbf{C}_t} U_{t+1}^{s_{t+1}(i)}\right)\left(\sqrt{\mathbf{C}_t} U_{t+1}^{s_{t+1}(i)}\right)^\top}_{\text{rank-mu update}} \quad , \tag{4.8}$$

where $c_1 \in [0, 1]$ and $c_\mu \in [0, 1]$ are the *learning rates* for the so-called *rank-one* and *rank-mu updates*, respectively. The weights obey $w_1^c \geqslant \ldots \geqslant w_\mu^c > 0 \geqslant w_{\mu+1}^c \geqslant \ldots \geqslant w_\lambda^c$, $\sum_{i=1}^\mu w_i^c = 1$ and $\sum_{i=1}^\lambda w_i^c \geqslant -c_1/c_\mu$ [65, p. 30]. The first coefficient of the RHS is chosen such that $\mathbb{E}[\mathbf{C}_{t+1} \mid \mathbf{C}_t] = \mathbf{C}_t$ when $s_{t+1} = \text{Id}_{\{1,\ldots,\lambda\}}$, and $p_{t+1}^c \sim \sqrt{\mathbf{C}_t} U_1^1$.

For our analysis, we require the strict inequalities $c_\mu > 0$ and $c_1 + c_\mu < 1$ in (4.8) and the recombination weights to obey the additional standard assumption

W1. $w_1^\star \geqslant \cdots \geqslant w_\mu^\star > w_{\mu+1}^\star = \ldots w_\lambda^\star = 0$ *and* $\sum_{i=1}^\mu w_i^\star = 1$, *for* $\star = m, c$.

The latter implies that all recombination weights are positive and the weights $w_{\mu+1}^\star, ..., w_\lambda^\star$ are zero. Without loss of generality, we assume $\sum_i w_i^\star = 1$, as we can absorb $\sum_i w_i^m$ into c_m of Eq. (4.3) and $\sum_i w_i^c$ into c_μ of Eq. (4.8). We finally update the stepsize in a multiplicative manner as

$$\sigma_{t+1} = \sigma_t \times \Gamma_{d_\sigma}\left(p_{t+1}^\sigma\right) \tag{4.9}$$

where p_{t+1}^σ is given by (4.7). We call $\Gamma_{d_\sigma} : \mathbb{R}^d \to \mathbb{R}_{++}$ the stepsize change and consider two variants. The default update, usually referred to as cumulative stepsize adaptation [63], reads

$$\Gamma_{\text{CSA}}^1(p) \coloneqq \exp\left(\frac{c_\sigma}{d_\sigma} \times \left(\frac{\|p\|}{\mathbb{E}\|\nu_U^d\|} - 1\right)\right) \tag{4.10}$$

where $d_\sigma > 0$ is generally chosen greater than or equal to 1 to serve as a *damping parameter* that scales down the stepsize changes. The update compares $\|p\| = \|p_{t+1}^\sigma\|$ to its expected value $\mathbb{E}\|\nu_U^d\|$ under neutral selection which depends on the distribution ν_U^d. We have $\mathbb{E}\|\nu_U^d\| \approx \sqrt{d}$ when $\nu_U^d = \mathcal{N}(0, \mathbf{I}_d)$ as by default. An alternative update [68, Eq. (1)] reads more elegantly

$$\Gamma_{\text{CSA}}^2(p) \coloneqq \exp\left(\frac{c_\sigma}{2d_\sigma} \times \left(\frac{\|p\|^2}{d} - 1\right)\right) \tag{4.11}$$

which is smooth in 0. We set d_σ proportional to $\sqrt{\mu_{\text{eff}}}$ for Γ_{CSA}^1 and proportional to μ_{eff} for Γ_{CSA}^2. This update was used in proofs of convergence for stepsize adaptive ES [141], that are based on similar approaches than ours. However, a recent generalization [52] of the methodology to obtain irreducibility and topological properties of Markov chains underlying ES algorithms extended the assumption of a smooth update to a locally Lipschitz update. Yet, CMA-ES variants with Γ_{CSA}^1 and Γ_{CSA}^2 have shown similar empirical performances [49]. We thus consider in this paper an abstract stepsize change $\Gamma_{d_\sigma} : \mathbb{R}^d \to \mathbb{R}_{++}$ (which depends on the damping parameter d_σ) that encompasses both of them and on which we use the following assumptions.

Γ1. *The stepsize change* $\Gamma_{d_\sigma} : \mathbb{R}^d \to \mathbb{R}$ *is locally Lipschitz, differentiable in every nonzero vector of* $\mathbb{R}^d$, *and in addition* $\Gamma_{d_\sigma}(0) < 1$, *and* $\liminf_{\|p\|\to\infty} \Gamma_{d_\sigma}(p) > (1 - c_c)^{-1}$ *when* $c_c < 1$, *and* $\liminf_{\|p\|\to\infty} \Gamma_{d_\sigma}(p) = +\infty$ *when* $c_c = 1$, *where* $c_c \in (0, 1]$ *is the decay rate used to update the path* p^c *in* (4.6).

Γ2. *When* $d_\sigma > 0$ *is large enough, we have* $\|1/\Gamma_{d_\sigma}\|_\infty \leqslant 1 + 2/d_\sigma$. *Loosely speaking, the stepsize decrease (by* Γ_{d_σ}*) is bounded from below and the bound approaches one with increasing* d_σ.

Γ3. *Given* $\delta_\sigma > 0$ *and* $k \in \mathbb{N}$, *if* $\delta_p > 0$ *and* $d_\sigma \geqslant 1$ *are large enough (i.e., larger than large enough thresholds* $\bar{\delta}_p > 0$ *and* $\bar{d}_\sigma \geqslant 1$, *respectively), then, for any random vector* $p \in \mathbb{R}^d$ *such that* $\|p\|$ *has finite moments, i.e.,* $\mathbb{E}\|p\|^n < +\infty$ *for* $n \in \mathbb{N}$, *and such that* $\mathbb{E}\|p\| \geqslant (1 + \delta_p)\mathbb{E}\|\nu_U^d\|$, *we have*

$$\mathbb{E}\left[\Gamma_{d_\sigma}(p)^{-k}\right] \leqslant 1 - k\frac{\delta_\sigma}{d_\sigma} \ . \tag{4.12}$$

Γ4. *The stepsize change is invariant to rotation, i.e., for every* $p \in \mathbb{R}^d$ *and every orthogonal matrix* $\mathbf{R}$ *of size* $d \times d$, *we have* $\Gamma_{d_\sigma}(\mathbf{R}p) = \Gamma_{d_\sigma}(p)$.

Assumption **Γ**1 is required to prove irreducibility and topological properties of a normalized Markov chain underlying the CMA-ES algorithm, as stated in [53]. Assumptions **Γ**2 and **Γ**3 are needed in this paper in the proof for the ergodicity of this chain. Assumption **Γ**3 translates that the stepsize increases when the (expected) path is "large enough", the increasing factor is controlled by controlling the expected inverse stepsize change to the power of k. It is used in the proofs mainly in Proposition 4.18 to find the cases in which the expected norm squared of the mean m_t divided by a normalization—which includes the stepsize—decreases. Moreover, we use **Γ**4 to extend results (see Proposition 4.6) from an objective function with spherical level sets to an objective function with ellipsoidal level sets. As shown in the next proposition and proven in Section B, the assumptions **Γ**1-**Γ**4 are satisfied by both Γ^1_{CSA} and Γ^2_{CSA}.

Proposition 4.1. Suppose that the sampling distribution ν_U^d satisfies **N**4. Then the stepsize changes Γ^1_{CSA} and Γ^2_{CSA} satisfy **Γ**1–**Γ**4. Moreover, **Γ**3 holds for any value of $k \in \mathbb{N}$ and $\delta_\sigma > 0$.

Overall, in this paper we study the following update of the states of CMA-ES

$$\begin{aligned}
m_{t+1} &= m_t + c_m \sigma_t \sqrt{\mathbf{C}_t} \textstyle\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \\
p_{t+1}^\sigma &= (1-c_\sigma) p_t^\sigma + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}} \textstyle\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \\
p_{t+1}^c &= (1-c_c) p_t^c + \sqrt{c_c(2-c_c)\mu_{\text{eff}}\mathbf{C}_t} \textstyle\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \\
\sigma_{t+1} &= \sigma_t \times \Gamma_{d_\sigma}\left(p_{t+1}^\sigma\right) \\
\mathbf{C}_{t+1} &= (1-c_1-c_\mu)\mathbf{C}_t + c_1 (p_{t+1}^c)(p_{t+1}^c)^\top + c_\mu \textstyle\sum_{i=1}^{\mu} w_i^c \left(\sqrt{\mathbf{C}_t} U_{t+1}^{s_{t+1}(i)}\right)\left(\sqrt{\mathbf{C}_t} U_{t+1}^{s_{t+1}(i)}\right)^\top
\end{aligned} \tag{4.13}$$

where s_{t+1} is the permutation which sorts almost surely the offspring $x_{t+1}^i = m_t + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^i$ for $i = 1, \ldots, \lambda$ and where the i.i.d. process $\mathbf{U} = \{U_t\}_{t \in \mathbb{N}^*}$ is independent from $(m_0, \sigma_0, \mathbf{C}_0, p_0^\sigma, p_0^c)$ and such that the U_1^i, $i = 1, \ldots, \lambda$, are i.i.d. random vectors following the distribution ν_U^d. We have assumed earlier a set of conditions on the hyperparameters of CMA-ES. We summarize them under the following assumption.

H1. *We set $c_1 \geqslant 0, c_m > 0, c_c > 0, c_\sigma > 0$, and $d_\sigma > 0$ covering the definition domains as given in CMA-ES. Additionally, we assume the strict inequalities $c_\mu > 0, c_1 + c_\mu < 1, c_c < 1$, and $c_\sigma < 1$ while CMA-ES also allows equality in these cases.*

Our analysis requires the rank-mu update of the covariance matrix ($c_\mu > 0$) and optionally allows the rank-one update too ($c_1 > 0$). Hence, the analysis does not comprise the case of a rank-one update alone (which however corresponds to a viable algorithm, in particular when λ is small [76]). As c_μ can be chosen arbitrarily small, this limitation has no practical relevance in itself, however we also require c_1/c_μ to be sufficiently small (Propositions 4.14 and 4.15 and Corollary 4.5).

2.2 *Objective function assumptions*

Our goal is to analyze the CMA-ES algorithm on strictly convex-quadratic functions composed with strictly increasing functions. This class of functions includes ill-conditioned problems. Here, we define several assumptions on f that we will use later. We recall first that a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is said to be *scaling-invariant* [142] with respect to a reference point $x^* \in \mathbb{R}^d$ when for every $x, y \in \mathbb{R}^d$ and $\rho > 0$

$$f(x^* + x) \leqslant f(x^* + y) \Rightarrow f(x^* + \rho x) \leqslant f(x^* + \rho y) \ . \tag{4.14}$$

The following first assumption allows to define a normalized Markov chain associated to CMA-ES via scaling-invariance (see Section 2.3) and to obtain its irreducibility (see Theorem 4.2).

F1. *The objective function f is an increasing transformation of a continuous function, is scaling-invariant with respect to a point $x^* \in \mathbb{R}^d$, and has Lebesgue-negligible level sets.*

The class of functions satisfying **F**1 includes composites of strictly increasing functions with positively homogeneous functions and thus all composites of strictly increasing functions with a norm. These latter functions are considered in the following assumptions which we will use to prove geometric ergodicity.

F2. *The objective function $f : \mathbb{R}^d \to \mathbb{R}$ obeys* **F**1 *and has spherical level sets; specifically, there exists an increasing map $g : \mathbb{R} \to \mathbb{R}$ and a point $x^* \in \mathbb{R}^d$ such that $f : x \mapsto g(\|x - x^*\|)$.*

F3. *The objective function $f : \mathbb{R}^d \to \mathbb{R}$ obeys* **F**1 *and has ellipsoidal level sets; specifically, there exists an increasing map $g : \mathbb{R} \to \mathbb{R}$, a point $x^* \in \mathbb{R}^d$, and a matrix $\mathbf{H} \in \mathcal{S}^d_{++}$ satisfying $\lambda_{\min}(\mathbf{H}) = 1$ (without loss of generality), such that $f : x \mapsto g((x - x^*)^\top \mathbf{H}(x - x^*))$.*

Assumption **F**3 is a generalization of **F**2 and equivalent to **F**2 when $\mathbf{H}$ is a homothety, i.e., proportional to $\mathbf{I}_d$. The matrix $\mathbf{H}$ is not the Hessian matrix of $x \mapsto (x - x^*)^\top \mathbf{H}(x - x^*)$ since it is the Hessian of $x \mapsto (x - x^*)^\top \mathbf{H}(x - x^*)/2$. This latter function also belongs to the class of functions obeying **F**3 (for which $\lambda_{\min}(\mathbf{H}) = 1$) which can be seen by using the increasing transformation $y \mapsto y/2$. Hence, in the rest of the paper, we will call the matrix $\mathbf{H}$ from **F**3 the quasi-Hessian matrix of a function f satisfying **F**3.

2.3 *Definition of a normalized chain underlying the CMA-ES algorithm*

To prove convergence of CMA-ES, we investigate the stability of a normalized Markov chain underlying the algorithm. This approach has been used before to prove linear convergence of stepsize adaptive ES without adaptation of a covariance matrix [17, 141].

Before to define a normalized chain, we recall the definition of a Markov chain and of a transition kernel. Consider the measurable space $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ where X is a topological space equipped with its Borel σ-field $\mathcal{B}(\mathsf{X})$. A map $P \colon \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}_+$ is called a *transition kernel* when for every $x \in \mathsf{X}$, the function $\mathsf{A} \in \mathcal{B}(\mathsf{X}) \mapsto P(x, \mathsf{A})$ is a probability measure, and for every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$, the function $x \in \mathsf{X} \mapsto P(x, \mathsf{A})$ is a measurable map.

For every $t \in \mathbb{N}$, consider a random variable θ_t valued in $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ (i.e., a measurable map from the outcome space Ω to the state space X). For each $x \in \mathsf{X}$, we equip the outcome space $(\Omega, \mathcal{A}) = (\mathsf{X}^{\mathbb{N}}, \mathcal{B}(\mathsf{X}^{\mathbb{N}}))$ with a probability measure $\mathbb{P}_x = \mathbb{P}[\,\cdot \mid \theta_0 = x]$ satisfying $\mathbb{P}[\theta_0 = x \mid \theta_0 = x] = 1$. Then, we say that $\Theta = \{\theta_t\}_{t \in \mathbb{N}}$ is a *time-homogeneous Markov chain* with transition kernel P when for every $k \geqslant 1$, $x \in \mathsf{X}$ and $\mathsf{A} \in \mathcal{B}(\mathsf{X})$, we have

$$\mathbb{P}\left[\theta_k \in \mathsf{A} \mid \theta_0 = x\right] = P^k(x, \mathsf{A}) \coloneqq \int_{\mathsf{X}^{k+1}} P(x_{k-1}, \mathsf{A}) P(x_{k-2}, \mathrm{d}x_{k-1}) \dots P(x, \mathrm{d}x_1) \ . \tag{4.15}$$

We can define now a Markov chain by normalizing the states of CMA-ES when minimizing a scaling-invariant objective function. The convergence in law of this Markov chain to an invariant probability measure at a geometric rate implies the linear convergence of CMA-ES as reminded in Theorem 4.1 proven in [53]. Remark that without normalization, the process $\{(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t \in \mathbb{N}}$ presented in Section 2.1 does not define a stationary Markov chain, and we expect $m_t - x^*$, σ_t, and $\mathbf{C}_t$ to tend almost surely to 0.

Consider a measurable function $R\colon \mathcal{S}_{++}^d \to \mathbb{R}_{++}$ that we call normalization function. We remind that R is *positively homogeneous* with degree one when $R(\rho\mathbf{A}) = \rho R(\mathbf{A})$ for every $\rho > 0$ and $\mathbf{A} \in \mathcal{S}_{++}^d$. For the sake of conciseness, we omit to write "with degree one" in the remainder. We use the following cases for R:

R1. $R\colon \mathcal{S}_{++}^d \to \mathbb{R}_{++}$ *is locally Lipschitz, differentiable on an open subset of* $\mathcal{S}_{++}^d$*, positively homogeneous, and satisfies* $R(\mathbf{I}_d) = 1$*,*

R2. $R(\cdot) = \lambda_{\min}(\cdot)$.

Given a positive definite matrix $\mathbf{H}$ with $\lambda_{\min}(\mathbf{H}) = 1$, we generalize **R**2 as

R3. $R(\cdot) = \lambda_{\min}(\mathbf{H}^{1/2} \times \cdot \times \mathbf{H}^{1/2})$.

In particular, **R**2 implies **R**1 [53, Proposition 2.1], and since $\lambda_{\min}(\mathbf{H}) = 1$, **R**3 implies **R**1. Throughout the paper, we consider the normalization **R**2 when the objective function f is spherical, i.e., satisfies **F**2, and the normalization **R**3 when the objective function is ellipsoidal, i.e., satisfies **F**3. Other examples of normalization functions which satisfy **R**1 are $R = \lambda_i(\cdot)$ for $i = 1, \ldots, d$, or $R = \det(\cdot)^{1/d}$.

We define the process $\Theta = \{\theta_t\}_{t\geqslant 1} = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ by the *normalized Markov chain*

$$z_t = \frac{m_t - x^*}{\sqrt{R(\mathbf{C}_t)}\sigma_t},\ p_t = p_t^\sigma,\ q_t = \frac{p_t^c}{\sqrt{R(\mathbf{C}_{t-1})}},\ \mathbf{\Sigma}_t = \frac{\mathbf{C}_t}{R(\mathbf{C}_t)},\ r_t = \frac{R(\mathbf{C}_t)}{R(\mathbf{C}_{t-1})} \quad , \tag{4.16}$$

where $x^* \in \mathbb{R}^d$ is the optimum of f that satisfies **F**1. We define $\mathsf{X} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ the state space of the process Θ that we equip with its Borel σ-field $\mathcal{B}(\mathsf{X})$. The process Θ is a time-homogeneous Markov chain when f is scaling-invariant and R is positively homogeneous. Then, its update rule can be defined independently of $\{(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$. This result was established in [53], and we remind it for the sake of completeness.

Proposition 4.2 ([53, Proposition 2.2])**.** Suppose that the objective function f is an increasing transformation of a continuous scaling-invariant function with Lebesgues-negligible level sets, i.e., satisfies **F**1 with optimum in $x^* \in \mathbb{R}^d$, and that the normalization function R satisfies **R**1. Then, the process $\Theta = \{\theta_t\}_{t\geqslant 1} = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ defined in (4.16) is a time-homogeneous Markov chain and obeys the recursion

$$\begin{aligned}
z_{t+1} &= \frac{z_t + c_m\sqrt{\mathbf{\Sigma}_t}\sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)}}{\sqrt{r_{t+1}}\Gamma_{d_\sigma}(p_{t+1})} \\
p_{t+1} &= (1-c_\sigma)p_t + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)} \\
q_{t+1} &= r_t^{-1/2}(1-c_c)q_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{\Sigma}_t^{1/2}\sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)} \\
\mathbf{\Sigma}_{t+1} &= \frac{\tilde{\mathbf{\Sigma}}_{t+1}}{r_{t+1}} \\
r_{t+1} &= R(\tilde{\mathbf{\Sigma}}_{t+1}) \ ,
\end{aligned} \tag{4.17}$$

where the matrix $\tilde{\mathbf{\Sigma}}_{t+1}$ is defined by

$$\tilde{\mathbf{\Sigma}}_{t+1} = (1 - c_1 - c_\mu)\mathbf{\Sigma}_t + c_1(q_{t+1})(q_{t+1})^\top + c_\mu\sum_{i=1}^\mu w_i^c\left(\sqrt{\mathbf{\Sigma}_t}U_{t+1}^{s_{t+1}(i)}\right)\left(\sqrt{\mathbf{\Sigma}_t}U_{t+1}^{s_{t+1}(i)}\right)^\top \tag{4.18}$$

and $\{U_{t+1}\}_{t\in\mathbb{N}}$ is the i.i.d. process independent of θ_0 used to update the state variables of CMA-ES introduced in Section 2.1, with $U_1 \sim (\nu_U^d)^{\otimes\lambda}$, and $s_{t+1} \in \mathfrak{S}_\lambda$ is a (almost surely

unique assuming that **N**1 holds) permutation that sorts the $f(x^* + z_t + \sqrt{\Sigma_t} U_{t+1}^i)$, $i = 1, \ldots, \lambda$, i.e.,

$$f\left(x^* + z_t + \Sigma_t^{1/2} U_{t+1}^{s_{t+1}(1)}\right) \leqslant \cdots \leqslant f\left(x^* + z_t + \Sigma_t^{1/2} U_{t+1}^{s_{t+1}(\lambda)}\right) \quad . \tag{4.19}$$

We will prove a drift condition associated to the Markov chain of this proposition when f is either a spherical function (satisfies **F**2) or an ellipsoidal function (satisfies **F**3).

When the normalization R is the smallest eigenvalue, the variable r_t satisfies $r_t \geqslant 1 - c_1 - c_\mu$ and we can consider a smaller state space for Θ. This will be useful in our analysis and is formalized in the next proposition.

Proposition 4.3. Suppose that the normalization function R satisfies either **R**2 or **R**3 for some matrix $\mathbf{H} \in \mathcal{S}_{++}^d$, and consider the process $\Theta = \{(z_t, p_t, q_t, \Sigma_t, r_t)\}_{t \geqslant 1}$ defined via (4.16). Then, for every $t \geqslant 1$, the following lower bound on r_t holds

$$r_t \geqslant 1 - c_1 - c_\mu \quad . \tag{4.20}$$

Therefore, under the conditions of Proposition 4.2, this process defines a time-homogeneous Markov chain with state space $\mathbb{R}^{3d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$.

Proof. Let $t \geqslant 1$. Since **R**2 is a particular case of **R**3 for $\mathbf{H} = \mathbf{I}_d$, we only need to consider the case where R satisfies **R**3. Using the update formula for $\mathbf{C}_t$ in (4.8) we have

$$R(\mathbf{C}_t) = \lambda_{\min}\left(\mathbf{H}^{1/2}\left((1 - c_1 - c_\mu)\mathbf{C}_{t-1} + c_1 p_t^c (p_t^c)^\top + c_\mu \mathbf{C}_{t-1}^{1/2} \sum_{i=1}^{\mu} w_i^c U_t^{s_t(i)} (U_t^{s_t(i)})^\top \mathbf{C}_{t-1}^{1/2}\right)\mathbf{H}^{1/2}\right) \quad .$$

Since, by semi-definite positiveness,

$$\lambda_{\min}\left(\mathbf{H}^{1/2}\left(c_1 p_t^c (p_t^c)^\top + c_\mu \mathbf{C}_{t-1}^{1/2} \sum_{i=1}^{\mu} w_i^c U_t^{s_t(i)} (U_t^{s_t(i)})^\top \mathbf{C}_{t-1}^{1/2}\right)\mathbf{H}^{1/2}\right) \geqslant 0$$

and by the inequality $\lambda_{\min}(\mathbf{A} + \mathbf{B}) \geqslant \lambda_{\min}(\mathbf{A}) + \lambda_{\min}(\mathbf{B})$ for symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ (see (4.41)), we obtain $R(\mathbf{C}_t) \geqslant (1 - c_1 - c_\mu)\lambda_{\min}\left(\mathbf{H}^{1/2}\mathbf{C}_{t-1}\mathbf{H}^{1/2}\right) = (1 - c_1 - c_\mu)R(\mathbf{C}_{t-1})$ and thus by definition of r_t, see (4.16), we have $r_t = R(\mathbf{C}_t)/R(\mathbf{C}_{t-1}) \geqslant 1 - c_1 - c_\mu$. □

2.4 *Main results*

We give in this section sufficient conditions for the normalized process underlying CMA-ES defined in (4.16) to be an ergodic Markov chain and to converge at a geometric rate towards an invariant probability distribution.

Before that, we recall some definitions that are useful to the understanding of our results. Let $\Theta = \{\theta_t\}_{t \in \mathbb{N}}$ be a time-homogeneous Markov chain valued in the state space $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$, and denote P its transition kernel and P^k its k-step transition kernel connected to P by induction as $P^{k+1}(x, A) = \int P(y, A) P^k(x, dy)$.

A transition kernel P is *irreducible* when there exists a nontrivial nonnegative measure φ on $\mathcal{B}(\mathsf{X})$ such that, for every $x \in \mathsf{X}$ and $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ with $\varphi(\mathsf{A}) > 0$, there exists a positive integer k, with

$P^k(x, \mathsf{A}) > 0$.[2] We then say that P is *φ-irreducible*. An irreducible transition kernel P is *aperiodic*[3] if for every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ with $\varphi(\mathsf{A}) > 0$ for some irreducibility measure φ of P, and every $x \in \mathsf{X}$, there exists $k_0 \in \mathbb{N}$ such that

$$\forall k \geqslant k_0, \quad P^k(x, \mathsf{A}) > 0 \ . \tag{4.21}$$

A set $\mathsf{K} \in \mathcal{B}(\mathsf{X})$ is a *small set* when there exists an integer $m \in \mathbb{N}$ and a nontrivial measure ν_m on $\mathcal{B}(\mathsf{X})$ such that $P^m(x, \mathsf{A}) \geqslant \nu_m(\mathsf{A})$ for every $x \in \mathsf{K}$ and every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$. Additionally, an *invariant measure* of P is a measure π on $\mathcal{B}(\mathsf{X})$ such that $\int_{\mathsf{X}} P(y, \mathsf{A})\pi(\mathrm{d}y) = \pi(\mathsf{A})$ for every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$. When there exists an invariant probability measure π of P, we say that P is *positive*.

Define for $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ the occupation time in A as the random variable $\eta_{\mathsf{A}} = \sum_{k\geqslant 1} \mathbb{1}\{\theta_k \in \mathsf{A}\}$ and suppose that P is φ-irreducible for some nontrivial measure φ on $\mathcal{B}(\mathsf{X})$. Then, P is *recurrent* if for $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ with $\varphi(\mathsf{A}) > 0$ and for $x \in \mathsf{A}$, we have $\mathbb{E}[\eta_{\mathsf{A}} \mid \theta_0 = x] = +\infty$, and P is *Harris-recurrent* if moreover $\mathbb{P}[\eta_{\mathsf{A}} = +\infty \mid \theta_0 = x] = 1$.

Consider a measurable function $V : \mathsf{X} \to [1, +\infty]$. The transition kernel P is said to be *V-geometrically ergodic* when it is positive Harris-recurrent, its invariant probability measure π is such that V is π-integrable, and there exists a real constant $r > 1$ such that for all $x \in \mathsf{X}$

$$\sum_{k=1}^{\infty} r^k \|P^k(x, \cdot) - \pi\|_V < +\infty \ . \tag{4.22}$$

In the previous equation, for every (signed) measure ν on $\mathcal{B}(\mathsf{X})$ we denote $\|\nu\|_V = \sup_{|f|\leqslant V} \int f \mathrm{d}\nu$. When $V(\cdot) = 1$, we simply say that P is *geometrically ergodic*.

We have introduced the notion of irreducibility, aperidocity, recurrence and geometric ergodicity for a transition kernel, however throughout the paper, we will also label the corresponding Markov chain respectively.

We will prove the geometric ergodicity of the normalized chain (which implies positive Harris-recurrence) and integrability properties with respect to the invariant measure π. Those properties are needed to prove the linear convergence of CMA-ES as stated in the next theorem whose proof can be found in [53].

Theorem 4.1 ([53, Proposition 2.3])**.** Suppose that the objective function f satisfies **F**1 and the normalization function R satisfies **R**1. Suppose that the process $\Theta = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ defined by (4.16) is a time-homogeneous aperiodic positive Harris recurrent Markov chain, with unique invariant probability measure denoted π. Assume moreover that the functions

$$(z, p, q, \mathbf{\Sigma}, r) \in \mathsf{X} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times \mathbb{R}_{++} \mapsto \begin{cases} \log \|z\| \\ \log \Gamma_{d_\sigma}(p) \\ \log R(\mathbf{\Sigma}) \end{cases}$$

are integrable with respect to π. Then, the distance of the mean m_t in CMA-ES to the optimum x^* of f behaves geometrically when $t \to +\infty$, and for every initialization $(m_0, p_0^\sigma, p_0^c, \sigma_0, \mathbf{C}_0) \in \mathbb{R}^{3d} \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d$, we have

$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty} \frac{1}{T} \log \frac{\sigma_T}{\sigma_0} + \frac{1}{2T} \log \frac{R(\mathbf{C}_T)}{R(\mathbf{C}_0)} = -\mathrm{CR} \tag{4.23}$$

[2] This is referred to as φ-irreducibility in other works [110, Section 4.2.1]. Alternative definitions for irreducibility can be found in the literature [41, Definition 9.2.1], which however are equivalent to the one we rely on here under mild assumptions [41, Theorem 9.2.6] .

[3] Equivalently, we can define the period of an irreducible Markov chain [110, Section 5.4.3], [41, Definition 9.3.5], the latter is aperiodic if and only if its period is one [41, Theorem 9.3.10].

and

$$\lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}\right] = \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2}\log \frac{R(\mathbf{C}_{t+1})}{R(\mathbf{C}_t)}\right] = -\mathrm{CR} \tag{4.24}$$

for some constant $\mathrm{CR} \in \mathbb{R}$. If furthermore $\mathrm{CR} > 0$, we say that CMA-ES converges linearly.

To prove geometric ergodicity and thus positivity and Harris-recurrence, we will prove a drift condition outside a small set. We thus need to identify some small sets of the state-space. In previous work, we have proven the irreducibility, aperiodicity and the property that compact sets are small for the normalized Markov chain (4.17) when f is an increasing transformation of a continuous scaling-invariant function with Lebesgue-negligible level sets and the normalization function R satisfies **R**1 [53, Theorem 3.1]. We remind below the theorem summarizing those results.

Theorem 4.2 ([53, Theorem 3.1])**.** Suppose that the objective function f, the normalization function R, the stepsize change Γ_{d_σ} and the sampling distribution ν_U^d satisfy **F**1, **R**1, **Γ**1, **N**1, respectively. Let $\Theta = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ be the Markov chain defined via (4.16). Then,

(i) if $c_c, c_\sigma \neq 1$, $c_\mu \neq 0$ and $1 - c_c \neq (1 - c_\sigma)\sqrt{1 - c_1 - c_\mu}$, then Θ is an irreducible aperiodic T-chain, and thus compact sets of $\mathsf{X} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ are small;

(ii) if $c_c \neq 1$, $c_\sigma = 1$ and $c_\mu \neq 0$, then $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ is an irreducible aperiodic T-chain, and thus compact sets of $\mathbb{R}^{2d} \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ are small;

(iii) if $c_c = 1$ and $c_\sigma \neq 1$, then $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\geqslant 1}$ is an irreducible aperiodic T-chain, and thus compact sets of $\mathbb{R}^{2d} \times R^{-1}(\{1\})$ are small;

(iv) if $c_c = c_\sigma = 1$, then $\{(z_t, \mathbf{\Sigma}_t)\}_{t\geqslant 1}$ is an irreducible aperiodic T-chain, and thus compact sets of $\mathsf{X} = \mathbb{R}^d \times R^{-1}(\{1\})$ are small.

Note that the statement of Theorem 4.2 slightly differs from the statement of [53, Theorem 3.1] because we have grouped assumptions together in this paper. For instance the assumption **R**1 on the normalization function R corresponds to three assumptions in [53].

Our first main result is the key to prove geometric ergodicity and V-integrability of the normalized process (4.16) from which we will deduce positivity, Harris recurrence, and integrability properties needed for the linear convergence of CMA-ES. It is formulated as a multi-step geometric drift towards a compact (and thus small set) of the state space that holds under restricted conditions on the hyperparameters of CMA-ES when optimizing ellipsoidal functions (obeying **F**3). The result is formulated in the next proposition and its proof will occupy a large part of Section 3.

Proposition 4.4. Consider hyperparameters of CMA-ES satisfying **H**1, an ellipsoidal objective function f via **F**3 for some quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}_{++}^d$ and a normalization function R satisfying **R**3 with the matrix $\mathbf{H}$. Suppose moreover that the sampling distribution ν_U^d is standard normal, and that the stepsize change Γ_{d_σ} and the weights $\mathbf{w}_m$, $\mathbf{w}_c$ satisfy **Γ**1–**Γ**4 and **W**1, respectively.

Suppose that $\mu \leqslant \lambda/2$, that μ_{eff} is sufficiently large, $c_1 + c_\mu < 1$, d_σ^{-1} is sufficiently larger than c_μ and $c_m^{3/2}$, and sufficiently smaller than c_m, both $2c_c$ and c_σ are larger than c_μ, and c_1 is sufficiently small. Moreover, assume that either $c_c = 1$ or $c_\sigma = 1$. Then, there exist $\varepsilon > 0$, a compact subset K of $\mathsf{X} := \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$ defined via (4.142) and

(4.137)–(4.141), and positive constants $\beta \geqslant 1$, $\gamma_p > 0$, $\gamma_q > 0$, $\gamma_r > 0$ such that the potential function $V : \mathsf{X} \to [1, +\infty)$ defined as

$$V(z, p, q, \boldsymbol{\Sigma}, r) = \|\mathbf{H}^{1/2} z\|^2 + \beta \lambda_1(\mathbf{H}^{1/2} \boldsymbol{\Sigma} \mathbf{H}^{1/2}) + \gamma_p \|p\| + \gamma_q \|\mathbf{H}^{1/2} q\|^2 + \gamma_r r \tag{4.25}$$

and the Markov chain $\Theta = \{\theta_t\}_{t \in \mathbb{N}} = \{(z_t, p_t, q_t, \boldsymbol{\Sigma}_t, r_t)\}_{t \in \mathbb{N}}$ obeying (4.17) with $\theta_0 \notin \mathsf{K}$ satisfies

$$\mathbb{E}_0 \left[V(z_t, p_t, q_t, \boldsymbol{\Sigma}_t, r_t) \right] \leqslant (1 - \varepsilon c_\mu) V(z_0, p_0, q_0, \boldsymbol{\Sigma}_0, r_0) \quad \text{for } t = 1 \text{ or } 2. \tag{D}$$

From this proposition, together with the irreducibility Theorem 4.2, we deduce as a consequence the second main result of the paper, Theorem 4.3, namely that under some conditions on the hyperparameters of CMA-ES, the normalized process (4.16) is a geometrically ergodic Markov chain.

Theorem 4.3. Consider an ellipsoidal objective function f satisfying **F**3 for some quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}^d_{++}$ and a normalization function R satisfying **R**3 with the matrix $\mathbf{H}$. Suppose moreover that the sampling distribution of CMA-ES ν_U^d is standard normal, and that the stepsize change Γ_{d_σ} and the weights $\mathbf{w}_m$, $\mathbf{w}_c$ satisfy **Γ**1–**Γ**4 and **W**1, respectively. Assume that the hyperparameters of CMA-ES satisfy **H**1.
Suppose additionally that $\mu \leqslant \lambda/2$, that μ_{eff} is sufficiently large, $c_1 + c_\mu < 1$, d_σ^{-1} is sufficiently larger than $c_\mu > 0$ and $c_m^{3/2} > 0$, and sufficiently smaller than c_m, both $2c_c$ and c_σ are larger than c_μ, and c_1 is sufficiently small.

(i) Consider the potential function $V : \mathbb{R}^{2d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty) \to [1, +\infty)$ defined by

$$V(z, q, \boldsymbol{\Sigma}, r) = \|\mathbf{H}^{1/2} z\|^2 + \beta \lambda_1(\mathbf{H}^{1/2} \boldsymbol{\Sigma} \mathbf{H}^{1/2}) + \gamma_q \|\mathbf{H}^{1/2} q\|^2 + \gamma_r r \tag{4.26}$$

for well-chosen constants $\beta \geqslant 1, \gamma_q, \gamma_r > 0$. If $c_c \neq 1$ and $c_\sigma = 1$, then the Markov chain $\Theta = \{(z_t, q_t, \boldsymbol{\Sigma}_t, r_t)\}_{t \in \mathbb{N}}$ defined via (4.17) and $\theta_0 \in \mathbb{R}^{2d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$ is geometrically ergodic and the function V is π-integrable, where π is the invariant probability measure of Θ.

(ii) Consider the potential function $V : \mathbb{R}^{2d} \times R^{-1}(\{1\}) \to [1, +\infty)$ defined by

$$V(z, p, \boldsymbol{\Sigma}) = \|\mathbf{H}^{1/2} z\|^2 + \beta \lambda_1(\mathbf{H}^{1/2} \boldsymbol{\Sigma} \mathbf{H}^{1/2}) + \gamma_p \|p\| \tag{4.27}$$

for well-chosen constants $\beta \geqslant 1, \gamma_p, > 0$. If $c_c = 1$ and $c_\sigma \neq 1$, then the Markov chain $\Theta = \{(z_t, p_t, \boldsymbol{\Sigma}_t)\}_{t \in \mathbb{N}}$ defined via (4.17) and $\theta_0 \in \mathbb{R}^{2d} \times R^{-1}(\{1\})$ is geometrically ergodic and the function V is π-integrable, where π is the invariant probability measure of Θ.

(iii) Consider the potential function $V : \mathbb{R}^d \times R^{-1}(\{1\}) \to [1, +\infty)$ defined by

$$V(z, \boldsymbol{\Sigma}) = \|\mathbf{H}^{1/2} z\|^2 + \beta \lambda_1(\mathbf{H}^{1/2} \boldsymbol{\Sigma} \mathbf{H}^{1/2}) \tag{4.28}$$

for a well-chosen constant $\beta \geqslant 1$. If $c_c = c_\sigma = 1$, then the Markov chain $\Theta = \{(z_t, \boldsymbol{\Sigma}_t)\}_{t \in \mathbb{N}}$ defined via (4.17) and $\theta_0 \in \mathbb{R}^d \times R^{-1}(\{1\})$ is geometrically ergodic and the function V is π-integrable, where π is the invariant probability measure of Θ.

The different subcases correspond to different variants of CMA-ES where c_σ or c_c are set to 1. The associated Markov chains have fewer variables, namely p_σ^t (when $c_\sigma = 1$) or p_c^t (when $c_c = 1$) are redundant. As we will explain in the next section, we will deduce the geometric ergodicity of those subcases by considering the single geometric decrease condition (D) proven in Proposition 4.4 associated to the (possibly redundant) Markov chain (4.16).

3 Proof of the main results

We establish in this section the geometric ergodicity of the Markov chain (4.16) defined by the normalization of the states of CMA-ES. More precisely, we provide in Section 3.1 results we rely on to prove Theorem 4.3 as well as the key ideas for its proof. In Section 3.2, we prove Theorem 4.3, and the rest of the section is devoted to the proof of Proposition 4.4.

3.1 Methodology to prove the geometric ergodicity of the normalized Markov chain

To prove the geometric ergodicity of the Markov chain (4.16), we rely on the following criterion for geometric ergodicity, that we call state-dependent (or multi-step) geometric drift condition, or Foster-Lyapunov condition.

Theorem 4.4[110, Theorem 19.1.3]**.** Suppose that $\Theta = \{\theta_t\}_{t\geqslant 0}$ is an irreducible aperiodic Markov chain valued in a measured space $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ and with transition kernel P. Let $n\colon \mathsf{X} \to \mathbb{N}^*$ and $V\colon \mathsf{X} \to [1, +\infty)$ be two measurable functions. If there exist a small set $\mathsf{K} \in \mathcal{B}(\mathsf{X})$ such that V is bounded on K and positive constants $\rho < 1$ and $b < +\infty$ such that

$$\int P^{n(x)}(x, \mathrm{d}y)V(y) \leqslant \rho^{n(x)} \times (V(x) + b\mathbb{1}_{x\notin\mathsf{K}}) \quad \text{for } x \in \mathsf{X} \ , \tag{4.29}$$

then Θ is geometrically ergodic. Moreover, we have

$$\sum_{t\in\mathbb{N}} r^t \|P^t(x, \cdot) - \pi\| \leqslant RV(x) \quad \text{for } x \in \mathsf{X} \tag{4.30}$$

for some constants $R < \infty$ and $r > 1$, where π is the unique invariant probability measure of Θ.

Contrary to a single-step drift condition, this theorem does not imply the V-geometric ergodicity of the Markov chain (only its 1-geometric ergodicity). We are however able to deduce that if Theorem 4.4 holds, the potential function V is integrable with respect to the invariant probability measure which is a property needed to apply a Law of Large Number in order to prove the convergence of CMA-ES.

Corollary 4.1. Assume that Theorem 4.4 applies. Then, there exists a measurable function $\tilde{n}\colon \mathsf{X} \to \mathbb{N}^*$ such that P is $\tilde{V}$-geometrically ergodic, where $\tilde{V}(x) = P^{\tilde{n}(x)-1}V(x)$ for $x \in \mathsf{X}$. If moreover, the function $n\colon \mathsf{X} \to \mathbb{N}^*$ in Theorem 4.4 is bounded, then V is π-integrable.

Proof. For $x \in \mathsf{X}$, define $\tilde{n}(x) = \min\{n \geqslant 1 \mid \int P^n(x, \mathrm{d}y)V(y) \leqslant \rho^n \times (V(x) + b\mathbb{1}_{x\notin\mathsf{K}})\}$. Since (4.29) holds, $\tilde{n}(x)$ is well-defined and is a positive integer. For $x \in \mathsf{X}$, define as well the potential function $\tilde{V}(x) = \int P^{\tilde{n}(x)-1}(x, \mathrm{d}y)V(y)$ if $\tilde{n}(x) > 1$ and $\tilde{V}(x) = V(x)$ when $\tilde{n}(x) = 1$. By definition of $\tilde{n}(x)$, we get (the term $b\mathbb{1}_{x\notin\mathsf{K}}$ does not remain in the lower bound

when $\tilde{n}(x) = 1$):

$$\tilde{V}(x) \geqslant \rho^{\tilde{n}(x)-1} \times V(x)$$

and consequently

$$\int P(x, \mathrm{d}y)\tilde{V}(y) = \int P^{\tilde{n}(x)}(x, \mathrm{d}y)V(y) \leqslant \rho^{\tilde{n}(x)} \times (V(x) + b\mathbb{1}_{x\notin\mathsf{K}}) \leqslant \rho \times \tilde{V}(x) + b\mathbb{1}_{x\notin\mathsf{K}} \ .$$

Therefore, $\tilde{V}$ satisfies a one-step drift condition [110, Theorem 15.0.1] which yields to the $\tilde{V}$-geometric ergodicity of P and thus to the π-integrability of $\tilde{V}$. Moreover, if $\tilde{n}$ is bounded by above by a constant $\bar{n} < +\infty$, then $V(x) \leqslant \rho^{1-\bar{n}}\tilde{V}(x)$ for every $x \in \mathsf{X}$ and thus V is π-integrable. □

Our candidate Foster-Lyapunov function V to satisfy (4.29) was given in Proposition 4.4 as

$$V(z, p, q, \mathbf{\Sigma}, r) = \|\mathbf{H}^{1/2}z\|^2 + \beta\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}) + \gamma_p\|p\| + \gamma_q\|\mathbf{H}^{1/2}q\|^2 + \gamma_r r \tag{4.25}$$

for some constants $\beta, \gamma_p, \gamma_q, \gamma_r$ that we will carefully choose. By Theorem 4.2, compact subsets of the state space $\mathsf{X} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$ are small sets for the Markov chains satisfying the assumptions of the theorem. Hence, we build a compact subset K of X that satisfies that there exists $\rho \in (0, 1)$ such that for every initial condition $\theta_0 = (z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \in \mathsf{X} \setminus \mathsf{K}$ taken outside K, there exists an iteration $t = t(\theta_0) > 0$ such that

$$\mathbb{E}_{\theta_0}\left[V(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\right] \leqslant \rho^t \times V(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0). \tag{4.31}$$

Here and throughout the paper, for $\theta \in \mathsf{X}$, $\mathbb{E}_\theta$ denotes the expectation associated to the probability function $\mathbb{P}_\theta = \mathbb{P}[\,\cdot \mid \theta_0 = \theta]$ satisfying $\mathbb{P}_\theta[\theta_0 = \theta] = 1$. We use the condensed notation $\mathbb{P}_{\theta_0} =: \mathbb{P}_0$ and $\mathbb{E}_{\theta_0} =: \mathbb{E}_0$ to refer to the conditional probability and expectation with respect to the random variable θ_0.

We partition the state space X in order to construct a compact set, outside of which a multistep drift condition holds. First, we consider the case when $\lambda_1(\mathbf{\Sigma}_0)$ is sufficiently large, i.e., larger than a well-chosen constant $M_{\mathbf{\Sigma}} > 0$, and also sufficiently larger than $\|\mathbf{\Sigma}_0^{1/2}z_0\|$, i.e., $\|\mathbf{\Sigma}_0^{1/2}z_0\| \leqslant M_z\lambda_1(\mathbf{\Sigma}_0)$ for a well-chosen constant $M_z > 0$ (see Section 3.4). Then, $\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2})$ dominates the drift and we prove its decrease in expectation. To this end, Proposition 4.12 shows that the mean has a comparatively small impact on the expected rank-mu update of the covariance matrix. In Corollary 4.5, we prove

$$\mathbb{E}_0\left[\lambda_1\Big(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\Big)\right] \leqslant (1 - \delta c_\mu)\lambda_1\Big(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\Big) + 3c_1\|\mathbf{H}^{1/2}q_0\|^2 \tag{4.32}$$

for some $\delta > 0$, when the learning rates $c_1 \geqslant 0$ and $c_\mu > 0$ are sufficiently small. As stated in Lemma 4.7, this allows to find a decrease of the conditional expectation of the potential function in one step, i.e., we obtain (4.31) with $t = 1$.

Second, we consider the case when the normalized mean dominates the largest eigenvalue of $\mathbf{\Sigma}_0$, i.e., when there exists M_z such that $\|\mathbf{\Sigma}_0^{1/2}z_0\| \geqslant M_z\lambda_1(\mathbf{\Sigma}_0)$. Proposition 4.18 (ii) and (iii) show that the expected norm squared of the mean (weighted by the matrix $\mathbf{H}^{1/2}$) can be bounded in one or two steps. As a consequence, Lemma 4.8 (i) finds a decrease of the conditional expectation of the potential function in one or two steps, i.e., we have (4.31) with $t = 1$ or 2.

Last, when $\lambda_1(\mathbf{\Sigma}_0) \leqslant M_{\mathbf{\Sigma}}$ and $\|\mathbf{\Sigma}_0^{1/2}z_0\| \leqslant M_z\lambda_1(\mathbf{\Sigma}_0)$, and when either $\|p_0\| \geqslant M_p$, $\|q_0\| \geqslant M_q$, or $r_0 \geqslant M_r$, for some well-chosen constants $M_p, M_q, M_r > 0$, Lemma 4.8 (ii), (iii), (iv) prove the

decrease of the conditional expectation of the potential function in one step in those cases. From these subcases, we define the compact set K as

$$\left\{(z,p,q,\boldsymbol{\Sigma},r)\in\mathsf{X}\;\middle|\;\lambda_1(\boldsymbol{\Sigma})<M_{\boldsymbol{\Sigma}},\|\boldsymbol{\Sigma}^{1/2}z\|<M_y\lambda_1(\boldsymbol{\Sigma}),\|p\|<M_p,\|q\|<M_q,r<M_r\right\} \tag{4.33}$$

outside of which a geometric (one or two steps) drift condition holds and conclude with Proposition 4.4.

From the geometric multistep drift condition outside K in Proposition 4.4, we deduce Theorem 4.3 which covers different hyperparameters settings yielding different normalized Markov chains. For instance, if the decay rate c_σ is set to 1, i.e., if we do not have cumulation for the stepsize update, the variable p_t in the Markov chain (4.17) can be dropped. In order to deduce geometric ergodicity for these different Markov chains at once, we use the theorem below that uses the concept of redundant and projected chains [53].

More precisely, if a Markov chain $\{(\phi_t,\xi_t)\}_{t\in\mathbb{N}}$ valued in a state space $\mathsf{X}\times\mathsf{Y}$ is such that the projected process $\{\phi_t\}_{t\in\mathbb{N}}$ valued in X is a Markov chain as well, then we call $\{(\phi_t,\xi_t)\}_{t\in\mathbb{N}}$ a *redundant* Markov chain, and $\{\phi_t\}_{t\in\mathbb{N}}$ a *projected* Markov chain. By definition for all $(x,y)\in\mathsf{X}\times\mathsf{Y}$, for all A, for all $\mathcal{B}(\mathsf{X})$, for all $m\geqslant 1$

$$P^m(x,A)=\tilde{P}^m((x,y),A\times Y)$$

where P^m and $\tilde{P}^m$ denote the m steps transition kernels of the Markov chains $\{\phi_t\}_{t\in\mathbb{N}}$ and $\{(\phi_t,\xi_t)\}_{t\in\mathbb{N}}$. We present in the next theorem how to transfer a geometric drift condition from a redundant Markov chain to a projected irreducible and aperiodic chain. More precisely assume that $(\phi,\chi)\mapsto\tilde{V}(\phi,\chi)$ is a potential function that satisfies a state-dependent geometric drift condition for the redundant chain and that there exists $\chi^*\in\mathsf{Y}$ such that for all $(\phi,\chi)\in\mathsf{X}\times\mathsf{Y}$,

$$\tilde{V}(\phi,\chi^*)\leqslant\tilde{V}(\phi,\chi)\ . \tag{4.34}$$

Then the function $\phi\mapsto V(\phi)=\tilde{V}(\phi,\chi^*)$ is a potential function that satisfies a state-dependent drift condition for the redundant process.

The condition (4.34) is satisfied for instance when the potential $\tilde{V}$ has a separated form $\tilde{V}(\phi,\chi)=V(\phi)+W(\chi)$ where W admits a global minimum χ^*.

Theorem 4.5. Consider a redundant Markov chain $\tilde{\Phi}=\{(\phi_t,\chi_t)\}_{t\in\mathbb{N}}$ valued in the product Polish space $\mathsf{X}\times\mathsf{Y}$, equipped with its Borelian σ-fields $\mathcal{B}(\mathsf{X}\times\mathsf{Y})$, i.e., such that the projected process $\Phi=\{\phi_t\}_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain valued in X. Denote P and $\tilde{P}$ the transition kernels of Φ and $\tilde{\Phi}$, respectively.

(i) Suppose that $\tilde{\mathsf{K}}\subset\mathsf{X}\times\mathsf{Y}$ is a small set for the redundant Markov chain $\tilde{\Phi}$. Then, for every $\mathsf{A}\in\mathcal{B}(\mathsf{Y})$, the set

$$\mathsf{K}=\{\phi\in\mathsf{X}\mid\exists\chi\in\mathsf{A},(\phi,\chi)\in\tilde{\mathsf{K}}\}=\pi_{\mathsf{X}}((\mathsf{X}\times\mathsf{A})\cap\tilde{\mathsf{K}})$$

where π_{X} is the projection on X, is small for the projected Markov chain Φ.

(ii) Suppose that Φ is irreducible and aperiodic. Let $\tilde{V}\colon\mathsf{X}\times\mathsf{Y}\to[1,+\infty)$ be a measurable function, and suppose that there exists $\chi^*\in\mathsf{Y}$ such that

$$\text{for every }(\phi,\chi)\in\mathsf{X}\times\mathsf{Y},\quad\tilde{V}(\phi,\chi^*)\leqslant\tilde{V}(\phi,\chi)\ . \tag{4.35}$$

Suppose that there exists a measurable function $\tilde{n}\colon\mathsf{X}\times\mathsf{Y}\to\mathbb{N}^*$, a small (respectively compact) set $\tilde{\mathsf{K}}\in\mathcal{B}(\mathsf{X}\times\mathsf{Y})$, positive constants $\tilde{\rho}<1$ and $\tilde{b}<+\infty$ such that $\tilde{V}$ is

bounded on $\tilde{\mathsf{K}}$ and

$$\int \tilde{P}^{\tilde{n}(x,y)}\left((x,y),\mathrm{d}(z,w)\right)\tilde{V}(z,w) \leqslant \tilde{\rho}^{\tilde{n}(x,y)} \times \left(\tilde{V}(x,y) + \tilde{b}\mathbb{1}_{(x,y)\in\tilde{\mathsf{K}}}\right) \quad \text{for } (x,y) \in \mathsf{X}\times\mathsf{Y}.$$

Then there exist a measurable function $n\colon \mathsf{X} \to \mathbb{N}^*$, a small (respectively compact) set $\mathsf{K} \in \mathcal{B}(\mathsf{X})$, positive constants $\rho < 1$ and $b < +\infty$ such that

$$\int P^{n(x)}\left(x,\mathrm{d}z\right)V(z) \leqslant \rho^{n(x)} \times \left(V(x) + b\mathbb{1}_{x\in\mathsf{K}}\right) \quad \text{for } x \in \mathsf{X},$$

where $V(x) = \tilde{V}(x,\chi^*)$ defines a measurable function bounded on K.

Proof. For (i), since $\tilde{\mathsf{K}}$ is small for $\tilde{\Phi}$, then there exist $m \in \mathbb{N}^*$ and a nontrivial measure $\tilde{\nu}_m$ on $\mathcal{B}(\mathsf{X}\times\mathsf{Y})$ such that

$$\tilde{P}^m((x,y),\tilde{\mathsf{B}}) \geqslant \tilde{\nu}_m(\tilde{\mathsf{B}}) \quad \text{for every } (x,y) \in \tilde{\mathsf{K}} \text{ and } \tilde{\mathsf{B}} \in \mathcal{B}(\mathsf{X}\times\mathsf{Y}) \ .$$

Define $\nu_m(\mathsf{B}) = \tilde{\nu}_m(\mathsf{B}\times\mathsf{Y})$ for $\mathsf{B} \in \mathcal{B}(\mathsf{X})$, which defines a nontrivial measure on $\mathcal{B}(\mathsf{X})$. Let $\mathsf{A} \in \mathcal{B}(\mathsf{Y})$, let $x \in \mathsf{K} = \{\phi \in \mathsf{X} \mid \exists \chi \in \mathsf{A}, (\phi,\chi) \in \tilde{K}\}$, let $a \in \mathsf{A}$ such that $(x,a) \in \tilde{\mathsf{K}}$, and $\mathsf{B} \in \mathcal{B}(\mathsf{X})$. Then,

$$P^m(x,\mathsf{B}) = \tilde{P}^m((x,a),\mathsf{B}\times\mathsf{Y}) \geqslant \tilde{\nu}_m(\mathsf{B}\times\mathsf{Y}) = \nu_m(\mathsf{B}).$$

In addition since $\mathsf{K} = \pi_{\mathsf{X}}((\mathsf{X}\times\mathsf{A})\cap\tilde{\mathsf{K}})$ with π_{X} the canonical projection on X, then K is measurable as the projection by π_{X} of the measurable set $(\mathsf{X}\times\mathsf{A})\cap\tilde{\mathsf{K}}$ [34, Theorem 2.12]. This proves that K is a small set for Φ.
For (ii), define for $x \in \mathsf{X}$

$$V(x) = \tilde{V}(x,\chi^*) \quad \text{and} \quad n(x) = \tilde{n}(x,\chi^*)$$

and set $b = \tilde{b}$, $\tilde{\rho} = \rho$ and $\mathsf{K} = \{x \in \mathsf{X} \mid (x,\chi^*) \in \tilde{\mathsf{K}}\}$. By (i), if $\tilde{\mathsf{K}}$ is small, then K is a small set for Φ (since K is the set introduced in (i) with $\mathsf{A} = \{\chi^*\}$). Moreover, if $\tilde{\mathsf{K}}$ is compact, then K is compact as well since K is the continuous image by π_{X} of $(\mathsf{X}\times\mathsf{A})\cap\tilde{\mathsf{K}}$ which is compact as the intersection of a compact and a closed set. Besides, we have

$$\begin{aligned}\int P^{n(x)}\left(x,\mathrm{d}z\right)V(z) &= \int P^{n(x)}\left(x,\mathrm{d}z\right)\tilde{V}(z,\chi^*) = \int \tilde{P}^{\tilde{n}(x,\chi^*)}\left((x,\chi^*),\mathrm{d}(z,w)\right)\tilde{V}(z,\chi^*)\\ &\leqslant \int \tilde{P}^{\tilde{n}(x,\chi^*)}\left((x,\chi^*),\mathrm{d}(z,w)\right)\tilde{V}(z,w) \leqslant \tilde{\rho}^{\tilde{n}(x,\chi^*)} \times \left(\tilde{V}(x,\chi^*) + \tilde{b}\mathbb{1}_{(x,\chi^*)\in\tilde{\mathsf{K}}}\right)\\ &= \rho^{n(x)} \times \left(V(x) + b\mathbb{1}_{x\in\mathsf{K}}\right) \ ,\end{aligned}$$

ending the proof. □

This theorem is useful to deduce the proof of Theorem 4.3 from Proposition 4.4.

3.2 Proof of Theorem 4.3

In the previous section, we have presented in Theorem 4.5 how the geometric ergodicity of a projected chain can be deduced from the geometric ergodicity of an associated redundant chain. This is needed

to prove our main result stated in Theorem 4.3 together with the multi-step drift condition presented in Proposition 4.4 that establishes the geometric ergodicity for the redundant chain. Since we have all the tools to prove Theorem 4.3 (except the proof of Proposition 4.4), we present here its proof while the rest of the paper will focus on proving Proposition 4.4.

Proof of Theorem 4.3. By Theorem 4.2, in all cases (i)-(iii), Θ is an irreducible aperiodic time-homogeneous Markov chain. Since by Proposition 4.4, $\beta \geqslant 1$ and since $\lambda_1(\mathbf{\Sigma}) \geqslant 1$ for $\mathbf{\Sigma} \in R^{-1}(\{1\})$, we have in cases (i)-(iii) that $V \geqslant 1$. We use Theorem 4.5 to prove our results. In case (i), the process $\tilde{\Theta} = \{(z_t, p_t, q_r, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ defines a redundant Markov chain such that the projected process $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ is a Markov chain as well, for which compact subsets of $\mathbb{R}^{2d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$ are small, by Theorem 4.2 and Proposition 4.3. Besides, the potential function $\tilde{V}$ given by (4.25) satisfies the drift condition (4.29) and for every $(z, p, q, \mathbf{\Sigma}, r) \in \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$, we have

$$\tilde{V}(z, 0, q, \mathbf{\Sigma}, r) \leqslant \tilde{V}(z, p, q, \mathbf{\Sigma}, t) \ .$$

Yet, the potential function (4.26) satisfies $V(z, q, \mathbf{\Sigma}, r) = \tilde{V}(z, 0, q, \mathbf{\Sigma}, t)$. Then, by Theorem 4.5, V satisfies condition (4.29) for some compact set $\mathsf{K} \subset \mathbb{R}^{2d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$ which is thus small. Therefore, by Theorem 4.4, Θ is geometrically ergodic, and by Corollary 4.1, V is π-integrable. The proofs in cases (ii) and (iii) are similar than for case (i): we define the potential function for the projected chain by setting the redundant parameter to 0. The details are left to the reader. □

3.3 *Preliminary definitions and results for the proof of Proposition 4.4*

The proof of Proposition 4.4, i.e., that a multi-step geometric drift condition holds for the normalized Markov chain defined earlier via (4.16), will occupy the rest of the paper. To start, in this section, we give a few preliminary results that will be used later.

3.3.1 *Some useful inequalities and their consequences*

Before we give more specific results, we remind the following inequality that we will use several times throughout the proof of Proposition 4.4. It is sometimes referred to as the Harris-Fortuin, Kasteleyn and Ginibre (Harris-FKG) inequality [58, Theorem 2.4], and elementary proofs can be found in the literature [139, 130]. We state here the following variant and detail its proof in Section C.

Theorem 4.6 (Harris-FKG inequality)**.** Let X be a real-valued random variable, Y a random variable valued in a space Y and independent of X, $g\colon \mathbb{R} \to \mathbb{R}$ a nondecreasing function, and $h\colon \mathbb{R} \times \mathsf{Y} \to \mathbb{R}$ a nonincreasing function with respect to its first variable (i.e., for Y-almost every y, $x \in \mathbb{R} \mapsto h(x, y)$ is nonincreasing) such that $g(X)$, $h(X, Y)$ and $g(X)h(X, Y)$ have finite expectations. Then

$$\mathrm{Cov}(g(X), h(X, Y)) \leqslant 0 \ . \tag{4.36}$$

If moreover g is increasing, and X and $h(X, Y)$ are not almost surely constant, then the above inequality is strict.

A consequence of the Harris-FKG inequality is the following upper bound for a weighted sum of ranked random variables with nonincreasing weights which is also proven in Section C.

Corollary 4.2. Consider some weights $\mathbf{w} = (w_1, \ldots, w_\mu, \ldots, w_\lambda)$ satisfying $w_1 \geqslant w_2 \geqslant \ldots \geqslant w_\mu > w_{\mu+1} = \ldots = w_\lambda = 0$ and summing to 1 (i.e., satisfying the assumption **W**1). Let $\xi^1, \ldots, \xi^\lambda$ be i.i.d. real-valued random variables, with probability distribution ν_U^1, and define $s \in \mathfrak{S}_\lambda$ a permutation such that $\xi^{s(1)} \leqslant \cdots \leqslant \xi^{s(\lambda)}$, i.e., $\{\xi^{s(i)}, i = 1, \ldots, \lambda\}$ are order statistics of random variables following ν_U^1. Then, almost surely $\sum_{i=1}^\mu w_i \xi^{s(i)} \leqslant \frac{1}{\mu} \sum_{i=1}^\mu \xi^{s(i)}$ and thus in expectation $\mathbb{E}\left[\sum_{i=1}^\mu w_i \xi^{s(i)}\right] \leqslant \mathbb{E}\left[\sum_{i=1}^\mu \frac{1}{\mu} \xi^{s(i)}\right]$. In addition, when $\mu \leqslant \lambda/2$, if ν_U^1 is integrable and symmetric with respect to 0, we have

$$\mathbb{E}\left[\sum_{i=1}^{\mu} w_i \xi^{s(i)}\right] \leqslant \mathbb{E}\left[\sum_{i=1}^{\mu} \frac{1}{\mu} \xi^{s(i)}\right] \leqslant \mathbb{E}\left[\min\{\xi^1, \xi^2\}\right]. \tag{4.37}$$

The first inequality of (4.37) states in particular that given some decreasing weights satisfying the assumption, the coefficient $\mathbb{E}\left[\sum_{i=1}^\mu w_i \xi^{s(i)}\right]$ is smaller than or equal to the coefficient using equal weights.

We remind now some inequalities on symmetric matrices that will be useful for estimates on the covariance matrix. Given a symmetric matrix $\mathbf{A}$, we denote $\mathbf{A} \succeq 0$ if $\mathbf{A}$ is positive semidefinite, i.e., $\mathbf{A} \in \mathcal{S}_+^d$. Naturally, given two matrices (not necessarily symmetric), we denote $\mathbf{A} \succeq \mathbf{B}$ if $\mathbf{A} - \mathbf{B} \succeq 0$ (equivalently if $\mathbf{A} - \mathbf{B}$ is (symmetric) positive semidefinite). Therefore, if $\mathbf{A} \succeq \mathbf{B}$, then for every matrix $\mathbf{C}$ we have

$$\mathbf{C} + \mathbf{A} \succeq \mathbf{C} + \mathbf{B} \ . \tag{4.38}$$

Here are some consequences of Weyl's inequality [79, Theorem 4.3.1]. First, if $\mathbf{A}$ and $\mathbf{B}$ are symmetric and $\mathbf{B}$ is positive semidefinite then for $k = 1, \ldots, d$ [79, Corollary 4.3.12]

$$\lambda_k(\mathbf{A}) \leqslant \lambda_k(\mathbf{A} + \mathbf{B}) \tag{4.39}$$

where $\lambda_k(\cdot)$ is the function that maps a symmetric matrix to its k-th largest eigenvalue counted with multiplicity. Consequently, if $\mathbf{A}$ and $\mathbf{B}$ are symmetric with $\mathbf{A} \succeq \mathbf{B}$ then

$$\lambda_k(\mathbf{A}) \geqslant \lambda_k(\mathbf{B}) \ . \tag{4.40}$$

Indeed we can write $\mathbf{A} = \mathbf{B} + \mathbf{A} - \mathbf{B}$ with $\mathbf{A} - \mathbf{B} \succeq 0$ and thus by (4.39) $\lambda_k(\mathbf{A}) = \lambda_k(\mathbf{B} + \mathbf{A} - \mathbf{B}) \geqslant \lambda_k(\mathbf{B})$. Another useful corollary [79, Corollary 4.3.15] of Weyl's inequality reads that if $\mathbf{A}$ and $\mathbf{B}$ are symmetric matrices then

$$\lambda_k(\mathbf{A}) + \lambda_d(\mathbf{B}) \leqslant \lambda_k(\mathbf{A} + \mathbf{B}) \leqslant \lambda_k(\mathbf{A}) + \lambda_1(\mathbf{B}) \ . \tag{4.41}$$

We also remind that given a symmetric matrix $\mathbf{A}$

$$\lambda_d(\mathbf{A}) \leqslant x^\top \mathbf{A} x \leqslant \lambda_1(\mathbf{A}) \text{ for any unit vector } x. \tag{4.42}$$

and moreover the maximum and minimal eigenvalue of $\mathbf{A}$ are given by

$$\lambda_1(\mathbf{A}) = \max_{x \neq 0} \frac{x^\top \mathbf{A} x}{x^\top x} \text{ and } \lambda_d(\mathbf{A}) = \min_{x \neq 0} \frac{x^\top \mathbf{A} x}{x^\top x} \tag{4.43}$$

(see e.g., [79, Theorem 4.2.2]). Remark also that if $\mathbf{A} \succeq \mathbf{B}$ with $\mathbf{A}$ and $\mathbf{B}$ symmetric, then for any $z \in \mathbb{R}^d$, $z^\top \mathbf{A}^{1/2} \mathbf{A}^{1/2} z \geqslant z^\top \mathbf{B}^{1/2} \mathbf{B}^{1/2} z$ and thus equivalently

$$\|\mathbf{A}^{1/2} z\| \geqslant \|\mathbf{B}^{1/2} z\|. \tag{4.44}$$

We additionally derive the following lemma which provides a useful matrix inequality to apply to the rank-one update.

Lemma 4.1. Let $u, v \in \mathbb{R}^d$. Then $2(uu^\top + vv^\top) \succeq (u+v)(u+v)^\top$.

Proof. We have $2(uu^\top + vv^\top) - (u+v)(u+v)^\top = uu^\top + vv^\top - uv^\top - vu^\top = u[u-v]^\top + v[v-u]^\top = [u-v][u-v]^\top$. Since $[u-v][u-v]^\top$ is positive semidefinite, $2(uu^\top + vv^\top) - (u+v)(u+v)^\top \succeq 0$. □

3.3.2 *On selection functions*

In the CMA-ES algorithm, the random input at each iteration U_{t+1} is used to sample candidate solutions $x_{t+1}^i = m_t + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^i$. When f is scaling-invariant with respect to x^*, ranking the candidate solutions x_{t+1}^i by their f-values is equivalent to ranking $x^* + z_t + \sqrt{\mathbf{\Sigma}_t} U_{t+1}^i$ for $i = 1, \ldots, \lambda$ [53, Lemma 2.2]. The permutation containing the order of the candidate solutions is then used to update the state of the algorithm. This motivates to define the selection function at iteration t as

$$F_{t+1}(u) := f\left(x^* + z_t + \sqrt{\mathbf{\Sigma}_t} u\right) \qquad \text{for } u \in \mathbb{R}^d \tag{4.45}$$

such that the permutation s_{t+1} that defines the order of the candidate solutions satisfies

$$F_{t+1}\left(U_{t+1}^{s_{t+1}(1)}\right) \leqslant F_{t+1}\left(U_{t+1}^{s_{t+1}(2)}\right) \leqslant \cdots \leqslant F_{t+1}\left(U_{t+1}^{s_{t+1}(\lambda)}\right) \ . \tag{4.46}$$

We also consider selection functions independently of the state of the algorithm. We introduce in particular the set of admissible selection functions defined as

$$\mathcal{F} = \{F \colon \mathbb{R}^d \to \mathbb{R} \text{ measurable with Lebesgue-negligible level sets}\}. \tag{4.47}$$

Given $U = (U^1, \ldots, U^\lambda)$ such that the random variables $U^1, \ldots, U^\lambda$ are i.i.d., and follow the probability distribution ν_U^d, the selection function F defines $s_{F;U}$ as the (random and almost sure unique) permutation of $\mathfrak{S}_\lambda$ such that[4]

$$F(U^{s_{F;U}(1)}) \leqslant \cdots \leqslant F(U^{s_{F;U}(\lambda)}). \tag{4.48}$$

Since the distribution of the random vector $(U_{t+1}^{s_{t+1}(1)}, \ldots, U_{t+1}^{s_{t+1}(\lambda)})$ is the only information about f which is put into the algorithm, we will need to understand the behavior of its distribution depending on different values of θ_t. For this, it will often be useful to work with an equivalent, typically simpler, selection function where naturally, a selection function G is equivalent to the selection function F if the permutation extracted via F or G (which is the information used to update the algorithm) is the same. We show in the next lemma that it is equivalent to having F and G that differ up to a strictly increasing transformation. The proof of Lemma 4.2 can be found in C.3.

Lemma 4.2. Let $U^1, \ldots, U^\lambda$ be i.i.d. random vectors following the distribution ν_U^d. Given F or G two selection functions, we define $s_{F;U}$ (respectively $s_{G;U}$) the random permutation of $\{1, \ldots, \lambda\}$ that sorts the $F(U^i)$ (respectively $G(U^i)$) for $i = 1, \ldots, \lambda$. Then, almost surely $s_{F;U} = s_{G;U}$ if and only if there exists a (strictly) increasing transformation $g \colon F(\mathbb{R}^d) \to G(\mathbb{R}^d)$ such that $G(x) = g \circ F(x)$ for ν_U^d-almost every $x \in \mathbb{R}^d$.

The previous lemma motivates to introduce the following equivalence relation on $\mathcal{F}$.

[4] In case of equality in the ordering of the F-values, i.e., there exist $i < j$ with $F(U^{s_{F;U}(i)}) = F(U^{s_{F;U}(j)})$, we impose $s_{F;U}(i) < s_{F;U}(j)$ as a tie-break

Definition 4.1. Consider two admissible selection functions F and G, then F is equivalent to G, denoted $F \overset{\text{sel}}{\sim} G$, when there exists a (strictly) increasing transformation $g\colon \mathbb{R} \to \mathbb{R}$ such that $G = g \circ F$.

We use a slight abuse of notation and denote $F(u) \overset{\text{sel}}{\sim} G(u)$ for $u \in \mathbb{R}^d$, which simplifies the definitions of some equivalent functions. Moreover, for $(z, \boldsymbol{\Sigma}) \in \mathbb{R}^d \times \mathcal{S}_{++}^d$, we define the selection function associated to spherical functions as

$$G_{z,\boldsymbol{\Sigma}}(u) := \|z + \sqrt{\boldsymbol{\Sigma}}u\|^2 \ . \tag{4.49}$$

We show in the next proposition that when minimizing a spherical function, the selection function F_{t+1} is equivalent to $G_{z_t,\boldsymbol{\Sigma}_t}(\cdot)$.

Proposition 4.5. Consider $(z, \boldsymbol{\Sigma}) \in \mathbb{R}^d \times \mathcal{S}_{++}^d$. We have the following equivalence between selection functions:

$$G_{z,\boldsymbol{\Sigma}}(u) \overset{\text{sel}}{\sim} \sum_{k=1}^{d} 2\sqrt{\lambda_k(\boldsymbol{\Sigma})}\langle z, e_k(\boldsymbol{\Sigma})\rangle\langle u, e_k(\boldsymbol{\Sigma})\rangle + \lambda_k(\boldsymbol{\Sigma})\langle u, e_k(\boldsymbol{\Sigma})\rangle^2 \ . \tag{4.50}$$

In particular, consider the Markov chain $\Theta = \{(z_t, p_t, q_t, \boldsymbol{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) with $\theta_0 \in \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times (0, +\infty)$. Let $t \in \mathbb{N}$ and $\mathcal{B}_t = (e_1(\boldsymbol{\Sigma}_t), \ldots, e_d(\boldsymbol{\Sigma}_t))$ be an orthonormal basis of $\mathbb{R}^d$ composed of eigenvectors of $\boldsymbol{\Sigma}_t$, such that for $k = 1, \ldots, d$, $e_k(\boldsymbol{\Sigma}_t)$ is an eigenvector of $\boldsymbol{\Sigma}_t$ associated to its k-th largest (counted with multiplicity) eigenvalue $\lambda_k(\boldsymbol{\Sigma}_t)$. Assume that the objective function f is spherical, i.e., satisfies **F**2. Then

$$F_{t+1}(u) \overset{\text{sel}}{\sim} \sum_{k=1}^{d} 2\sqrt{\lambda_k(\boldsymbol{\Sigma}_t)}\langle z_t, e_k(\boldsymbol{\Sigma}_t)\rangle\langle u, e_k(\boldsymbol{\Sigma}_t)\rangle + \lambda_k(\boldsymbol{\Sigma}_t)\langle u, e_k(\boldsymbol{\Sigma}_t)\rangle^2 \ . \tag{4.51}$$

Proof. Indeed, we have $G_{z,\boldsymbol{\Sigma}}(u) = \|z + \sqrt{\boldsymbol{\Sigma}}u\|^2 = \|z_t\|^2 + 2\langle z, \sqrt{\boldsymbol{\Sigma}}u\rangle + \|\sqrt{\boldsymbol{\Sigma}}u\|^2 = \|z\|^2 + 2\sum_{k=1}^d \langle e_k(\boldsymbol{\Sigma}), z\rangle\sqrt{\lambda_k(\boldsymbol{\Sigma})}\langle e_k(\boldsymbol{\Sigma}), u\rangle + \sum_{k=1}^d \lambda_k(\boldsymbol{\Sigma})\langle e_k(\boldsymbol{\Sigma}), u\rangle^2$. Consider the increasing function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(v) = v - \|z\|^2$, applied to the expression of $G_{z,\boldsymbol{\Sigma}}(u)$, we find $G_{z,\boldsymbol{\Sigma}}(u) \overset{\text{sel}}{\sim} \sum_{k=1}^d 2\sqrt{\lambda_k(\boldsymbol{\Sigma})}\langle z, e_k(\boldsymbol{\Sigma})\rangle\langle u, e_k(\boldsymbol{\Sigma})\rangle + \lambda_k(\boldsymbol{\Sigma})\langle u, e_k(\boldsymbol{\Sigma})\rangle^2$. Moreover, (4.51) follows from that, by **F**2 and (4.45), we have $F_{t+1}(u) \overset{\text{sel}}{\sim} \|z_t + \sqrt{\boldsymbol{\Sigma}_t}u\|^2 = G_{z_t,\boldsymbol{\Sigma}_t}(u)$. □

We now derive an equivalent expression to the selection function in the case of an ellipsoidal objective function, i.e., when f satisfies **F**3. We derive it using the property that $x^\top \mathbf{H} x = \|\mathbf{H}^{1/2}x\|^2$ as a consequence of the previous proposition giving an equivalent selection function for spherical functions. Therefore we state it as a corollary of Proposition 4.5.

Corollary 4.3. Consider an ellipsoidal objective function f satisfying **F**3 with quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}_{++}^d$, and the normalized Markov chain $\Theta = \{(z_t, p_t, q_t, \boldsymbol{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) with $\theta_0 \in \mathbb{R}^{3d} \times \mathcal{S}_{++}^d \times (0, +\infty)$. Let $t \in \mathbb{N}$ and $\mathcal{B}_t = (e_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}), \ldots, e_d(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}))$ be an orthonormal basis of $\mathbb{R}^d$ composed of eigenvectors of $\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}$ such that for $k = 1, \ldots, d$, $e_k(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2})$ is an eigenvector of $\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}$ associated to its k-th largest (counted with multiplicity) eigenvalue $\lambda_k(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2})$. Then, two equivalent expressions of

the selection function F_{t+1} are given by

$$F_{t+1}(u) \overset{\text{sel}}{\sim} \sum_{k=1}^{d} 2\sqrt{\lambda_k\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}\right)}\langle \mathbf{H}^{1/2}z_t\rangle_k\langle \mathbf{R}_t^{\mathbf{H}}u\rangle_k + \lambda_k\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}\right)\langle \mathbf{R}_t^{\mathbf{H}}u\rangle_k^2 \quad (4.52)$$

$$\overset{\text{sel}}{\sim} \left\|\mathbf{H}^{1/2}z_t + \left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}\right)^{1/2}\mathbf{R}_t^{\mathbf{H}}u\right\|^2 \quad (4.53)$$

where $\langle\cdot\rangle_k = \left\langle \cdot, e_k\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}\right)\right\rangle$ for $k = 1, \dots, d$ and

$$\mathbf{R}_t^{\mathbf{H}} = \left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}\right)^{-1/2}\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t^{1/2} \quad (4.54)$$

is an orthogonal matrix of $\mathbb{R}^d$.

Proof. Since the objective function satisfies **F**3 and by definition of the selection function (4.45), we have

$$\begin{aligned} F_{t+1}(u) &\overset{\text{sel}}{\sim} \left(z_t + \sqrt{\boldsymbol{\Sigma}_t}u\right)^\top \mathbf{H}\left(z_t + \sqrt{\boldsymbol{\Sigma}_t}u\right) = \left(\mathbf{H}^{1/2}z_t + \mathbf{H}^{1/2}\sqrt{\boldsymbol{\Sigma}_t}u\right)^\top \left(\mathbf{H}^{1/2}z_t + \mathbf{H}^{1/2}\sqrt{\boldsymbol{\Sigma}_t}u\right) \\ &= \left\|\mathbf{H}^{1/2}z_t + \mathbf{H}^{1/2}\sqrt{\boldsymbol{\Sigma}_t}u\right\|^2 = \left\|\mathbf{H}^{1/2}z_t + \left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}\right)^{1/2}\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}\right)^{-1/2}\mathbf{H}^{1/2}\sqrt{\boldsymbol{\Sigma}_t}u\right\|^2 \\ &= \left\|\mathbf{H}^{1/2}z_t + \left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}\right)^{1/2}\mathbf{R}_t^{\mathbf{H}}u\right\|^2 . \end{aligned}$$

We thus obtain that $F_{t+1}(u) \overset{\text{sel}}{\sim} G_{\mathbf{H}^{1/2}z_t, \mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}}(R_t^{\mathbf{H}}u)$ so that (4.52) follows from (4.50). Besides, observe that $\mathbf{R}_t^{\mathbf{H}}\left(\mathbf{R}_t^{\mathbf{H}}\right)^\top = \left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}\right)^{-1/2}\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\Sigma}_t^{1/2}\mathbf{H}^{1/2}\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2}\right)^{-1/2} = \mathbf{I}_d$ and thus $\mathbf{R}_t^{\mathbf{H}}$ is an orthogonal rotation of $\mathbb{R}^d$. □

The following proposition connects the normalized Markov chain (4.17) when the objective function is spherical to when it is ellipsoidal. More precisely, it provides the right change of state variables (except for the normalized path p_t) to go from optimizing a spherical function to optimizing an ellipsoidal function between two consecutive iterations. It is instrumental in many proofs to extend the results from the spherical to any ellipsoidal function. Its proof relies on the equivalences between selection functions given in Proposition 4.5 and Corollary 4.3 previously stated and is delayed to C.4.

Proposition 4.6 (Change of variables under affine transformation)**.** Consider the Markov chain $\{\theta_t\}_{t\in\mathbb{N}} = \{(z_t, p_t, q_t, \boldsymbol{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) when minimizing a spherical objective function f (satisfying **F**2) and normalized by $R(\cdot) = \lambda_{\min}(\cdot)$ (i.e., satisfying **R**2). Moreover, let $\{\hat\theta_t\}_{t\in\mathbb{N}} = \{(\hat z_t, \hat p_t, \hat q_t, \hat{\boldsymbol{\Sigma}}_t, \hat r_t)\}_{t\in\mathbb{N}}$ be the Markov chain obeying (4.17) with an ellipsoidal objective function $\hat f$ satisfying **F**3 with quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}_{++}^d$ and normalized by $\hat R(\cdot) = \lambda_{\min}(\mathbf{H}^{1/2}\cdot\mathbf{H}^{1/2})$ (i.e., satisfying **R**3). Suppose that the sampling distribution ν_U^d and that the stepsize change Γ_{d_σ} are invariant to rotation (i.e., they satisfy **N**2 and **Γ**4, respectively). If we couple the initial states θ_0 and $\hat\theta_0$ via

$$\hat z_0 = \mathbf{H}^{-1/2}z_0\ ,\ \hat p_0 = \left(\mathbf{R}_0^{\mathbf{H}}\right)^{-1}p_0\ ,\ \hat q_0 = \mathbf{H}^{-1/2}q_0\ ,\ \hat{\boldsymbol{\Sigma}}_0 = \mathbf{H}^{-1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{-1/2}\ ,\ \hat r_0 = r_0\ , \quad (4.55)$$

where $\mathbf{R}_0^{\mathbf{H}}$ is the orthogonal matrix $\mathbf{R}_0^{\mathbf{H}} = (\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})^{-1/2}\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0^{1/2}$ (see (4.54)). Then, at the next iteration, the following equalities in distribution hold:

$$\left(\mathbf{H}^{1/2}\hat{z}_1, \|\hat{p}_1\|, \mathbf{H}^{1/2}\hat{q}_1, \mathbf{H}^{1/2}\hat{\boldsymbol{\Sigma}}_1\mathbf{H}^{1/2}, \hat{r}_1, \mathbf{R}_0^{\mathbf{H}}U_1\right) \overset{\text{dist.}}{=} (z_1, \|p_1\|, q_1, \boldsymbol{\Sigma}_1, r_1, U_1) \quad , \tag{4.56}$$

and

$$U_1^{s_1(i)} \overset{\text{dist.}}{=} U_1^{\hat{s}_1(i)} \quad \text{for } i = 1, \dots, \lambda, \tag{4.57}$$

where $U_1 = (U_1^1, \dots, U_1^\lambda) \sim (\nu_U^d)^{\otimes\lambda}$ and the permutations s_1 and $\hat{s}_1$ sort the $f(z_0 + \boldsymbol{\Sigma}_0^{1/2}U_1^i)$ and $\hat{f}(\hat{z}_0 + \hat{\boldsymbol{\Sigma}}_0^{1/2}U_1^i)$, respectively, for $i = 1, \dots, \lambda$.

This proposition is tightly connected to affine-invariance. More precisely if we would be able to obtain a change of variable as in (4.56) with p_1, $\hat{p}_1$ instead of $\|p_1\|$, $\|\hat{p}_1\|$, then the corollary would imply affine-invariance of the (normalized chain associated to the) CMA-ES algorithm. For other variants of CMA-ES (that are using another or a modified stepsize update), such a change of variable involving all state variables exists and the algorithms are affine-invariant [70, 15].

The following technical proposition is useful in the sequel. It will imply in particular a continuity[5] property of the permutation s_{t+1} with respect to the selection function F_{t+1}. We exploit this later to deduce continuity of the rank-mu update matrix with respect to the current state of the process (4.17), see Proposition 4.13, and to uniformly bound the expected length of one step from below, see Lemma 4.4. The proof of Proposition 4.7 is delayed to Section C.5.

Proposition 4.7. Consider a random vector $U = (U^1, \dots, U^\lambda)$, such that $U^1, \dots, U^\lambda$ are i.i.d. and follow the sampling distribution ν_U^d satisfying **N**1. For every measurable F in $\mathcal{F}$ (see (4.47)), let $s_{F;U} \in \mathfrak{S}_\lambda$ be the almost surely unique permutation which sorts the $F(U^i)$ for $i = 1, \dots, \lambda$, i.e., satisfies (4.48).
Let $\{F_x\}_{x\in\mathsf{X}}$ be a family of functions in $\mathcal{F}$ that depend continuously on a parameter x in a metric space X, i.e., F_x converges pointwise to $F_{\bar{x}}$ when x tends to $\bar{x} \in \mathsf{X}$. Then, for any $\phi \in \mathsf{L}^1(\nu_U^{d\lambda})$, the map $x \in \mathsf{X} \mapsto \mathbb{E}[\phi((U^{s_{F_x;U}(i)})_{i=1,\dots,\lambda})]$ is continuous.

3.3.3 *Updated covariance matrix*

In this section, we provide preliminary results about the update of the covariance matrix in the Markov chain (4.17). We find lower and upper bounds for the eigenvalues of the updated covariance matrix (Proposition 4.8 and Corollary 4.4) and an upper bound on the projection of its eigenvectors on the coordinate system of the initial covariance matrix (Proposition 4.9). Given a basis $\mathcal{B} = (e_1, \dots, e_d)$ of $\mathbb{R}^d$ and a vector $v \in \mathbb{R}^d$, we denote $[v]_\mathcal{B}$ the vector v written in the basis $\mathcal{B}$:

$$[v]_\mathcal{B} = [\langle v, e_1\rangle, \dots, \langle v, e_d\rangle]^\top \quad . \tag{4.58}$$

Likewise, for a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, $[\mathbf{A}]_\mathcal{B}$ denotes the matrix rewritten in the basis $\mathcal{B}$, i.e. $[\mathbf{A}]_\mathcal{B} = \mathbf{P}_\mathcal{B}^{-1}\mathbf{A}\mathbf{P}_\mathcal{B}$ where $\mathbf{P}_\mathcal{B} = [e_1|\dots|e_d]$ is the change-of-base matrix to $\mathcal{B}$. We consider $\tilde{\boldsymbol{\Sigma}}_{t+1}$ the update of the normalized covariance matrix (but before renormalization), with rank-mu and rank-one updates defined in (4.18). Let $\lambda_1(\boldsymbol{\Sigma}_t), \dots, \lambda_d(\boldsymbol{\Sigma}_t)$ be the decreasingly ordered eigenvalues of $\boldsymbol{\Sigma}_t$, counted with multiplicity, i.e.,

$$\lambda_1(\boldsymbol{\Sigma}_t) \geqslant \dots \geqslant \lambda_d(\boldsymbol{\Sigma}_t) \tag{4.59}$$

[5] with respect to the discrete topology $\mathcal{P}(\mathfrak{S}_\lambda)$ of the set of permutations $\mathfrak{S}_\lambda$

and let $e_1(\mathbf{\Sigma}_t), \dots, e_d(\mathbf{\Sigma}_t)$ be eigenvectors of $\mathbf{\Sigma}_t$ associated to the eigenvalues $\lambda_1(\mathbf{\Sigma}_t), \dots, \lambda_d(\mathbf{\Sigma}_t)$ respectively, such that $\mathcal{B}_t = (e_1(\mathbf{\Sigma}_t), \dots, e_d(\mathbf{\Sigma}_t))$ results in an orthonormal basis of $\mathbb{R}^d$. The matrix $\tilde{\mathbf{\Sigma}}_{t+1}$ expressed in the basis $\mathcal{B}_t$ reads

$$\left[\tilde{\mathbf{\Sigma}}_{t+1}\right]_{\mathcal{B}_t} = (1 - c_1 - c_\mu)\Lambda(\mathbf{\Sigma}_t) + c_1[q_{t+1}]_{\mathcal{B}_t}[q_{t+1}]_{\mathcal{B}_t}^\top + c_\mu\sqrt{\Lambda(\mathbf{\Sigma}_t)}\tilde{\mathrm{M}}_{t+1}^\mu\sqrt{\Lambda(\mathbf{\Sigma}_t)} \tag{4.60}$$

where $\tilde{\mathrm{M}}_{t+1}^\mu$ is the so-called rank-mu update matrix defined by

$$\tilde{\mathrm{M}}_{t+1}^\mu = \sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right]_{\mathcal{B}_t}\left[U_{t+1}^{s_{t+1}(i)}\right]_{\mathcal{B}_t}^\top \quad , \tag{4.61}$$

and $\Lambda(\mathbf{\Sigma}_t) = \mathrm{diag}(\lambda_1(\mathbf{\Sigma}_t), \dots, \lambda_d(\mathbf{\Sigma}_t))$ is the diagonal matrix composed of the eigenvalues of $\mathbf{\Sigma}_t$, counted with multiplicity and decreasingly ordered.

We define then the rank-mu update matrix associated to a selection function $F \in \mathcal{F}$ (see (4.47)), acting on i.i.d. random variables $U^1, \dots, U^\lambda$ following ν_U^d, as the matrix

$$\tilde{\mathrm{M}}^\mu(F; U) \coloneqq \sum_{i=1}^{\mu} w_i^c [U^{s_{F;U}(i)}][U^{s_{F;U}(i)}]^\top \tag{4.62}$$

where the permutation $s_{F;U} \in \mathfrak{S}_\lambda$ is defined via (4.48). When the selection function defined in (4.51) satisfies $F_{t+1}([\cdot]_{\mathcal{B}_t}) \overset{\text{sel}}{\sim} F$, we have $\tilde{\mathrm{M}}_{t+1}^\mu = \tilde{\mathrm{M}}^\mu(F, [U_{t+1}]_{\mathcal{B}_t})$. Note that the dependency of $\tilde{\mathrm{M}}^\mu(F; U)$ on F is fully determined by its dependency on $s_{F;U}$.

We discuss now how the eigenvalues and eigenvectors of the matrix $\mathbf{\Sigma}_1$ depend on the initial covariance matrix $\mathbf{\Sigma}_0$. We derive in the next proposition some upper and lower bounds on the eigenvalues of the updated covariance matrix $\mathbf{\Sigma}_1$. The second inequality proven relies on a consequence of the min-max principle [51, Eq. (5)] that reads: given $\mathbf{A} \in \mathcal{S}_{++}^d$ and $v_1, \dots, v_\mu \in \mathbb{R}^d$ the following inequality holds

$$\lambda_k\left(\mathbf{A} + \sqrt{\mathbf{A}}\sum_{i=1}^{\mu} v_i v_i^\top \sqrt{\mathbf{A}}\right) \leqslant \lambda_k(\mathbf{A}) \times \left(1 + \mu d \max_{i=1,\dots,\mu} \|v_i\|_\infty^2\right) \quad . \tag{4.63}$$

The proof of Proposition 4.8 is presented in C.6.

Proposition 4.8. Consider the Markov chain $\Theta = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) with initial state $(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \in \mathbb{R}^{3d} \times \mathcal{S}_{++}^d \times (0, +\infty)$ where the permutation is given by (4.19) for a measurable objective function f. For $k = 1, \dots, d$, the k^{th} eigenvalue (counted with its multiplicity) of $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ satisfies the following inequalities

$$(1 - c_1 - c_\mu)\lambda_k(\mathbf{\Sigma}_0) \leqslant r_1\lambda_k(\mathbf{\Sigma}_1) \tag{4.64}$$

$$r_1\lambda_k(\mathbf{\Sigma}_1) \leqslant \left(1 - c_1 - c_\mu + (2c_1\mu_{\text{eff}} + c_\mu)d\mu\|U_1\|_\infty^2\right)\lambda_k(\mathbf{\Sigma}_0) + 2c_1 r_0^{-1}(1 - c_c)^2\|q_0\|^2 \tag{4.65}$$

where $\|U_1\|_\infty = \max_{i=1,\dots,\lambda}\|U_1^i\|_\infty$. In particular, when the normalization R satisfies **R**2, we have $\lambda_d(\mathbf{\Sigma}_1) = \lambda_d(\mathbf{\Sigma}_0) = 1$ and thus $r_1 \geqslant 1 - c_1 - c_\mu$.

The bounds on the eigenvalue of the normalized covariance matrix derived in the previous proposition hold for any f. When the objective function is ellipsoidal satisfying **F**3 (giving thus the

same selection than a convex quadratic function with quasi-Hessian matrix $\mathbf{H}$), we need (for instance in Lemmas 4.7 and 4.8) bounds on $\lambda_k\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right)$. We derive them in the next corollary as a consequence of the previous proposition (applied when the objective function is spherical) and of the change of variable under affine transformations given in Proposition 4.6.

Corollary 4.4. Suppose that the objective function f is an increasing transformation of $(x-x^*)^\top\mathbf{H}(x-x^*)$ with $\lambda_{\min}(\mathbf{H})=1$, i.e., satisfies **F**3, that the normalization function $R(\cdot)=\lambda_{\min}(\mathbf{H}^{1/2}\cdot\mathbf{H}^{1/2})$ (satisfies **R**3), that the sampling distribution ν_U^d is invariant to rotation, i.e., **N**2 holds, and that the stepsize change is invariant to rotation, i.e., Γ_{d_σ} satisfies **Γ**4. Consider the Markov chain $\Theta=\{(z_t,p_t,q_t,\mathbf{\Sigma}_t,r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) with initial state $(z_0,p_0,q_0,\mathbf{\Sigma}_0,r_0)\in\mathbb{R}^{3d}\times\mathcal{S}_{++}^d\times(0,+\infty)$. For $k=1,\dots,d$, the k^{th} eigenvalue (counted with its multiplicity) of $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ satisfies the following inequalities almost surely

$$(1-c_1-c_\mu)\lambda_k\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)\leqslant r_1\lambda_k\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right) \tag{4.66}$$

$$r_1\lambda_k\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right)\leqslant\left(1-c_1-c_\mu+(2c_1\mu_{\text{eff}}+c_\mu)d\mu\|U_1\|^2\right)\lambda_k\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)+2c_1r_0^{-1}(1-c_c)^2\|\mathbf{H}^{1/2}q_0\|^2 \tag{4.67}$$

where $\|U_1\|=\max_{i=1,\dots,\lambda}\|U_1^i\|$. In particular, since $\lambda_d\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)=\lambda_d\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right)=1$, we have $r_1\geqslant 1-c_1-c_\mu$.

Proof. First, if $\mathbf{H}=\mathbf{I}_d$, we apply Proposition 4.8 which gives the desired result. For the general case, we use Proposition 4.6. Consider $\{\theta_t\}_{t\in\mathbb{N}}=\{(z_t,p_t,q_t,\mathbf{\Sigma}_t,r_t)\}_{t\in\mathbb{N}}$ associated to f which is an increasing transformation of $x\mapsto(x-x^*)^\top\mathbf{H}(x-x^*)$. Likewise consider $\{\hat\theta_t\}_{t\in\mathbb{N}}=\{(\hat z_t,\hat p_t,\hat q_t,\hat{\mathbf{\Sigma}}_t,\hat r_t)\}_{t\in\mathbb{N}}$ the Markov chain defined by (4.17) with a spherical objective function $\hat f$ (i.e., satisfying **F**2) and normalization function $\hat R$ satisfying **R**2 such that the initial states of $\{\theta_t\}_{t\in\mathbb{N}}$ and $\{\hat\theta_t\}_{t\in\mathbb{N}}$ satisfy

$$(\hat z_0,\hat p_0,\hat q_0,\hat{\mathbf{\Sigma}}_0,\hat r_0)=(\mathbf{H}^{1/2}z_0,\mathbf{R}_0^{\mathbf{H}}p_0,\mathbf{H}^{1/2}q_0,\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2},r_0)$$

where $\mathbf{R}_0^{\mathbf{H}}$ is the orthogonal matrix given in Proposition 4.6. By the equality (4.56) in distribution in Proposition 4.6, we obtain

$$\mathbb{P}\left[(1-c_1-c_\mu)\lambda_k\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)\leqslant r_1\lambda_k\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right)\right]=\mathbb{P}\left[(1-c_1-c_\mu)\lambda_k(\hat{\mathbf{\Sigma}}_0)\leqslant\hat r_1\lambda_k(\hat{\mathbf{\Sigma}}_1)\right]=1$$

and thus (4.66) holds almost surely. Likewise, (4.67) holds almost surely. We leave the proof of (4.67) to the reader (note that we use that $\|u\|_\infty\leqslant\|u\|$ and that for an orthogonal matrix $\mathbf{R}$, $\|\mathbf{R}u\|=\|u\|$). □

The next proposition provides upper bounds on the projection of the eigenvectors of the updated covariance matrix $\mathbf{\Sigma}_1$ on the eigenvectors of the initial covariance matrix $\mathbf{\Sigma}_0$. It is applied later in Section 3.4 to prove Proposition 4.15. It relies on the following upper bound for matrices that write as a sum of a matrix $\mathbf{A}$ and a sum of rank-one matrices $\sqrt{\mathbf{A}}\sum_{i=1}^{\mu}v_iv_i^\top\sqrt{\mathbf{A}}$ [51, Theorem 1]. More

precisely, given $\mathbf{A} \in \mathcal{S}^d_{++}$ and $v_1, \dots, v_\mu \in \mathbb{R}^d$, we will use the following bound for $j, k = 1, \dots, d$:

$$\left| \left\langle e_j(\mathbf{A}), e_k\left(\mathbf{A} + \sqrt{\mathbf{A}} \sum_{i=1}^{\mu} v_i v_i^\top \sqrt{\mathbf{A}}\right) \right\rangle \right| \leqslant P_{d,\mu}\left(\max_{i=1,\dots,\mu} \|v_i\|\right) \times \sqrt{\frac{\min\{\lambda_j(\mathbf{A}), \lambda_j(\mathbf{A})\}}{\max\{\lambda_j(\mathbf{A}), \lambda_j(\mathbf{A})\}}} \quad (4.68)$$

where $P_{d,\mu}(\cdot)$ is a polynomial function and $e_j(\mathbf{B})$ denotes a normalized eigenvector of a symmetric matrix $\mathbf{B}$ associated to its j^{th} largest eigenvalue (counted with multiplicity). Additionally, denoting ν_i the eigenvalues of $\mathbf{A} + \sqrt{\mathbf{A}} \sum_{i=1}^{\mu} u_i u_i^\top \sqrt{\mathbf{A}}$, and λ_i the eigenvalues of $\mathbf{A}$, the proof also uses the inequality [85, Theorem 2.7], [135, Theorem 2.1, p. 175]

$$\nu_i \leqslant \lambda_i (1 + \mu d \times \max_{k=1,\dots,\mu} \|u_i\|^2_\infty) \ . \quad (4.69)$$

The proof of Proposition 4.9 is given in C.7.

Proposition 4.9. Consider an objective function f satisfying **F**1, a normalization function R satisfying **R**1 and the associated Markov chain $\Theta = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t \in \mathbb{N}}$ obeying (4.17). Suppose moreover that the distribution ν^d_U satisfies **N**4 and let $\bar{q} > 0$. Then, there exists a positive random variable ρ_1 with finite moments, such that, if $c_\mu \in [0, 1/4]$ and $c_1 \in [0, \max(\mu_{\text{eff}}^{-1} c_\mu, 1 - c_\mu)]$, then for any initial condition $(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \in \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times (0, +\infty)$ where $\|q_0\|_\infty \leqslant \bar{q}$, and for $i, j = 1, \dots, d$, when $\frac{\min\{\lambda_i, \lambda_j\}}{\max\{\lambda_i, \lambda_j\}} \leqslant 1 - \sqrt{c_\mu}$, we have that

$$|\langle e_i(\mathbf{\Sigma}_1), e_j(\mathbf{\Sigma}_0) \rangle| \leqslant \sqrt{c_\mu} \rho_1 \times \sqrt{\frac{\min\{\lambda_i, \lambda_j\}}{\max\{\lambda_i, \lambda_j\}}} \ . \quad (4.70)$$

3.3.4 *Sufficient condition on the initial state and algorithm parameters for the expected stepsize change to be smaller than one after one or two iterations*

In assumption **Γ**3, we formulated the condition that for large enough δ_p, $\delta_\sigma \geqslant 1$, if the norm of a path is larger than $(1 + \delta_p)\mathbb{E}\|\nu^d_U\|$ in expectation, the stepsize change associated to the path is larger 1 in expectation, in the sense that $\mathbb{E}\left[\Gamma_{d_\sigma}(p)^{-k}\right] \leqslant 1 - k\frac{\delta_\sigma}{d_\sigma}$. As was established earlier, this condition is satisfied by both the CSA and the modified CSA. In this section, we prove in Proposition 4.10 a sufficient condition on the initial state $(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0)$ of the Markov chain satisfying (4.17) such that at iteration $t = 1$ (case (i) with $\alpha = 0$) or 2 (case (ii)) the path condition $\mathbb{E}\|p_t\| \geqslant (1+\delta_p)\mathbb{E}\|\nu^d_U\|$ is satisfied for a large enough $\delta_p > 0$. This will allow us to use assumption **Γ**3 to prove Proposition 4.18, in which we control the updated mean z_1 or z_2, depending on the current mean z_0.

Proposition 4.10 relies on the assumption that the effective mass μ_{eff} defined in (4.4) is large enough for the following inequality to hold

$$\sqrt{\mu_{\text{eff}}} \left| I_{\nu^1_U} \right| + \frac{\sqrt{2}}{2} \mathbb{E}\|\nu^{d-1}_U\| > (1 + 2\delta_p) \frac{\mathbb{E}\|\nu^d_U\|}{\sqrt{c_\sigma(2 - c_\sigma)}} \ , \quad (4.71)$$

where

$$I_{\nu^1_U} = \sqrt{2} \int \mathbb{1}_{\xi_1 \leqslant \xi_2} \xi_1 \nu^1_U(\mathrm{d}\xi_1) \nu^1_U(\mathrm{d}\xi_2) = \frac{\sqrt{2}}{2} \mathbb{E}_{(\xi^1, \xi^2) \sim \nu^2_U}[\min(\xi^1, \xi^2)] < 0 \quad (4.72)$$

is proportional to the expectation of the first order statistics of two independent random variables with probability distribution ν^1_U. If $\mu \leqslant \lambda$ is large enough, we can always find weights such that (4.71) is

satisfied, as we have $\mu_{\text{eff}} = \mu$ with equal weights [72]. When $\mu = \lambda/2$, we find that $\mu_{\text{eff}} \approx \mu/2$ for the default weights, whereas $\mu_{\text{eff}} \approx \lambda/\pi$ for the optimal weights [12].

In the next lemma, we give a lower bound on the expected shift of the mean with a linear selection function that is given a vector $l \in \mathbb{R}^d$, $F_l(u) = 2\langle l, u\rangle$. Remark that a linear selection function occurs with a linear objective function $f(x) = 2\langle \tilde{l}, x\rangle$ since in this case $F_{t+1}(u) = f\left(x^* + z_t + \sqrt{\boldsymbol{\Sigma}_t}u\right) = 2\langle \tilde{l}, x^* + z_t\rangle + 2\langle \tilde{l}, \sqrt{\boldsymbol{\Sigma}_t}u\rangle \overset{\text{sel}}{\sim} 2\langle \sqrt{\boldsymbol{\Sigma}_t}\tilde{l}, u\rangle$. In the sequel it will appear as the limit of other selection functions. The proof of Lemma 4.3 is given in C.8.

Lemma 4.3. Suppose that $\mu \leqslant \lambda/2$, the weights $\mathbf{w}_m$ satisfy **W1**, and that the probability distribution ν_U^d is a standard multivariate normal distribution (satisfies **N5**). Let $l \in \mathbb{R}^d$ be a nonzero vector and define the selection function $F_l \colon \mathbb{R}^d \to \mathbb{R}$ with $F_l(\cdot) = 2\langle l, \cdot\rangle$. Let $U = (U_1, \ldots, U_\lambda) \sim \nu_U^{d\lambda}$ and $s_{F_l;U} \in \mathfrak{S}_\lambda$ be the permutation defined by the selection function via (4.48). Then

$$\mathbb{E}\left\|\sum_{i=1}^{\mu} w_i^m U^{s_{F_l;U}(i)}\right\| \geqslant \left|I_{\nu_U^1}\right| + \sqrt{\frac{1}{2\mu_{\text{eff}}}}\mathbb{E}\|\nu_U^{d-1}\| \ , \tag{4.73}$$

where $I_{\nu_U^1}$ is defined in (4.72).

Consequently, we obtain the following bounds on the expected step length when the initial state is such that $\|\boldsymbol{\Sigma}_0^{1/2} z_0\|$ is sufficiently large which will be useful to estimate the expected length of the path variable p_t in the proof of Proposition 4.10.

Lemma 4.4. Suppose that $\mu \leqslant \lambda/2$, that the objective function f is ellipsoidal (satisfies **F3**) with quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}_{++}^d$, that the weights $\mathbf{w}_m$ satisfy **W1**, and that the distribution ν_U^d is a standard multivariate normal distribution (satisfies **N5**). Let $I_{\nu_U^1}$ be defined in (4.72). When the initial state is such that $\|\lambda_1(\boldsymbol{\Sigma}_0)^{-1}\boldsymbol{\Sigma}_0^{1/2} z_0\| \to \infty$, the norm of the weighted sum of the selected samples satisfies

$$\liminf_{\|\lambda_1(\boldsymbol{\Sigma}_0)^{-1}\boldsymbol{\Sigma}_0^{1/2} z_0\| \to \infty} \mathbb{E}_0\left\|\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right\| \geqslant \left|I_{\nu_U^1}\right| + \frac{\sqrt{2}}{2}\sqrt{\frac{1}{\mu_{\text{eff}}}}\mathbb{E}\|\nu_U^{d-1}\| \ . \tag{4.74}$$

Consequently, for every $\varepsilon > 0$, there exists $M_\sigma > 0$ such that, if $\|\boldsymbol{\Sigma}_0^{1/2} z_0\| \geqslant M_\sigma \lambda_1(\boldsymbol{\Sigma}_0)$, then

$$\mathbb{E}_0\left\|\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right\| \geqslant \left|I_{\nu_U^1}\right| + \frac{\sqrt{2}}{2}\sqrt{\frac{1}{\mu_{\text{eff}}}}\mathbb{E}\|\nu_U^{d-1}\| - \varepsilon \ . \tag{4.75}$$

Proof. Consider a sequence of states $\{\theta_0^k\}_{k\in\mathbb{N}} = \{(z_0^k, p_0^k, q_0^k, \boldsymbol{\Sigma}_0^k, r_0^k)\}_{k\in\mathbb{N}} \in \mathsf{X}^{\mathbb{N}}$ such that

$$\lim_{k\to\infty} \|\lambda_1(\boldsymbol{\Sigma}_0^k)^{-1}(\boldsymbol{\Sigma}_0^k)^{1/2} z_0^k\| = +\infty \ .$$

Let $k \in \mathbb{N}$. Without loss of generality, up to considering a coordinate system composed of eigenvectors of $\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0^k\mathbf{H}^{1/2}$, since the Euclidean norm of $\lambda_(\boldsymbol{\Sigma}_0^k)^{-1}(\boldsymbol{\Sigma}_0^k)^{1/2} z_0^k$ is independent of the coordinate system we choose, we assume that $\mathbf{D}_k := \mathbf{H}^{1/2}\boldsymbol{\Sigma}_0^k\mathbf{H}^{1/2}$ is diagonal with diagonal elements that are decreasingly ordered. Assume moreover without loss of generality, up to change the sign of some coordinates in our coordinate system, that $[y_k]_j := [\mathbf{H}^{1/2} z_0^k]_j \geqslant 0$ for all $j = 1, \ldots, d$. Denote $F_1^k(u) = f(z_0^k + \sqrt{\boldsymbol{\Sigma}_0^k}u)$. By Corollary 4.3, denoting $\mathbf{R}_{0,k}^{\mathbf{H}}$ the

orthogonal matrix of $\mathbb{R}^d$ defined in (4.54), since we have assumed $\mathbf{D}_k$ diagonal, the coordinates in the basis $e_k(\mathbf{D}_k)$ correspond to the coordinates in the original basis and we find that

$$F_1^k((\mathbf{R}_{0,k}^{\mathbf{H}})^{-1}u) \overset{\text{sel}}{\sim} \sum_{j=1}^{d} 2\sqrt{\lambda_j(\mathbf{D}_k)}\,[y_k]_j\,[u]_j + \lambda_j(\mathbf{D}_k)\,[u]_j^2 \ .$$

Then, by dividing by $C_k = 2\max_{j=1,\dots,d}\sqrt{\lambda_j(\mathbf{D}_k)}\lambda_1(\mathbf{D}_k)^{-1/2}[y_k]_j$ and by $\sqrt{\lambda_1(\mathbf{D}_k)}$, we obtain

$$F_1^k((\mathbf{R}_{0,k}^{\mathbf{H}})^{-1}u) \overset{\text{sel}}{\sim} \sum_{j=1}^{d} \frac{\sqrt{\lambda_j(\mathbf{D}_k)}}{C_k}\left(2\lambda_1(\mathbf{D}_k)^{-1/2}\,[y_k]_j\,[u]_j + \sqrt{\frac{\lambda_j(\mathbf{D}_k)}{\lambda_1(\mathbf{D}_k)}}[u]_j^2\right) =: G^k(u) \ . \tag{4.76}$$

Since $\|\lambda_1(\mathbf{\Sigma}_0^k)^{-1}(\mathbf{\Sigma}_0^k)^{1/2}z_0^k\|$ tends to $+\infty$ when k goes to $+\infty$, using the matrix inequalities $\lambda_d(\mathbf{H})\mathbf{\Sigma}_0^k \preceq \mathbf{D}_k \preceq \lambda_1(\mathbf{H})\mathbf{\Sigma}_0^k$, we have that

$$\lambda_1(\mathbf{D}_k)^{-1/2}C_k = 2\|\lambda_1(\mathbf{D}_k)^{-1}\sqrt{\mathbf{D}_k}y_k\|_\infty \geqslant \mathrm{Cond}(\mathbf{H})^{-1}\|\lambda_1(\mathbf{\Sigma}_0^k)^{-1}(\mathbf{\Sigma}_0^k)^{1/2}z_0^k\|_\infty$$

tends to $+\infty$ as well. Moreover, up to considering subsequences of $\{\theta_0^k\}_{k\in\mathbb{N}}$, we suppose that the following limit is well defined and, by definition of C_k, is equal to 1 for at least one $j \in \{1,\dots,d\}$:

$$l_j = \lim_{k\to\infty} \frac{2\sqrt{\lambda_j(\mathbf{D}_k)}\lambda_1(\mathbf{D}_k)^{-1/2}\,[y_k]_j}{C_k} \in [0,1] \ .$$

Furthermore,

$$\lim_{k\to\infty} \frac{\sqrt{\lambda_j(\mathbf{D}_k)}\sqrt{\frac{\lambda_j(\mathbf{D}_k)}{\lambda_1(\mathbf{D}_k)}}}{C_k} \leqslant \lim_{k\to\infty} \frac{\sqrt{\lambda_1(\mathbf{D}_k)}}{C_k} = 0 \ .$$

Then, the sequence of functions $\{G^k\}_{k\in\mathbb{N}}$ defined in (4.76) converges point-wise to $F_l\colon \mathbb{R}^d \to \mathbb{R}$, where

$$F_l(u) = 2\sum_{j=1}^{d} l_j[u]_j = 2\langle l, u\rangle \ .$$

Then, by Proposition 4.7 applied to the function $\phi\colon (u_1,\dots,u_\lambda) \in \mathbb{R}^{d\lambda} \mapsto \|\sum_{i=1}^{\mu} w_i^m u_i\|$ which is integrable with respect to $\nu_U^{d\lambda}$ by **N**4, we have

$$\lim_{k\to\infty} \mathbb{E}_0\left\|\sum_{i=1}^{\mu} w_i^m U_1^{s_{F_1^k;U_1}(i)}\right\| = \mathbb{E}_0\left\|\sum_{i=1}^{\mu} w_i^m U_1^{s_{F_l;U_1}(i)}\right\| \ .$$

By Lemma 4.3, this proves (4.74) and (4.75).

□

We are now ready to state and prove Proposition 4.10 where in (ii) we state a sufficient condition on the initial state for the expected updated path to be large enough, and thus such that the expected inverse stepsize change is smaller than 1 after one or two iterations (the consequence after one iteration holds when the inequality $(1-c_\sigma)\mathbb{E}\|p_1\| \leqslant \delta_p\mathbb{E}\|\nu_U^d\|$ is not satisfied). The statement (i) of Proposition 4.10 is used to prove (ii).

Proposition 4.10. Consider hyperparameters of CMA-ES satisfying **H**1. Suppose that $\mu \leqslant \lambda/2$, that the objective function f is ellipsoidal (satisfying **F**3) with quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}^d_{++}$, that the weights $\mathbf{w}_m$ satisfy **W**1 and that the distribution ν_U^d is a standard multivariate normal distribution (i.e., satisfies **N**5). The following holds:

(i) For every $\delta_p > 0$, if $(1-c_\sigma)\mathbb{E}\|p_0\| \leqslant \delta_p \mathbb{E}\|\nu_U^d\|$ and if μ_{eff} is large enough for the inequality (4.71) to hold, then there exists $M_\sigma > 0$, such that if $\mathbb{P}\left[\left\|\boldsymbol{\Sigma}_0^{1/2} z_0\right\| \geqslant M_\sigma \lambda_1\left(\boldsymbol{\Sigma}_0\right)\right] \geqslant 1-\alpha$, then

$$\mathbb{E}\|p_1\| \geqslant (1-2\alpha)(1+\delta_p)\mathbb{E}\|\nu_U^d\| \tag{4.77}$$

for any $\alpha \in [0,1]$.

(ii) Assume moreover that the stepsize change Γ_{d_σ} satisfies **Γ**1. Let $\delta_p > 0$ and $\bar{q} > 0$, and suppose that μ_{eff} is large enough for the inequality (4.71) to hold and that c_1 is sufficiently small. Then, there exists $M_\sigma > 0$ such that, for every $(z_0, p_0, q_0, \boldsymbol{\Sigma}_0, r_0) \in \mathsf{X} = \mathbb{R}^{3d} \times \mathcal{S}^d_{++} \times (0, +\infty)$, if $(1-c_\sigma)\mathbb{E}\|p_1\| \leqslant \delta_p \mathbb{E}\|\nu_U^d\|$, $\|\boldsymbol{\Sigma}_0^{1/2} z_0\| \geqslant M_\sigma \lambda_1\left(\boldsymbol{\Sigma}_0\right)$ and $r_0^{-1/2}(1-c_c)\|q_0\| \leqslant \bar{q}\sqrt{\lambda_1\left(\boldsymbol{\Sigma}_0\right)}$, then

$$\mathbb{E}\|p_2\| \geqslant (1+\delta_p)\mathbb{E}\|\nu_U^d\| \ . \tag{4.78}$$

Proof. We start by proving (i). Let $\delta_p > 0$ and suppose $(1-c_\sigma)\mathbb{E}\|p_0\| \leqslant \delta_p \mathbb{E}\|\nu_U^d\|$ and that μ_{eff} is large enough for the inequality (4.71) to hold. Let $M_\sigma > 0$ be large enough so that, if $\|\boldsymbol{\Sigma}_0^{1/2} z_0\| \geqslant M_\sigma \lambda_1\left(\boldsymbol{\Sigma}_0\right)$, then, by Lemma 4.4 and inequality (4.71), we obtain

$$\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbb{E}_0\left\|\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right\| \geqslant (1+2\delta_p)\mathbb{E}\|\nu_U^d\| \ . \tag{4.79}$$

If instead $\|\boldsymbol{\Sigma}_0^{1/2} z_0\| \geqslant M_\sigma \lambda_1\left(\boldsymbol{\Sigma}_0\right)$ happens with probability at least $(1-\alpha)$, then we obtain

$$\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbb{E}\left\|\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right\| \geqslant \mathbb{P}\left[\|\boldsymbol{\Sigma}_0^{1/2} z_0\| \geqslant M_\sigma \lambda_1\left(\boldsymbol{\Sigma}_0\right)\right] \times$$
$$\mathbb{E}\left[\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbb{E}_0\left\|\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right\| \ \middle| \ \|\boldsymbol{\Sigma}_0^{1/2} z_0\| \geqslant M_\sigma \lambda_1\left(\boldsymbol{\Sigma}_0\right)\right] \geqslant (1-\alpha)(1+2\delta_p)\mathbb{E}\|\nu_U^d\| \ .$$

Then, we obtain by the inequality $\|x+y\| \geqslant \|y\| - \|x\|$ applied to $p_1 = (1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}$:

$$\mathbb{E}\|p_1\| \geqslant \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbb{E}\left\|\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right\| - (1-c_\sigma)\mathbb{E}\|p_0\| \tag{4.80}$$
$$\geqslant (1-\alpha)(1+2\delta_p)\mathbb{E}\|\nu_U^d\| - \delta_p\mathbb{E}\|\nu_U^d\| \geqslant (1-2\alpha)(1+\delta_p)\mathbb{E}\|\nu_U^d\| \ . \tag{4.81}$$

We now prove (ii). Let $\delta_p' = 2\delta_p$ and let $M_\sigma' > 0$ be such that (i) holds with δ_p' and M_σ'. We prove below that for an arbitrary small $\alpha > 0$, if the assumptions in (ii) are satisfied for a sufficiently large $M_\sigma > 0$, then

$$\mathbb{P}\left[\left\|\boldsymbol{\Sigma}_1^{1/2} z_1\right\| \geqslant M_\sigma' \lambda_1\left(\boldsymbol{\Sigma}_1\right)\right] > 1-\alpha \ .$$

Let $M_\sigma > 0$ (we fix its value later) and assume that $\|\mathbf{\Sigma}_0^{1/2} z_0\| \geqslant M_\sigma \lambda_1(\mathbf{\Sigma}_0)$. We have, by definition of z_1 in (4.17) $z_1 = r_1^{-1/2}\Gamma_{d_\sigma}(p_1)^{-1} \times \left(z_0 + c_m \mathbf{\Sigma}_0^{1/2} \sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right)$ then by multiplying by $\mathbf{\Sigma}_1^{1/2}$ and rearranging we obtain

$$\mathbf{\Sigma}_1^{1/2} z_1 = r_1^{-1/2}\Gamma_{d_\sigma}(p_1)^{-1} \times \left(\mathbf{\Sigma}_1^{1/2} z_0 + c_m \mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_0^{1/2} \sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right) \quad . \tag{4.82}$$

We want to lower-bound $\|\mathbf{\Sigma}_1^{1/2} z_1\|$ and first derive bounds on the two RHS summands. On the one hand, by definition of $\mathbf{\Sigma}_1$ in (4.17) we have $\mathbf{\Sigma}_1^{1/2} \succeq r_1^{-1/2}(1 - c_1 - c_\mu)^{1/2}\mathbf{\Sigma}_0^{1/2}$ and thus by (4.44)

$$\left\|r_1^{-1/2}\mathbf{\Sigma}_1^{1/2} z_0\right\| \geqslant r_1^{-1}(1 - c_1 - c_\mu)^{1/2} \left\|\mathbf{\Sigma}_0^{1/2} z_0\right\| \geqslant r_1^{-1}(1 - c_1 - c_\mu)^{1/2} M_\sigma \lambda_1(\mathbf{\Sigma}_0) \quad . \tag{4.83}$$

On the other hand, by Proposition 4.8,

$$\lambda_1(\mathbf{\Sigma}_1) \leqslant r_1^{-1}\left[\left(1 - c_1 - c_\mu + (2c_1\mu_{\text{eff}} + c_\mu) d\mu \|U_1\|_\infty^2\right)\lambda_1(\mathbf{\Sigma}_0) + 2c_1 r_0^{-1}(1 - c_c)^2 \|q_0\|^2\right]$$

where $r_0^{-1}(1-c_c)^2\|q_0\|^2 \leqslant \bar{q}^2 \lambda_1(\mathbf{\Sigma}_0)$. Denoting $\kappa_1 = (1 - c_1 - c_\mu + (2c_1\mu_{\text{eff}} + c_\mu) d\mu \|U_1\|_\infty^2)$ we obtain

$$\lambda_1(\mathbf{\Sigma}_1) \leqslant r_1^{-1}(\kappa_1 + 2c_1\bar{q}^2)\lambda_1(\mathbf{\Sigma}_0) \tag{4.84}$$

and thus since for $a \geqslant 0, b \geqslant 0$, $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$

$$\sqrt{\lambda_1(\mathbf{\Sigma}_1)} \leqslant r_1^{-1/2}(\sqrt{\kappa_1} + \sqrt{2c_1}\bar{q})\sqrt{\lambda_1(\mathbf{\Sigma}_0)} \quad . \tag{4.85}$$

We can now bound the norm of $\left\|r_1^{-1/2} c_m \mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_0^{1/2} \sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right\|$ using the property that for a symmetric matrix $\|\mathbf{A}x\| \leqslant |||\mathbf{A}||| \|x\|$ where the subordinate matrix norm $|||\mathbf{A}||| = \lambda_1(\mathbf{A})$ and use the previous equation to obtain

$$\begin{aligned}\left\|r_1^{-1/2} c_m \mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_0^{1/2} \sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right\| &\leqslant r_1^{-1/2} c_m \lambda_1(\mathbf{\Sigma}_1^{1/2})\lambda_1(\mathbf{\Sigma}_0^{1/2})\|U_1\| \\ &\leqslant c_m r_1^{-1}\left(\sqrt{\kappa_1} + \sqrt{2c_1}\bar{q}\right)\lambda_1(\mathbf{\Sigma}_0)\|U_1\| \end{aligned} \tag{4.86}$$

where we denote $\|U_1\| = \max_{i=1,\ldots,\lambda} \|U_1^i\|$. Therefore, applying the triangular inequality consequence $\|a+b\| \geqslant \|a\| - \|b\|$ to (4.82) and using (4.83) and (4.86), we obtain:

$$\left\|\mathbf{\Sigma}_1^{1/2} z_1\right\| \geqslant r_1^{-1}\Gamma_{d_\sigma}(p_1)^{-1} \times \left((1 - c_1 - c_\mu)^{1/2} M_\sigma - c_m\left(\sqrt{\kappa_1} + \sqrt{2c_1}\bar{q}\right)\|U_1\|\right)\lambda_1(\mathbf{\Sigma}_0) \tag{4.87}$$

and therefore, by (4.84)

$$\left\|\mathbf{\Sigma}_1^{1/2} z_1\right\| \geqslant \frac{\left((1 - c_1 - c_\mu)^{1/2} M_\sigma - c_m\left(\sqrt{\kappa_1} + \sqrt{2c_1}\bar{q}\right)\|U_1\|\right)}{\Gamma_{d_\sigma}(p_1)(\kappa_1 + 2c_1\bar{q}^2)}\lambda_1(\mathbf{\Sigma}_1) \quad . \tag{4.88}$$

We denote $\mathsf{U}(c_1)$ the event[a]

$$\frac{1}{2}(1 - c_1 - c_\mu)^{1/2} M_\sigma - c_m\left(\sqrt{\kappa_1} + \sqrt{2c_1}\bar{q}\right)\|U_1\| \geqslant 0$$

and we notice that for sufficiently large values of M_σ, that the probability that $\mathsf{U}(c_1)$ occurs tends to 1 when c_1 tends to 0. For $M > 0$, we denote $\mathsf{V}(M)$ the event $\Gamma_{d_\sigma}(p_1) \leqslant M$. However, since $\mathbb{E}\|p_1\| \leqslant \alpha_\sigma := \delta_p \mathbb{E}\|\nu_U^d\|/(1-c_\sigma)$, we have by Markov's inequality $\mathbb{P}[\|p_1\| \leqslant P] \geqslant 1 - \alpha_\sigma/P$ for $P > 0$. Moreover, if $\|p_1\| \leqslant P$, then since by **Γ**1, $\Gamma_{d_\sigma}^{-1}$ is locally Lipschitz and thus continuous we have that $\Gamma_{d_\sigma}^{-1}(p) \leqslant M_P$, where $M_P = \max\{\Gamma_{d_\sigma}^{-1}(p) \mid \|p\| \leqslant P\} < +\infty$. Therefore, since when $M \to \infty$, the probability of $M_P \leqslant M$ tends to 1, then the probability of $\mathsf{V}(M)$ tends to 1 when $M \to +\infty$. Remark that, as a consequence of (4.87) and the inequality $\|\mathbf{\Sigma}_0^{1/2} z_0\| \geqslant M_\sigma \lambda_1(\mathbf{\Sigma}_0)$, when $\mathsf{U}(c_1)$ and $\mathsf{V}(M)$ occur, we have

$$\left\|\mathbf{\Sigma}_1^{1/2} z_1\right\| \geqslant \frac{(1-c_1-c_\mu)^{1/2} M_\sigma}{2M(\kappa_1 + 2c_1\bar{q}^2)} \lambda_1(\mathbf{\Sigma}_0) \quad .$$

Finally, consider $\mathsf{W}(M_\sigma)$ the event $(1-c_1-c_\mu)^{1/2} M_\sigma \geqslant M'_\sigma \times 2M(\kappa_1 + 2c_1\bar{q}^2)$ which probability tends to 1 when M_σ goes to $+\infty$. We have then

$$\mathbb{P}\left[\left\|\mathbf{\Sigma}_1^{1/2} z_1\right\| \geqslant M'_\sigma \lambda_1(\mathbf{\Sigma}_1)\right] \geqslant \mathbb{P}\left[\mathsf{U}(c_1) \cap \mathsf{V}(M) \cap \mathsf{W}(M_\sigma)\right]$$

tends to 1 when c_1 to 0 and M and M_σ to $+\infty$. Choose $\alpha > 0$, $c_1 \geqslant 0$ and $M_\sigma > 0$ such that

$$\mathbb{P}\left[\left\|\mathbf{\Sigma}_1^{1/2} z_1\right\| \geqslant M'_\sigma \lambda_1(\mathbf{\Sigma}_1)\right] > 1 - \alpha$$

and $(1-2\alpha)(1+\delta'_p) = (1-2\alpha)(1+2\delta_p) = 1 + \delta_p$. Then, applying (i) ends the proof. □

[a] Obviously the event $\mathsf{U}(c_1)$ does not only depend on the parameter c_1 but on c_μ, c_m and $\bar{q}$ as well. However since we only discuss the choice of c_1 in this proof, we omit the other dependencies in this notation.

We end this section by showing that $\|p_t\|$ has finite moments which is needed in order to use the assumption **Γ**3 on the stepsize change, we require that the moments of $\|p_t\|$ are finite.

Proposition 4.11. Consider the Markov chain (4.17) and suppose that the sampling distribution ν_U^d satisfies **N**4. Let $k \in \mathbb{N}$. If p_0 is a random variable such that $\mathbb{E}\|p_0\|^k < \infty$, then for every $t \in \mathbb{N}$, $\mathbb{E}\|p_t\|^k < \infty$.

Proof. We proceed by induction. Indeed, for $t \in \mathbb{N}$, if $\mathbb{E}\|p_t\|^k < \infty$ (and thus $\mathbb{E}\|p_t\|^j < +\infty$ for $j \leqslant k$), since $p_{t+1} = (1-c_\sigma)p_t + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}} \sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)}$, then

$$\|p_{t+1}\|^k \leqslant \sum_{j=0}^k \binom{k}{j} (1-c_\sigma)^j \|p_t\|^j \times [c_\sigma(2-c_\sigma)\mu_{\text{eff}}]^{(k-j)/2} \left\|\sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)}\right\|^{k-j} \quad .$$

By Hölder's inequality, we obtain:

$$\mathbb{E}\|p_{t+1}\|^k \leqslant \sum_{j=0}^k \binom{k}{j} (1-c_\sigma)^j [c_\sigma(2-c_\sigma)\mu_{\text{eff}}]^{(k-j)/2} \left[\mathbb{E}\|p_t\|^k\right]^{j/k} \left[\mathbb{E}\left\|\sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)}\right\|^k\right]^{(k-j)/k} \quad .$$

Since $\mathbb{E}\|\sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)}\|^k \leqslant \mathbb{E}\max_{i=1}^\lambda \|U_{t+1}^i\|^k \leqslant \sum_{i=1}^\lambda \mathbb{E}\|U_{t+1}^i\|^k$ is finite by **N**4, then, by sum, $\mathbb{E}\|p_{t+1}\|^k < +\infty$, ending the proof. □

3.4 *Bounding the expected largest eigenvalue of the (normalized) covariance matrix*

In this section, we give sufficient conditions on the hyperparameters of CMA-ES and on the initial condition of the Markov chain (4.17) to obtain

$$\mathbb{E}[\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right)] \leqslant \rho\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)$$

for some constant $\rho \in (0,1)$. We start by assuming that the objective function is spherical, i.e., that $\mathbf{H} = \mathbf{I}_d$, and state the main result of the section in Proposition 4.15. In Corollary 4.4, we extend Proposition 4.15 to any positive definite matrix $\mathbf{H}$, based on the change of variable property in Proposition 4.6. This is the first step in order to partition the state space of the Markov chain $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$, and find the areas in which the Lyapunov function V evaluated in $(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)$ decreases in expectation.

We investigate the case where $\lambda_1(\mathbf{\Sigma}_0)$ is sufficiently large and is sufficiently larger than $\|\mathbf{\Sigma}_0^{1/2} z_0\|$. We establish first that when $\|\mathbf{\Sigma}_0^{1/2} z_0\|/\lambda_1(\mathbf{\Sigma}_0)$ converges to zero, the selection function associated to the spherical function which is equal to $u \mapsto \|z_0 + \sqrt{\mathbf{\Sigma}_0}u\|^2$ produces a ranking permutation that converges to the ranking permutation associated to $(z_0 = 0, \mathbf{\Sigma}_0)$ on the sphere function (i.e., to the selection function $u \mapsto \|\sqrt{\mathbf{\Sigma}_0}u\|^2$).

Lemma 4.5. Suppose that the sampling distribution ν_U^d satisfies **N**1 and **N**2. Consider λ random variables $U^1, \ldots, U^\lambda$ i.i.d. following ν_U^d. For $(z, \mathbf{\Sigma}) \in \mathbb{R}^d \times \mathcal{S}_{++}^d$, we consider the selection function $G_{z,\mathbf{\Sigma}}$ defined in (4.49), and $s_{z,\mathbf{\Sigma};U} = s_{G_{z,\mathbf{\Sigma}};U}$ the (almost surely unique) permutation of $\mathfrak{S}_\lambda$ which sorts the $G_{z,\mathbf{\Sigma}}$-values of U^i, for $i = 1, \ldots, \lambda$. Then, $\mathbb{P}[s_{z,\mathbf{\Sigma};U} \neq s_{0,\mathbf{\Sigma};U}]$ converges to 0 when $\frac{\|\mathbf{\Sigma}^{1/2}z\|}{\lambda_1(\mathbf{\Sigma})} \to 0$.

Proof. Let $\lambda_1(\mathbf{\Sigma}), \ldots, \lambda_d(\mathbf{\Sigma})$ be the decreasingly ordered eigenvalues (counted with multiplicity) of $\mathbf{\Sigma}$, and $(e_1(\mathbf{\Sigma}), \ldots, e_d(\mathbf{\Sigma}))$ a corresponding orthonormal basis of eigenvectors of $\mathbf{\Sigma}$. We use the notation $\langle\cdot\rangle_i = \langle\cdot, e_i(\mathbf{\Sigma})\rangle$ for $i = 1, \ldots, d$. The probability that $s_{z,\mathbf{\Sigma};U} \neq s_{0,\mathbf{\Sigma};U}$ equals $\mathbb{P}[s_{z,\mathbf{\Sigma};U} \neq s_{0,\mathbf{\Sigma};U}] = \mathbb{P}\left[\exists j,k \in \{1,\ldots,\lambda\} : G_{0,\mathbf{\Sigma}}(U^j) < G_{0,\mathbf{\Sigma}}(U^k) \text{ and } G_{z,\mathbf{\Sigma}}(U^j) > G_{z,\mathbf{\Sigma}}(U^k)\right]$. Using the union bound and the identical distribution of the U^i we can upper-bound the probability by

$$\mathbb{P}[s_{z,\mathbf{\Sigma};U} \neq s_{0,\mathbf{\Sigma};U}] \leqslant \binom{\lambda}{2} \times \mathbb{P}\left[G_{0,\mathbf{\Sigma}}(U^1) < G_{0,\mathbf{\Sigma}}(U^2) \text{ and } G_{z,\mathbf{\Sigma}}(U^1) > G_{z,\mathbf{\Sigma}}(U^2)\right].$$

By Proposition 4.5 the event $\{G_{0,\mathbf{\Sigma}}(U^1) < G_{0,\mathbf{\Sigma}}(U^2) \text{ and } G_{z,\mathbf{\Sigma}}(U^1) > G_{z,\mathbf{\Sigma}}(U^2)\}$ equals

$$\left\{\sum_{i=1}^d \lambda_i(\mathbf{\Sigma})\langle U^1\rangle_i^2 < \sum_{i=1}^d \lambda_i(\mathbf{\Sigma})\langle U^2\rangle_i^2\right\} \bigcap \left\{\sum_{i=1}^d 2\langle\mathbf{\Sigma}^{1/2}z\rangle_i\langle U^1\rangle_i + \lambda_i(\mathbf{\Sigma})\langle U^1\rangle_i^2 > \sum_{i=1}^d 2\langle\mathbf{\Sigma}^{1/2}z\rangle_i\langle U^2\rangle_i + \lambda_i(\mathbf{\Sigma})\langle U^2\rangle_i^2\right\}.$$

By subtracting $\sum_{i=1}^d \lambda_i(\mathbf{\Sigma})\langle U^1\rangle_i^2$ in the first event and by subtracting $\sum_{i=1}^d 2\langle\mathbf{\Sigma}^{1/2}z\rangle_i\langle U^1\rangle_i + \lambda_i(\mathbf{\Sigma})\langle U^2\rangle_i^2$ in the second event, the event in the previous equation equals $\{0 < \sum_{i=1}^d \lambda_i(\mathbf{\Sigma}) \times (\langle U^2\rangle_i^2 - \langle U^1\rangle_i^2) < 2\sum_{i=1}^d \langle\mathbf{\Sigma}^{1/2}z\rangle_i \times \langle U^1 - U^2\rangle_i\}$. By Cauchy-Schwarz inequality $\sum_{i=1}^d \langle\mathbf{\Sigma}^{1/2}z\rangle_i \times \langle U^1 - U^2\rangle_i \leqslant \|\mathbf{\Sigma}^{1/2}z\|\|U^1 - U^2\|$ so that the previous event is included into $\{0 < \sum_{i=1}^d \lambda_i(\mathbf{\Sigma}) \times (\langle U^2\rangle_i^2 - \langle U^1\rangle_i^2) < 2\|\mathbf{\Sigma}^{1/2}z\|\|U^1 - U^2\|\}$. Since by **N**1,

almost surely $\|U^1 - U^2\| > 0$, we can divide both sides of the inequality in the previous event by $\|U^1 - U^2\|$ and obtain

$$\mathbb{P}[s_{z,\mathbf{\Sigma};U} \neq s_{0,\mathbf{\Sigma};U}] = \binom{\lambda}{2} \times \mathbb{P}\left[0 < \frac{\langle \Lambda(\mathbf{\Sigma}), \langle U^2\rangle^2 - \langle U^1\rangle^2\rangle}{\|U^1 - U^2\|} < 2\|\mathbf{\Sigma}^{1/2} z\|\right]$$

where $\Lambda(\mathbf{\Sigma}) = (\lambda_1(\mathbf{\Sigma}), \ldots, \lambda_d(\mathbf{\Sigma}))$, and $\langle U^2\rangle^2 - \langle U^1\rangle^2 = (\langle U^2\rangle_1^2 - \langle U^1\rangle_1^2, \ldots, \langle U^2\rangle_d^2 - \langle U^1\rangle_d^2)$. However, by **N**2, the distribution of U^1 and U^2 is invariant to rotation, hence we have the following equality in distribution by Lemma 4.9

$$\frac{\langle \Lambda(\mathbf{\Sigma}), \langle U^2\rangle^2 - \langle U^1\rangle^2\rangle}{\|U^1 - U^2\|} \overset{\text{dist}}{=} \|\Lambda(\mathbf{\Sigma})\| \times \frac{[U^2]_1^2 - [U^1]_1^2}{\|U^1 - U^2\|} \ .$$

Denote Y the random variable $Y = \dfrac{[U^2]_1^2 - [U^1]_1^2}{\|U^1 - U^2\|}$. Since $\|\Lambda(\mathbf{\Sigma})\| = \sqrt{\lambda_1(\mathbf{\Sigma})^2 + \cdots + \lambda_d(\mathbf{\Sigma})^2} \geqslant \lambda_1(\mathbf{\Sigma}) \geqslant 0$, the even $\{0 < \|\Lambda(\mathbf{\Sigma})\| Y < 2\|\mathbf{\Sigma}^{1/2} z\|\}$ is included in the even $\{0 < \lambda_1(\mathbf{\Sigma}) Y < 2\|\mathbf{\Sigma}^{1/2} z\|\} = \{0 < Y < 2\|\mathbf{\Sigma}^{1/2} z\|/\lambda_1(\mathbf{\Sigma})\}$. Overall we have shown that $\mathbb{P}[s_{z,\mathbf{\Sigma};U} \neq s_{0,\mathbf{\Sigma};U}] \leqslant \binom{\lambda}{2} \times \mathbb{P}\left[0 < Y < 2\|\mathbf{\Sigma}^{1/2} z\|/\lambda_1(\mathbf{\Sigma})\right]$. The latter probability tends to 0 when $\|\mathbf{\Sigma}^{1/2} z\|/\lambda_1(\mathbf{\Sigma}) \to 0$. □

Using the previous lemma, we bound in the next proposition the perturbation induced by the mean on the diagonal coordinates of the rank-mu update matrix when the mean is sufficiently smaller than the largest eigenvalue of the covariance matrix. In other words, when $\|z_0\|/\lambda_1(\mathbf{\Sigma}_0)$ is small, the diagonal elements of the rank-mu update matrix become close to those when $\|z_0\| = 0$.

Proposition 4.12. Suppose that the probability distribution ν_U^d satisfies **N**1, **N**2 and **N**4, and that the weights $\mathbf{w}_c = (w_i^c)_{i=1,\ldots,\mu}$ are such that **W**1 holds. Let $\varepsilon > 0$, and $U = (U^1, \ldots, U^\lambda)$ be such that $U^1, \ldots, U^\lambda$ are i.i.d. and follow ν_U^d. Then, there exists $M_y > 0$, such that for all $\Lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}_{++}^d$ and $y = (y_1, \ldots, y_d) \in \mathbb{R}_+^d$ with $\|y\|_\infty \leqslant M_y \times \|\Lambda\|_\infty$, we have, for $i = 1, \ldots, d$

$$\left|\mathbb{E}\left[\tilde{\mathrm{M}}_{ii}^\mu(F_{y,\Lambda}; U) - \tilde{\mathrm{M}}_{ii}^\mu(F_{0,\Lambda}; U)\right]\right| \leqslant \varepsilon \tag{4.89}$$

where we use the following notation for the selection functions

$$F_{\alpha,\varphi}(u) \coloneqq \sum_{k=1}^d 2\alpha_k [u]_k + \varphi_k |[u]_k|^2 \qquad \text{for } (\alpha, \varphi) \in \mathbb{R}^d \times \mathbb{R}^d \tag{4.90}$$

and the notation for the rank-mu update matrix introduced in (4.62).

Proof. Let $y = (y_1, \ldots, y_d) \in \mathbb{R}_+^d$, $\Lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}_{++}^d$ and take $U = (U^1, \ldots, U^\lambda)$ a random vector with λ i.i.d. random vectors following ν_U^d. Let $s_{y,\Lambda;U}$ be the (almost surely unique) permutation that sorts the $F_{y,\Lambda}(U^i)$, $i = 1, \ldots, \lambda$. Likewise, let $s_{0,\Lambda;U}$ be the (almost surely unique) permutation that sorts the $F_{0,\Lambda}(U^i)$, $i = 1, \ldots, \lambda$. We can express the difference

in the diagonal elements of the covariance matrix as

$$\left|\mathbb{E}\left[\tilde{\mathrm{M}}_{ii}^{\mu}\left(F_{y,\Lambda};U\right)-\tilde{\mathrm{M}}_{ii}^{\mu}\left(F_{0,\Lambda};U\right)\right]\right| = \left|\mathbb{E}\left[\sum_{k=1}^{\mu} w_k^c\left(\left[U^{s_{F_{y,\Lambda};U}(k)}\right]_i^2-\left[U^{s_{F_{0,\Lambda};U}(k)}\right]_i^2\right)\times\mathbb{1}\{s_{F_{0,\Lambda};U}\neq s_{F_{y,\Lambda};U}\}\right]\right|. \quad (4.91)$$

Since the weights $\{w_k^c, k = 1\ldots,\mu\}$ have been assumed to sum to 1 and are nonnegative, they are smaller 1 and we can use the bound $\sum_{k=1}^{\mu} w_k^c\left(\left[U^{s_{F_{y,\Lambda};U}(k)}\right]_i^2-\left[U^{s_{F_{0,\Lambda};U}(k)}\right]_i^2\right)\leqslant 2\max_{k=1,\ldots,\lambda}\left[U^k\right]_i^2$ to upper-bound the previous equation by $\mathbb{E}\left[2\max_{k=1,\ldots,\lambda}\left[U^k\right]_i^2\times\mathbb{1}\{s_{F_{0,\Lambda};U}\neq s_{F_{y,\Lambda};U}\}\right]$. Using the Cauchy-Schwarz inequality we find that

$$\left|\mathbb{E}\left[\tilde{\mathrm{M}}_{ii}^{\mu}\left(F_{y,\Lambda};U\right)-\tilde{\mathrm{M}}_{ii}^{\mu}\left(F_{0,\Lambda};U\right)\right]\right|\leqslant 2\sqrt{\mathbb{P}[s_{0,\Lambda;U}\neq s_{y,\Lambda;U}]}\times\sqrt{\mathbb{E}\left[\max_{k=1,\ldots,\lambda}[U^k]_i^4\right]}. \quad (4.92)$$

By assumption **N**4, $\mathbb{E}\left[\max_{k=1,\ldots,\lambda}[U^k]_i^4\right]<+\infty$. Denote $\boldsymbol{\Lambda}$ the diagonal matrix composed of the diagonal elements Λ, by Proposition 4.5, $F_{y,\Lambda}\overset{\text{sel}}{\sim}G_{\boldsymbol{\Lambda}^{-1/2}y,\boldsymbol{\Lambda}}$ (see (4.50)). Since according to Lemma 4.5, $\mathbb{P}[s_{y,\Lambda;U}\neq s_{0,\Lambda;U}]$ goes to zero when $\|y\|/\|\boldsymbol{\Lambda}\|=\|\mathbf{y}\|/\lambda_{\mathbf{1}}(\boldsymbol{\Lambda})$ goes to zero, there exists $\tilde{M}_y>0$ sufficiently small such that, if $\|y\|\leqslant\tilde{M}_y\|\boldsymbol{\Lambda}\|$, then (4.89) holds. Since all norms are equivalent in finite-dimension Euclidean spaces, there exists $M_y>0$ such that, if $\|y\|_\infty\leqslant M_y\|\Lambda\|_\infty$, then (4.89) holds. □

We can now state the next result which describes the diagonal elements of the rank-mu update matrix when the initial covariance matrix $\boldsymbol{\Sigma}_0$ is sufficiently ill-conditioned, i.e., when it admits eigenvalues that satisfy $M_{\boldsymbol{\Sigma}}\lambda_j\left(\boldsymbol{\Sigma}_0\right)\leqslant\lambda_k\left(\boldsymbol{\Sigma}_0\right)$ and $m\lambda_k\left(\boldsymbol{\Sigma}_0\right)\geqslant\lambda_1\left(\boldsymbol{\Sigma}_0\right)$ (where the constants $M_{\boldsymbol{\Sigma}}$ and m are fixed in the next lemma), and additionally when the normalized mean is sufficiently small compared to the covariance matrix as explicated below. In this case, we find that the j^{th} diagonal element (i.e., corresponding to smaller eigenvalues) of the rank-mu update of the covariance matrix is expected to be larger than its k^{th} diagonal element (i.e., corresponding to larger eigenvalues), and we are able to uniformly bound the gap between these expected diagonal elements from below. Figure 4.1 illustrates this behavior.

Proposition 4.13. Consider the Markov chain $\Theta=\{(z_t,p_t,q_t,\boldsymbol{\Sigma}_t,r_t)\}_{t\in\mathbb{N}}$ obeying (4.16), when minimizing a spherical objective function f, i.e., satisfying **F**2, with a normalization function $R(\cdot)=\lambda_{\min}(\cdot)$, i.e., satisfying **R**2. We suppose moreover that the probability distribution ν_U^d is standard normal, i.e., satisfies **N**5, and that the weights $\mathbf{w}_c=(w_i^c)_{i=1}^{\mu}$ satisfy **W**1.
Let $m>1$ and $k,j\in\{1,\ldots,d\}$ such that $k<j$. Then, there exists constants $M_{\boldsymbol{\Sigma}}>m$, $M_y>0$ and $\delta>0$ such that for all $(z_0,p_0,q_0,\boldsymbol{\Sigma}_0,r_0)\in\mathsf{X}$, if $M_{\boldsymbol{\Sigma}}\lambda_j\left(\boldsymbol{\Sigma}_0\right)\leqslant\lambda_k\left(\boldsymbol{\Sigma}_0\right)$ and $\lambda_1\left(\boldsymbol{\Sigma}_0\right)\leqslant m\lambda_k\left(\boldsymbol{\Sigma}_0\right)$, and if $\|\boldsymbol{\Sigma}_0^{1/2}z_0\|_\infty\leqslant M_y\lambda_1\left(\boldsymbol{\Sigma}_0\right)$, then the j^{th} and k^{th} diagonal elements of the rank-mu matrix (defined in (4.61)) at iteration 1, $\tilde{\mathrm{M}}_1^{\mu}$, satisfy

$$\mathbb{E}\left[[\tilde{\mathrm{M}}_1^{\mu}]_{jj}-[\tilde{\mathrm{M}}_1^{\mu}]_{kk}\right]>\delta\ . \quad (4.93)$$

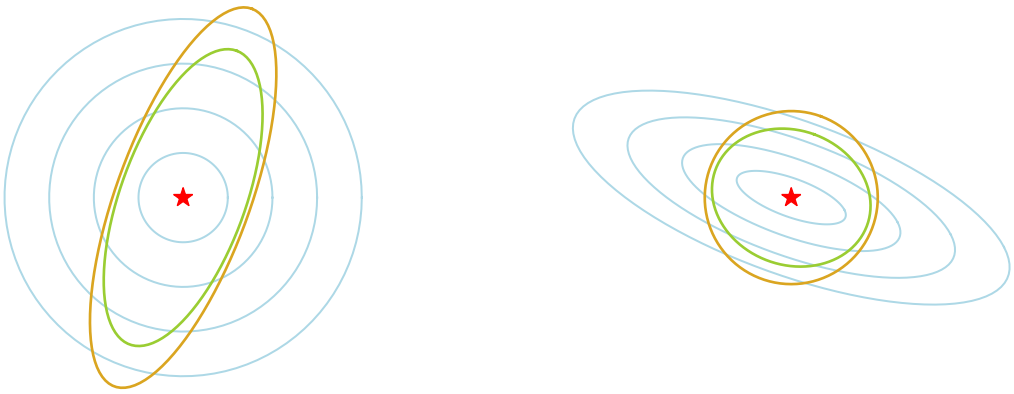


Figure 4.1: Shown are level sets of objective functions in cyan, of the density of sampling distributions with covariance matrix $\boldsymbol{\Sigma}_0$ in orange and of the expected rank-mu update matrix, $\mathbb{E}\tilde{\mathrm{M}}_1^\mu$, in light lime, where $z_0 = 0$ is in the optimum (red star). Left figure: when the objective function is spherical and $\boldsymbol{\Sigma}_0$ is ill conditioned, the average rank-mu matrix is shrinking the longest sampling axis. Right figure: transformation of the left figure in a coordinate system where $\boldsymbol{\Sigma}_0$ is the identity matrix while the orientation of eigenvectors is conserved.

Proof. Let $\Lambda = (\lambda_1, \ldots, \lambda_d) \in \mathcal{K}_m^k := [m^{-1}, 1]^k \times [0,1]^{d-k}$. Denote $F_\Lambda = F_{0,\Lambda}$, where we use the notation (4.90) for the selection functions. We express $\mathbb{E}\left[\tilde{\mathrm{M}}_{kk}^\mu(F_\Lambda; U)\right] = \sum_{i=1}^\mu \mathbb{E}\left[w^c_{s^{-1}_{F_\Lambda;U}(i)}{}^2 [U^i]_k^2\right]$ where $U = (U^1, \ldots, U^\lambda) \sim \nu_U^{d\lambda}$, and given $F \in \mathcal{F}$ (see (4.47)), $s_{F;U}$ is a permutation that satisfies (4.48). Hence

$$\begin{aligned}\mathbb{E}\left[\tilde{\mathrm{M}}_{kk}^\mu(F_\Lambda; U)\right] &= \sum_{i=1}^\lambda \mathbb{E}\left[w^c_{s^{-1}_{F_\Lambda;U}(i)} [U^i]_k^2\right] = \lambda \mathbb{E}\left[w^c_{1+\sum_{i\geqslant 2} \mathbb{1}\{F_\Lambda(U^i) < F_\Lambda(U^1)\}} [U^1]_k^2\right] \\ &= \lambda \times \int [u^1]_k^2 \left(\int w^c_{1+\sum_{i\geqslant 2} \mathbb{1}\{F_\Lambda(u^i) < F_\Lambda(u^1)\}} \nu_U^{d\lambda-1}(\mathrm{d}[u^i]_l, (i,l) \neq (1,k)) \right) \nu_U^1(\mathrm{d}[u^1]_k)\end{aligned}$$

Consider the random variables $X = [U^1]_k^2$, $Y = ([U^i]_j^2)_{(i,j)\neq(1,k)}$, $g(x) = x$ and $h(z) = h(x,y) = w^c_{1+\sum_{i\geqslant 2} \mathbb{1}\{\langle \Lambda, z^i\rangle < \langle \Lambda, z^1\rangle\}}$, for every $z = (z^1, \ldots, z_\lambda) \in (\mathbb{R}^d)^\lambda$, and by denoting $x = [z^1]_k$ and $y = ([z^i]_j)_{(i,j)\neq(1,k)}$. But, since $z^1 \mapsto \langle \Lambda, z^1\rangle$ is increasing with respect to $[z_1]_k$, that $i \mapsto w_i^c$ is nonincreasing and not constant by **W**1, then, by composition, the function $x \in \mathbb{R} \mapsto h(x,y)$ is nonincreasing and not almost everywhere constant for every $y \in \mathbb{R}^{d\lambda-1}$. Thus, by Theorem 4.6, we have

$$\mathrm{Cov}(g(X), h(X,Y)) = \mathrm{Cov}_{(u^1,\ldots,u^\lambda)\sim\nu_U^{d\lambda}} \left([u^1]_k^2, w^c_{1+\sum_{i\geqslant 2} \mathbb{1}\{F_\Lambda(u^i) < F_\Lambda(u^1)\}} \right) < 0 \ .$$

Hence since we have $\mathbb{E}\left[\tilde{\mathrm{M}}_{kk}^\mu(F_\Lambda; U)\right] = \lambda \mathrm{Cov}(g(X), h(X,Y)) + \lambda \mathbb{E}[g(X)]\mathbb{E}[h(X,Y)]$ we

obtain

$$\mathbb{E}\left[\tilde{\mathrm{M}}^{\mu}_{kk}(F_\Lambda;U)\right] < \lambda\mathbb{E}[g(X)]\mathbb{E}[h(X,Y)] =$$
$$\lambda\left(\int \nu_U^1(\mathrm{d}[u^1]_k)[u^1]_k^2\right)\left(\int \nu_U^{d\lambda}(\mathrm{d}u) w^c_{1+\sum_{i\geqslant 2}\mathbb{1}\{F_\Lambda(u^i)<F_\Lambda(u^1)\}}\right) = \lambda \times 1 \times \mathbb{E}\left[w^c_{s^{-1}_{F_\Lambda;U}(1)}\right] = 1$$

where we have used that $\lambda\mathbb{E}\left[w^c_{s^{-1}_{F_\Lambda;U}(1)}\right] = \mathbb{E}\left[\sum_{i=1}^{\lambda} w^c_{s^{-1}_{F_\Lambda;U}(i)}\right] = \sum_{i=1}^{\lambda} w_i^c = 1$ since the random variables $w^c_{s^{-1}_{F_\Lambda;U}(i)}$ for $i = 1,\dots,\lambda$ are identically distributed. Furthermore, this is true for all $\Lambda \in \mathcal{K}_m^k$. Since, by Proposition 4.7, $\Lambda \mapsto \mathbb{E}\left[\tilde{\mathrm{M}}^{\mu}_{kk}(F_\Lambda;U)\right]$ is continuous and $\mathcal{K}_m^k$ is compact, then

$$1 - 4\delta := \max_{\Lambda\in\mathcal{K}_m^k} \mathbb{E}\left[\tilde{\mathrm{M}}^{\mu}_{kk}(F_\Lambda;U)\right] < 1 \ . \tag{4.94}$$

Likewise, we have

$$\mathbb{E}\left[\tilde{\mathrm{M}}^{\mu}_{jj}(F_\Lambda;U)\right]$$
$$= \lambda\int [u^1]_j^2\left(\int \nu_U^{d\lambda-1}(\mathrm{d}[u^i]_l,(i,l)\neq(1,j)) w^c_{1+\sum_{i\geqslant 2}\mathbb{1}\{F_\Lambda(u^i)<F_\Lambda(u^1)\}}\right)\nu_U^1(\mathrm{d}[u^1]_j) \ .$$

Suppose that λ_j becomes sufficiently small, i.e., suppose that Λ tends to $\Lambda^* = (\lambda_1^*,\dots,\lambda_d^*)$ with $\lambda_j^* = 0$. Note that for every $\Lambda \in \mathbb{R}^d$,

$$w^c_{1+\sum_{i\geqslant 2}\mathbb{1}\{F_\Lambda(u^i)<F_\Lambda(u^1)\}}[u^1]_j^2 \leqslant w_1^c[u^1]_j^2$$

where the RHS defines an integrable function for $(u^1,\dots,u^\lambda) \sim \nu_U^{d\lambda}$. Thus by dominated convergence:

$$\lim_{\Lambda\to\Lambda^*}\mathbb{E}\left[\tilde{\mathrm{M}}^{\mu}_{jj}(F_\Lambda;U)\right]$$
$$= \lambda\int \nu_U^1(\mathrm{d}[u^1]_j)[u^1]_j^2 \times \int \nu_U^{d\lambda-1}(\mathrm{d}[u^i]_l,(i,l)\neq(1,j)) w^c_{1+\sum_{i\geqslant 2}\mathbb{1}\{F_{\Lambda^*}(u^i)<F_{\Lambda^*}(u^1)\}} = 1 \ .$$

To see that the term $\lambda\int \nu_U^1(\mathrm{d}[u^1]_j)[u^1]_j^2 \times \int \nu_U^{d\lambda-1}(\mathrm{d}[u^i]_l,(i,l)\neq(1,j)) w^c_{1+\sum_{i\geqslant 2}\mathbb{1}\{F_{\Lambda^*}(u^i)<F_{\Lambda^*}(u^1)\}}$ equals 1, we can go back to the expression of the term as $\sum_{i=1}^{\mu}\mathbb{E}\left[w^c_{U^{s_{F^*_\Lambda;U}(i)}_j}{}^2\right]$. However since $\lambda_j^* = 0$, then the selection does not influence the coordinates j of the random vectors such that $\{[U^{s_{F^*_\Lambda;U}(i)}]_j, i = 1\dots,\lambda\}$ are i.i.d. following a standard normal distribution and thus $\sum_{i=1}^{\mu}\mathbb{E}\left[w^c_{U^{s_{F^*_\Lambda;U}(i)}_j}{}^2\right] = 1$.
Hence, if $\lambda_j > 0$ is sufficiently small, i.e., given $M_{\boldsymbol{\Sigma}} > 0$ large enough, if $\lambda_j \leqslant M_{\boldsymbol{\Sigma}}^{-1}$, we have

$$\mathbb{E}\left[\tilde{\mathrm{M}}^{\mu}_{jj}(F_\Lambda;U)\right] > 1 - \delta \ . \tag{4.95}$$

Moreover, by Proposition 4.12, there exists $M_y > 0$ such that for all $\Lambda \in [0,1]^d$ we have for $l = 1,\dots,d$

$$\mathbb{E}\left|\tilde{\mathrm{M}}^{\mu}_{ll}(F_{y,\Lambda};U) - \tilde{\mathrm{M}}^{\mu}_{ll}(F_\Lambda;U)\right| \leqslant \delta \quad \text{for } y \in \mathbb{R}^d_+,\ \|y\|_\infty \leqslant M_y \tag{4.96}$$

where we use the notation (4.90) for $\tilde{\mathrm{M}}_{ll}^{\mu}(F_{y,\Lambda};U)$. Then, combining (4.94), (4.95) and (4.96), we get, for all $(y,\Lambda)\in\mathbb{R}_+^d\times\mathbb{R}_{++}^d$ such that $\lambda_k\geqslant m^{-1}$, $\lambda_j\leqslant M_{\boldsymbol{\Sigma}}^{-1}$ and $\|y\|_\infty\leqslant M_y$, that

$$\mathbb{E}\left[\tilde{\mathrm{M}}_{jj}^{\mu}(F_{y,\Lambda};U)-\tilde{\mathrm{M}}_{kk}^{\mu}(F_{y,\Lambda};U)\right]>\delta\ . \tag{4.97}$$

When $y=\sqrt{\boldsymbol{\Sigma}_0}z_0/\lambda_1(\boldsymbol{\Sigma}_0)$ and $\Lambda=(\lambda_1(\boldsymbol{\Sigma}_0)/\lambda_1(\boldsymbol{\Sigma}_0),\dots,\lambda_d(\boldsymbol{\Sigma}_0)/\lambda_1(\boldsymbol{\Sigma}_0))$, we obtain the desired result by Proposition 4.5. □

We derive now the upper bound (4.98) on the largest eigenvalue of the updated normalized covariance matrix $\boldsymbol{\Sigma}_1$ as a function of $\boldsymbol{\Sigma}_0$. It is in particular useful when c_μ is sufficiently small and the learning rate c_1 for the rank-one update is negligible compared to c_μ.

Proposition 4.14. Consider the CMA-ES algorithm optimizing a spherical objective function f satisfying **F**2 with a sampling distribution ν_U^d which is standard normal (**N**5) and that the weights $\mathbf{w}_c$ satisfy **W**1. Consider the associated normalized Markov chain $\Theta=\{(z_t,p_t,q_t,\boldsymbol{\Sigma}_t,r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) with the normalization function $R(\cdot)=\lambda_{\min}(\cdot)$ (satisfying **R**2). Assume that $c_1\in[0,1-c_\mu]$ is such that c_1/c_μ is sufficiently small.
Let $\tilde{\mathrm{M}}_1^{\mu}$ be the rank-mu update matrix defined in (4.61). Then, there exist a function $\varepsilon_{\text{frac}}\colon(0,1)\to\mathbb{R}_{++}$ with $\varepsilon_{\text{frac}}(c)/c\to 0$ when $c\to 0$, and a real-valued random variable ρ_1 with finite moments, such that for any initial condition $(z_0,p_0,q_0,\boldsymbol{\Sigma}_0,r_0)\in\mathsf{X}=\mathbb{R}^{3d}\times R^{-1}(\{1\})\times[1-c_1-c_\mu,+\infty)$

$$\lambda_1(\boldsymbol{\Sigma}_1)\leqslant\lambda_1(\boldsymbol{\Sigma}_0)\times\left(1+c_\mu\max_{k=1,\dots,k_1}[\tilde{\mathrm{M}}_1^{\mu}]_{kk}-c_\mu\min_{k=k_d,\dots,d}[\tilde{\mathrm{M}}_1^{\mu}]_{kk}+\varepsilon_{\text{frac}}(c_\mu)\times(1+\rho_1^2)\|U_1\|^2\right)$$
$$+2(1-c_1-c_\mu)^{-1}c_1r_0^{-1}\|q_0\|^2 \tag{4.98}$$

where $k_1=\max\{k=1,\dots,d:\lambda_k(\boldsymbol{\Sigma}_0)\geqslant(1-\sqrt{c_\mu})\lambda_1(\boldsymbol{\Sigma}_0)\}$ and $k_d=\min\{k=1,\dots,d:(1-\sqrt{c_\mu})\lambda_k(\boldsymbol{\Sigma}_0)\leqslant 1\}$.

Proof. Since R satisfies **R**2, it satisfies **R**1. Additionally, as the sampling distribution is Gaussian, it satisfies **N**4. Hence the assumptions of Proposition 4.9 are satisfied and its conclusion holds. Let ρ_1 be a positive random variable whose existence is proven in Proposition 4.9 and satisfying (4.70), hence such that ρ_1 has finite moments and is independent of the initial condition $(z_0,p_0,q_0,\boldsymbol{\Sigma}_0,r_0)\in\mathsf{X}$.
Without loss of generality, suppose that $\boldsymbol{\Sigma}_0$ is diagonal with decreasing diagonal elements.
By assumption on c_1, denote $c_1=\varepsilon_1(c_\mu)$ with $\varepsilon_1(c_\mu)=o(c_\mu)$ when $c_\mu\to 0$. Since, by definition of q_1 in (4.17)

$$q_1=r_0^{-1/2}(1-c_c)q_0+\sqrt{c_c(2-c_c)\mu_{\text{eff}}}\boldsymbol{\Sigma}_0^{1/2}\sum_{i=1}^{\mu}w_i^mU_1^{s_1(i)}\ ,$$

then by triangular inequality, using $\|\sum_{i=1}^{\mu}w_i^mU_1^{s_1(i)}\|^2\leqslant\|U_1\|^2=\sum_{i=1}^{\lambda}\|U_1^i\|^2$ and $c_c(2-c_c)\mu_{\text{eff}}\leqslant\mu_{\text{eff}}$, we obtain

$$\lambda_1\left(c_1q_1q_1^\top\right)=c_1\|q_1\|^2\leqslant 2c_1\|r_0^{-1/2}q_0\|^2+2c_1\times\lambda_1(\boldsymbol{\Sigma}_0)\left(\sqrt{\mu_{\text{eff}}}\|U_1\|\right)^2$$
$$=:2c_1r_0^{-1}\|q_0\|^2+\varepsilon_1(c_\mu)\|U_1\|^2\lambda_1(\boldsymbol{\Sigma}_0)\ . \tag{4.99}$$

Let $\tilde{\mathbf{\Sigma}}_1$ be defined as in (4.18). Then

$$\begin{aligned}\lambda_1(\tilde{\mathbf{\Sigma}}_1) &= e_1\left(\mathbf{\Sigma}_1\right)^\top \tilde{\mathbf{\Sigma}}_1 e_1\left(\mathbf{\Sigma}_1\right)\\ &= e_1\left(\mathbf{\Sigma}_1\right)^\top \left((1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_\mu\sqrt{\mathbf{\Sigma}_0}\tilde{\mathrm{M}}_1^\mu\sqrt{\mathbf{\Sigma}_0} + c_1 q_1 q_1^\top\right) e_1\left(\mathbf{\Sigma}_1\right)\\ &\leqslant (1-c_1-c_\mu) e_1\left(\mathbf{\Sigma}_1\right)^\top \mathbf{\Sigma}_0 e_1\left(\mathbf{\Sigma}_1\right) + c_\mu e_1\left(\mathbf{\Sigma}_1\right)^\top \sqrt{\mathbf{\Sigma}_0}\tilde{\mathrm{M}}_1^\mu\sqrt{\mathbf{\Sigma}_0} e_1\left(\mathbf{\Sigma}_1\right)\\ &\qquad + c_1 e_1\left(\mathbf{\Sigma}_1\right)^\top q_1 q_1^\top e_1\left(\mathbf{\Sigma}_1\right).\end{aligned}$$

With (4.42), we have $e_1\left(\mathbf{\Sigma}_1\right)^\top \mathbf{\Sigma}_0 e_1\left(\mathbf{\Sigma}_1\right) \leqslant \lambda_1\left(\mathbf{\Sigma}_0\right)$ and $c_1 e_1\left(\mathbf{\Sigma}_1\right)^\top q_1 q_1^\top e_1\left(\mathbf{\Sigma}_1\right) \leqslant \lambda_1\left(c_1 q_1 q_1^\top\right) \leqslant 2c_1 r_0^{-1}\|q_0\|^2 + \varepsilon_1(c_\mu)\|U_1\|^2\lambda_1\left(\mathbf{\Sigma}_0\right)$ where we have used (4.99) for the last inequality. We thus obtain

$$\begin{aligned}\lambda_1(\tilde{\mathbf{\Sigma}}_1) &\leqslant (1-c_1-c_\mu)\lambda_1\left(\mathbf{\Sigma}_0\right) + c_\mu \underbrace{e_1\left(\mathbf{\Sigma}_1\right)^\top \sqrt{\mathbf{\Sigma}_0}\tilde{\mathrm{M}}_1^\mu\sqrt{\mathbf{\Sigma}_0} e_1\left(\mathbf{\Sigma}_1\right)}_{\sum_{i=1}^\mu w_i^c (\sqrt{\mathbf{\Sigma}_0} e_1(\mathbf{\Sigma}_1))^\top U_1^{s_1(i)} U_1^{s_1(i)\top}\sqrt{\mathbf{\Sigma}_0} e_1(\mathbf{\Sigma}_1)}\\ &\qquad + \varepsilon_1(c_\mu)\|U_1\|^2\lambda_1\left(\mathbf{\Sigma}_0\right) + 2c_1 r_0^{-1}\|q_0\|^2\\ &= (1-c_1-c_\mu)\lambda_1\left(\mathbf{\Sigma}_0\right) + c_\mu\sum_{i=1}^\mu w_i^c |\langle\sqrt{\mathbf{\Sigma}_0} e_1\left(\mathbf{\Sigma}_1\right), U_1^{s_1(i)}\rangle|^2 + \varepsilon_1(c_\mu)\|U_1\|^2\lambda_1\left(\mathbf{\Sigma}_0\right)\\ &\qquad + 2c_1 r_0^{-1}\|q_0\|^2 \ .\end{aligned}$$

However, we have, by a triangular inequality, for $k_1 \in \{1,\dots,d\}$:

$$\begin{aligned}|\langle\sqrt{\mathbf{\Sigma}_0} e_1\left(\mathbf{\Sigma}_1\right), U_1^{s_1(i)}\rangle|^2 &\leqslant |\langle[\sqrt{\mathbf{\Sigma}_0} e_1\left(\mathbf{\Sigma}_1\right)]_{1:k_1}, [U_1^{s_1(i)}]_{1:k_1}\rangle|^2\\ &\qquad + |\langle[\sqrt{\mathbf{\Sigma}_0} e_1\left(\mathbf{\Sigma}_1\right)]_{k_1+1:d}, [U_1^{s_1(i)}]_{k_1+1:d}\rangle|^2\\ \leqslant \lambda_1\left(\mathbf{\Sigma}_0\right)|\langle[e_1\left(\mathbf{\Sigma}_1\right)]_{1:k_1}, [U_1^{s_1(i)}]_{1:k_1}\rangle|^2 &+ c_d\lambda_1\left(\mathbf{\Sigma}_0\right)\sum_{k=k_1+1}^d |\langle e_1\left(\mathbf{\Sigma}_1\right), e_k\left(\mathbf{\Sigma}_0\right)\rangle|^2\|U_1^{s_1(i)}\|^2 \ .\end{aligned}$$

where $c_d > 0$ only depends on the dimension d.
Let $m = (1-\sqrt{c_\mu})^{-1}$ and $k_1 = \max\{k = 1,\dots,d \mid m\lambda_k\left(\mathbf{\Sigma}_0\right) \geqslant \lambda_1\left(\mathbf{\Sigma}_0\right)\}$. Using Proposition 4.9, we get then

$$\begin{aligned}\lambda_1(\tilde{\mathbf{\Sigma}}_1) &\leqslant (1-c_1-c_\mu)\lambda_1\left(\mathbf{\Sigma}_0\right) + c_\mu\lambda_1\left(\mathbf{\Sigma}_0\right)\sum_{i=1}^\mu w_i^c |\langle[e_1\left(\mathbf{\Sigma}_1\right)]_{1:k_1}, [U_1^{s_1(i)}]_{1:k_1}\rangle|^2\\ &\qquad + c_\mu^2 c_d\lambda_1\left(\mathbf{\Sigma}_0\right)\rho_1^2 d\|U_1\|^2 + \varepsilon_1(c_\mu)\|U_1\|^2\lambda_1\left(\mathbf{\Sigma}_0\right) + 2c_1 r_0^{-1}\|q_0\|^2 \ .\end{aligned}$$

Then

$$\begin{aligned}\lambda_1(\tilde{\mathbf{\Sigma}}_1) \leqslant \lambda_1\left(\mathbf{\Sigma}_0\right) \times &\left(1-c_1-c_\mu + c_\mu\max_{k=1,\dots,k_1}[\tilde{\mathrm{M}}_1^\mu]_{kk} + c_\mu^2 c_d\rho_1^2 d\|U_1\|^2\right)\\ &+ \varepsilon_1(c_\mu)\|U_1\|^2\lambda_1\left(\mathbf{\Sigma}_0\right) + 2c_1 r_0^{-1}\|q_0\|^2 \ . \quad (4.100)\end{aligned}$$

Likewise, we have

$$\begin{aligned}\lambda_d(\tilde{\mathbf{\Sigma}}_1) &= e_d\left(\mathbf{\Sigma}_1\right)^\top \tilde{\mathbf{\Sigma}}_1 e_d\left(\mathbf{\Sigma}_1\right)\\ &= e_d\left(\mathbf{\Sigma}_1\right)^\top \left((1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_\mu\sqrt{\mathbf{\Sigma}_0}\tilde{\mathrm{M}}_1^\mu\sqrt{\mathbf{\Sigma}_0} + c_1 q_1 q_1^\top\right) e_d\left(\mathbf{\Sigma}_1\right)\\ &\geqslant (1-c_1-c_\mu)\underbrace{e_d\left(\mathbf{\Sigma}_1\right)^\top \mathbf{\Sigma}_0 e_d\left(\mathbf{\Sigma}_1\right)}_{\geqslant\lambda_d(\mathbf{\Sigma}_0)=1} + c_\mu e_d\left(\mathbf{\Sigma}_1\right)^\top \sqrt{\mathbf{\Sigma}_0}\tilde{\mathrm{M}}_1^\mu\sqrt{\mathbf{\Sigma}_0} e_d\left(\mathbf{\Sigma}_1\right)\end{aligned}$$

with

$$e_d\left(\mathbf{\Sigma}_1\right)^\top \sqrt{\mathbf{\Sigma}_0}\tilde{\mathrm{M}}_1^\mu\sqrt{\mathbf{\Sigma}_0}e_d\left(\mathbf{\Sigma}_1\right) = \sum_{i=1}^{\mu} w_i^c|\langle\sqrt{\mathbf{\Sigma}_0}e_d\left(\mathbf{\Sigma}_1\right), U_1^{s_1(i)}\rangle|^2$$

and, for $k_d \in \{1,\dots,d\}$:

$$\begin{aligned}|\langle\sqrt{\mathbf{\Sigma}_0}e_d\left(\mathbf{\Sigma}_1\right), U_1^{s_1(i)}\rangle|^2 &\geqslant |\langle[\sqrt{\mathbf{\Sigma}_0}e_d\left(\mathbf{\Sigma}_1\right)]_{k_d:d}, [U_1^{s_1(i)}]_{k_d:d}\rangle|^2 \\ &\quad - |\langle[\sqrt{\mathbf{\Sigma}_0}e_d\left(\mathbf{\Sigma}_1\right)]_{1:k_d-1}, [U_1^{s_1(i)}]_{1:k_d-1}\rangle|^2 \\ &\geqslant |\langle[\sqrt{\mathbf{\Sigma}_0}e_d\left(\mathbf{\Sigma}_1\right)]_{k_d:d}, [U_1^{s_1(i)}]_{k_d:d}\rangle|^2 - c_d\sum_{k=1}^{k_d-1}|\langle\sqrt{\mathbf{\Sigma}_0}e_d\left(\mathbf{\Sigma}_1\right), e_k\left(\mathbf{\Sigma}_0\right)\rangle|^2\|U_1^{s_1(i)}\|^2 \ .\end{aligned}$$

Let $k_d = \min\{k = 1,\dots,d \mid \lambda_k\left(\mathbf{\Sigma}_0\right) \leqslant m\}$. By Proposition 4.9, since $m \geqslant (1-\sqrt{c_\mu})^{-1}$, we thus obtain:

$$\lambda_d(\tilde{\mathbf{\Sigma}}_1) \geqslant 1 - c_1 - c_\mu + c_\mu(1-\sqrt{c_\mu})^{-1}\min_{k=k_d,\dots,d}[\tilde{\mathrm{M}}_1^\mu]_{kk} - c_\mu^2 c_d\rho_1^2 d\|U_1\|^2 \tag{4.101}$$

where ρ_1 is a real-valued random variable, independent of the initial condition $(z_0, p_0, q_0 1, \mathbf{\Sigma}_0, r_0)$, and with finite moments. Therefore, combining (4.100) and (4.101), we get

$$\begin{aligned}\lambda_1\left(\mathbf{\Sigma}_1\right) &= \frac{\lambda_1(\tilde{\mathbf{\Sigma}}_1)}{\lambda_d(\tilde{\mathbf{\Sigma}}_1)} \\ &\leqslant \lambda_1\left(\mathbf{\Sigma}_0\right) \times \frac{1 - c_1 - c_\mu + c_\mu\max_{k=1,\dots,k_1}[\tilde{\mathrm{M}}_1^\mu]_{kk} + c_\mu^2 c_d\rho_1^2\|U_1\|^2 + \varepsilon_1(c_\mu)\|U_1\|^2\lambda_1\left(\mathbf{\Sigma}_0\right)}{1 - c_1 - c_\mu + c_\mu(1-\sqrt{c_\mu})^{-1}\min_{k=k_d,\dots,d}[\tilde{\mathrm{M}}_1^\mu]_{kk} - c_\mu^2 c_d\rho_1^2 d\|U_1\|^2} \\ &\qquad + 2(1-c_1-c_\mu)^{-1}c_1 r_0^{-1}\|q_0\|^2 \\ &\coloneqq \lambda_1\left(\mathbf{\Sigma}_0\right) \times \frac{1 - c_1 - c_\mu + c_\mu\max_{k=1,\dots,k_1}[\tilde{\mathrm{M}}_1^\mu]_{kk} + \varepsilon_{\mathrm{num}}(c_\mu) \times (1+\rho_1^2)\|U_1\|^2}{1 - c_1 - c_\mu + c_\mu(1-\sqrt{c_\mu})^{-1}\min_{k=k_d,\dots,d}[\tilde{\mathrm{M}}_1^\mu]_{kk} - \varepsilon_{\mathrm{den}}(c_\mu) \times \rho_1^2\|U_1\|^2} \\ &\qquad + 2(1-c_1-c_\mu)^{-1}c_1 r_0^{-1}\|q_0\|^2\end{aligned} \tag{4.102}$$

where $\varepsilon_{\mathrm{num}}(c_\mu)/c_\mu$ and $\varepsilon_{\mathrm{den}}(c_\mu)/c_\mu$ tend to 0 when c_μ to 0. Thus, by a first-order Taylor expansion of (4.102) when c_μ to 0, there exists $\varepsilon_{\mathrm{frac}}(c_\mu) > 0$ such that $\varepsilon_{\mathrm{frac}}(c_\mu)/c_\mu$ tends to 0 when c_μ to 0, such that

$$\begin{aligned}\lambda_1\left(\mathbf{\Sigma}_1\right) \leqslant \lambda_1\left(\mathbf{\Sigma}_0\right) \times &\left(1 + c_\mu\max_{k=1,\dots,k_1}[\tilde{\mathrm{M}}_1^\mu]_{kk} - c_\mu\min_{k=k_d,\dots,d}[\tilde{\mathrm{M}}_1^\mu]_{kk} + \varepsilon_{\mathrm{frac}}(c_\mu) \times (1+\rho_1^2)\|U_1\|^2\right) \\ &+ 2(1-c_1-c_\mu)^{-1}c_1 r_0^{-1}\|q_0\|^2 \ .\end{aligned} \tag{4.103}$$

□

We can now state the following proposition, which gives sufficient conditions on the learning rates c_1 and c_μ, as well as on the initial condition of the Markov chain (4.17), for the expected largest eigenvalue of the normalized covariance matrix to decrease after one step.

Proposition 4.15. Consider hyperparameters of CMA-ES satisfying **H**1, a spherical objective function f satisfying **F**2 and the normalization function $R(\cdot) = \lambda_{\min}(\cdot)$ satisfying **R**2. Suppose moreover that the sampling distribution ν_U^d is standard normal (**N**5) and that the weights $\mathbf{w}_c$ satisfy **W**1. Consider the normalized Markov chain $\Theta = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) with initial state $(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \in \mathbb{R}^{3d} \times \mathcal{S}_{++}^d \times (0, +\infty)$.
Assume that $c_1 \in [0, 1 - c_\mu]$ is such that c_1/c_μ is sufficiently small. Then there exist a map $\varepsilon_\mu \colon (0,1) \to \mathbb{R}_+$ such that $\varepsilon_\mu(c_\mu)/c_\mu$ tends to 0 when c_μ goes to 0, and constants $M_{\mathbf{\Sigma}} > 0$, $M_y > 0$ and $\delta > 0$, such that for any initial condition $(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \in \mathsf{X} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$ such that $\|\mathbf{\Sigma}_0^{1/2} z_0\|_\infty \leqslant M_y \lambda_1(\mathbf{\Sigma}_0)$, $\lambda_1(\mathbf{\Sigma}_0) > M_{\mathbf{\Sigma}}$, we have that

$$\mathbb{E}_0[\lambda_1(\mathbf{\Sigma}_1)] \leqslant \lambda_1(\mathbf{\Sigma}_0) \times (1 - \delta c_\mu + \varepsilon_\mu(c_\mu)) + 2c_1(1-c_\mu)^{-1} r_0^{-1} \|q_0\|^2 \ . \tag{4.104}$$

Therefore, when $c_\mu > 0$ is sufficiently small and (4.104) holds, we obtain

$$\mathbb{E}_0[\lambda_1(\mathbf{\Sigma}_1)] \leqslant (1 - \delta c_\mu)\lambda_1(\mathbf{\Sigma}_0) + 3c_1 r_0^{-1} \|q_0\|^2 \ . \tag{4.105}$$

In particular, when $M_q > 0$ and $c_\mu > 0$ are sufficiently small and $\|r_0^{-1/2} \mathbf{\Sigma}_0^{-1/2} q_0\|_\infty \leqslant M_q$, we have

$$\mathbb{E}_0[\lambda_1(\mathbf{\Sigma}_1)] \leqslant \lambda_1(\mathbf{\Sigma}_0) \times \left(1 - \frac{\delta c_\mu}{2}\right) \ . \tag{4.106}$$

Proof. By Propositions 4.13 and 4.14 (with $m = (1 - \sqrt{c_\mu})^{-1}$), if $\lambda_{k_1}(\mathbf{\Sigma}_0)/\lambda_{k_d}(\mathbf{\Sigma}_0)$ is larger than a constant $(1 - \sqrt{c_\mu})^2 M_{\mathbf{\Sigma}} > 0$, we have that there exists $\delta > 0$ such that

$$\mathbb{E}\left[\max_{k=1,\dots,k_1} [\tilde{\mathrm{M}}_1^\mu]_{kk} - \min_{k=k_d,\dots,d} [\tilde{\mathrm{M}}_1^\mu]_{kk}\right] \leqslant -\delta$$

where we adopt the notations k_1 and k_d from Proposition 4.14. Plus, by definition of k_1 and k_d, the conditions $\lambda_1(\mathbf{\Sigma}_0) > M_{\mathbf{\Sigma}}$ and $\|\mathbf{\Sigma}_0^{1/2} z_0\|_\infty \leqslant M_y \lambda_1(\mathbf{\Sigma}_0)$ gives, since ρ_1 and $\|U_1\|$ have finite moments, by taking the expectation in (4.103):

$$\mathbb{E}_0[\lambda_1(\mathbf{\Sigma}_1)] \leqslant \lambda_1(\mathbf{\Sigma}_0) \times (1 - \delta c_\mu + \varepsilon_\mu(c_\mu)) + 2(1 - c_1 - c_\mu)^{-1} c_1 r_0^{-1} \|q_0\|^2 \tag{4.107}$$

where $\varepsilon_\mu(c_\mu)/c_\mu$ tends to 0 when c_μ to 0. This ends the proof of (4.104). By observing that $(1 - c_\mu)^{-1} \leqslant 3/2$ when $c_\mu > 0$ is sufficiently small, we prove (4.105), and by taking $M_q \leqslant \delta/6$, we prove (4.106). □

Consequently, the change of variables under affine transformations given in Proposition 4.6 allows us to extend Proposition 4.15 to ellipsoidal objective functions.

Corollary 4.5. Consider an ellipsoidal objective function f satisfying **F**3 with quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}_{++}^d$ and the normalization function $R(\cdot)$ satisfying **R**3 with $\mathbf{H}$. Suppose moreover that the sampling distribution ν_U^d is standard normal (**N**5) and that the stepsize change function Γ_{d_σ} and the weights $\mathbf{w}_c$ satisfy **Γ**4 and **W**1, respectively. Consider the normalized Markov chain $\Theta = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) with the hyperparameters of CMA-ES satisfying **H**1. Assume in addition that $c_1 \in [0, 1 - c_\mu]$ is such that c_1/c_μ is sufficiently small.
Then there exist a map $\varepsilon_\mu \colon (0,1) \to \mathbb{R}_+$ such that $\varepsilon_\mu(c_\mu)/c_\mu$ tends to 0 when c_μ goes to 0, and

constants $M_{\mathbf{\Sigma}} > 0$, $M_y > 0$ and $\delta > 0$, such that for any initial condition $(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \in \mathsf{X} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$ such that $\|\mathbf{\Sigma}_0^{1/2} z_0\| \leqslant M_y \lambda_1(\mathbf{\Sigma}_0)$, $\lambda_1(\mathbf{\Sigma}_0) > M_{\mathbf{\Sigma}}$, we have that

$$\mathbb{E}_0\left[\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right)\right] \leqslant \lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) \times (1 - \delta c_\mu + \varepsilon_\mu(c_\mu)) + 2c_1(1-c_\mu)^{-1} r_0^{-1} \|\mathbf{H}^{1/2} q_0\|^2 \ . \tag{4.108}$$

Therefore, when $c_1 \geqslant 0$ and $c_\mu > 0$ are sufficiently small, we obtain

$$\mathbb{E}_0\left[\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right)\right] \leqslant (1 - \delta c_\mu)\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) + 3c_1 r_0^{-1} \|\mathbf{H}^{1/2} q_0\|^2 \ . \tag{4.109}$$

In particular, when $M_q > 0$ and $c_\mu > 0$ are sufficiently small and $\|r_0^{-1/2}\mathbf{\Sigma}_0^{-1/2} q_0\|_\infty \leqslant M_q$, we have

$$\mathbb{E}_0\left[\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right)\right] \leqslant \lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) \times \left(1 - \frac{\delta c_\mu}{2}\right) \ . \tag{4.110}$$

Proof. It follows from Proposition 4.15 and Proposition 4.6. Indeed, denote $M'_{\mathbf{\Sigma}} > 0$ and $M'_z > 0$ the constants which satisfy Proposition 4.15, and define $M_{\mathbf{\Sigma}} = M'_{\mathbf{\Sigma}} > 0$ and $M_z = \mathrm{Cond}(\mathbf{H}) M'_z$, so that if $\lambda_1(\mathbf{\Sigma}_0) > M_{\mathbf{\Sigma}}$, then $\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}) \geqslant \lambda_d(\mathbf{H})\lambda_1(\mathbf{\Sigma}_0) = \lambda_1(\mathbf{\Sigma}_0) \geqslant M_{\mathbf{\Sigma}} = M'_{\mathbf{\Sigma}}$, and if $\|\mathbf{\Sigma}_0^{1/2} z_0\| \leqslant M_z \lambda_1(\mathbf{\Sigma}_0)$, then, $\|(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2})^{1/2}\mathbf{H}^{1/2} z_0\| \leqslant \lambda_1(\mathbf{H})\|\mathbf{\Sigma}_0^{1/2} z_0\| \leqslant \lambda_1(\mathbf{H}) M_z \lambda_1(\mathbf{\Sigma}_0) \geqslant \mathrm{Cond}(\mathbf{H}) M_z \lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}) = M'_z \lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2})$. □

3.5 *Bounding the expected (normalized) mean*

This section is devoted to establishing upper bounds on the expected updated mean of the normalized Markov chain (4.17). First, we characterize the expected direction of the mean update and prove that it has the opposite sign of $(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2})^{-1/2}\mathbf{H}^{1/2} z_0$ in each coordinate. Moreover, the update is upper-bounded by a decreasing function in the coordinate of the normalized mean $(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2})^{-1/2}\mathbf{H}^{1/2} z_0$ (roughly speaking the further away the coordinate of the normalized mean $(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2})^{-1/2}\mathbf{H}^{1/2} z_0$ is, the more negative it is). The next proposition holds when we assume $\mathbf{H} = \mathbf{I}_d$.

Proposition 4.16. Consider the normalized Markov chain $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) with initial state $(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \in \mathbb{R}^{3d} \times \mathcal{S}^d_{++} \times (0, +\infty)$. Suppose that the objective function f is spherical, i.e., satisfies **F**2, that the weights $\mathbf{w}_m$ satisfy **W**1, and that the distribution ν_U^d is a standard multivariate normal distribution, i.e., satisfies **N**5. Assume that $\mathbf{\Sigma}_0$ is diagonal. For $j \in \{1, \ldots, d\}$, if $[\mathbf{\Sigma}_0^{-1/2} z_0]_j > 0$, then

$$\mathbb{E}_0\left[\sum_{i=1}^{\mu} w_i^m \left[U_1^{s_1(i)}\right]_j\right] < 0 \ . \tag{4.111}$$

Furthermore, there exists a decreasing function $\omega_{\mathbf{w}_m, \nu_U^1} : \mathbb{R}_{++} \to \mathbb{R}_{--}$, such that for $j \in \{1, \ldots, d\}$, if $[\mathbf{\Sigma}_0^{-1/2} z_0]_j = \|\mathbf{\Sigma}_0^{-1/2} z_0\|_\infty > 0$, then,

$$\mathbb{E}_0\left[\sum_{i=1}^{\mu} w_i^m \left[U_1^{s_1(i)}\right]_j\right] \leqslant \omega_{\mathbf{w}_m, \nu_U^1}([\mathbf{\Sigma}_0^{-1/2} z_0]_j) < 0 \ . \tag{4.112}$$

Proof. We have

$$\mathbb{E}\left[\sum_{i=1}^{\mu} w_i^m \left[U_1^{s_1(i)}\right]_j\right] = \sum_{k=1}^{\lambda} \mathbb{E}\left[w^m_{s_1^{-1}(k)}[U_1^k]_j\right] \tag{4.113}$$

$$= \lambda \times \int w^m_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}[u^1]_j \nu_U^{d\lambda}(\mathrm{d}u^1,\dots,\mathrm{d}u^\lambda) \ . \tag{4.114}$$

Consider $u \in \mathbb{R}^d$, and denote $u^* = -\boldsymbol{\Sigma}_0^{-1/2} z_0$ whose coordinate j is assumed to be negative. Then, using (4.45) and Definition 4.1,

$$F_1(u^*+u) \overset{\text{sel}}{\sim} G_1(u^*+u) := \left\| z_0 + \boldsymbol{\Sigma}_0^{1/2}\left(-\boldsymbol{\Sigma}_0^{-1/2} z_0 + u\right)\right\|^2 = \|\boldsymbol{\Sigma}_0^{1/2} u\|^2$$
$$= \| - \boldsymbol{\Sigma}_0^{1/2} u\|^2 \overset{\text{sel}}{\sim} F_1(u^*-u) \ ,$$

so that $u \in \mathbb{R}^d \mapsto F_1(u)$ is symmetric with respect to $u^* = -\boldsymbol{\Sigma}_0^{-1/2} z_0$. In addition since $G_1(u^1) = \|z_0 + \sqrt{\boldsymbol{\Sigma}_0} u^1\|^2 = \| - \sqrt{\boldsymbol{\Sigma}_0} u^* + \sqrt{\boldsymbol{\Sigma}_0} u^1\|^2 = \|\sqrt{\boldsymbol{\Sigma}_0}(u^1 - u^*)\|^2$ and since $\boldsymbol{\Sigma}_0$ is assumed to be diagonal, then

$$G_1(u^1) = \sum_{k=1}^{d} \lambda_k\left(\boldsymbol{\Sigma}_0\right)\left([u^1]_k - [u^*]_k\right)^2.$$

However $[u^*]_j < 0 < -[u^*]_j$, hence on the interval $[[u^*]_j, -[u^*]_j]$ the maps $[u^1]_j \mapsto G_1(u^1)$ and thus $[u^1]_j \mapsto F_1(u^1)$ are increasing (since they are increasing on $([u^*]_j, +\infty)$).
Consider a set of λ candidate solutions that we gather in a vector $u = (u^1,\dots,u^\lambda) \in \mathbb{R}^{d\lambda}$ such that $[u^1]_j > -[u^*]_j > 0$ and another set of λ candidate solutions gathered in $v = (v^1,\dots,v^\lambda) \in \mathbb{R}^{d\lambda}$ such that $[v^1]_j = -[u^1]_j$ and $[v^i]_k = [u^i]_k$ for $(i,k) \neq (1,j)$ that are ranked according to the function F_1. Then $F_1(v^i) = F_1(u^i)$ for $i \neq 1$ and

$$G_1(v^1) = \sum_{k=1}^{d} \lambda_k\left(\boldsymbol{\Sigma}_0\right)\left([v^1]_k - [u^*]_k\right)^2 = \lambda_j\left(\boldsymbol{\Sigma}_0\right)\left(-[u^1]_j - [u^*]_k\right)^2 + \sum_{k\neq j} \lambda_k\left(\boldsymbol{\Sigma}_0\right)\left([u^1]_k - [u^*]_k\right)^2$$
$$< \lambda_j\left(\boldsymbol{\Sigma}_0\right)\left([u^1]_j - [u^*]_k\right)^2 + \sum_{k\neq j} \lambda_k\left(\boldsymbol{\Sigma}_0\right)\left([u^1]_k - [u^*]_k\right)^2 = G_1(u^1) \ .$$

Hence, the ranking of v^1 among the candidate solutions in v is better than or equal to the ranking of u^1 among the candidate solutions in u and since the weights $i \mapsto w_i^m$ are nonincreasing (by **W**1), the weight associated to v^1 is larger than or equal to the weight associated to u^1 and thus

$$w^m_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}} \leqslant w^m_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(v^k)<F_1(v^1)\}} \ . \tag{4.115}$$

We can now split the following integral into three parts

$$\int w^m_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}[u^1]_j \nu_U^{d\lambda}(\mathrm{d}u) = \int_{[u^1]_j > -[u^*]_j} w^m_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}[u^1]_j \nu_U^{d\lambda}(\mathrm{d}u)$$
$$+ \int_{[v^1]_j < [u^*]_j} w^m_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(v^k)<F_1(v^1)\}}[v^1]_j \nu_U^{d\lambda}(\mathrm{d}v)$$
$$+ \int_{-[u^*]_j \geqslant [u^1]_j \geqslant [u^*]_j} w^m_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}[u^1]_j \nu_U^{d\lambda}(\mathrm{d}u)$$

and by the symmetry of the distribution $\nu_U^{d\lambda}$ and by property (4.115)

$$\int_{[u^1]_j>-[u^*]_j} w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}^m [u^1]_j \nu_U^{d\lambda}(\mathrm{d}u) + \int_{[v^1]_j<[u^*]_j} w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(v^k)<F_1(v^1)\}}^m [v^1]_j \nu_U^{d\lambda}(\mathrm{d}v) \leqslant 0$$

so that

$$\int w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}^m [u^1]_j \nu_U^{d\lambda}(\mathrm{d}u) \leqslant \int_{-[u^*]_j\geqslant[u^1]_j\geqslant[u^*]_j} w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}^m [u^1]_j \nu_U^{d\lambda}(\mathrm{d}u) \ .$$

We can now use this inequality into the expression of the weighted selected steps given in (4.113)

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{\mu} w_i^m \left[U_1^{s_1(i)}\right]_j\right] &= \lambda \times \int w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}^m [u^1]_j \nu_U^{d\lambda}(\mathrm{d}u^1,\ldots,\mathrm{d}u^\lambda) \\
&\leqslant \lambda \int_{-[u^*]_j\geqslant[u^1]_j\geqslant[u^*]_j} w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}^m [u^1]_j \nu_U^{d\lambda}(\mathrm{d}u^1,\ldots,\mathrm{d}u^\lambda) \\
&= \lambda \int_{-[u^*]_j\geqslant[u^1]_j\geqslant[u^*]_j} w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}^m \nu_U^{d\lambda}(\mathrm{d}u) \underbrace{\int_{-[u^*]_j\geqslant[u^1]_j\geqslant[u^*]_j} [u^1]_j \nu_U^{d\lambda}(\mathrm{d}u)}_{=0} \\
&\quad + \lambda \mathrm{Cov}_{u\sim\nu_U^{d\lambda}}\left(w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}^m, [u^1]_j \mathbb{1}\{-[u^*]_j \geqslant [u^1]_j \geqslant [u^*]_j\}\right) \\
&= \lambda \mathrm{Cov}_{u\sim\nu_U^{d\lambda}}\left(w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}^m, [u^1]_j \mathbb{1}\{-[u^*]_j \geqslant [u^1]_j \geqslant [u^*]_j\}\right) \ .
\end{aligned}$$

However, by Theorem 4.6 applied to the random variables $X = [u^1]_j$ following the distribution ν_U^1 restricted to $[-[u^*]_j, [u^*]_j]$, $Y = ([u^i]_k)_{(i,k)\neq(1,j)}$, $g(X) = X$, and $h(X,Y) = w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}^m$, where $h(\cdot, y)$ for every $y \in \mathbb{R}^{d\lambda-1}$ is nonincreasing since $i \mapsto w_i^m$ is nonincreasing and $[u^1]_j \mapsto F_1(u^1)$ is increasing on $[-[u^*]_j, [u^*]_j]$, we have then that:

$$\mathrm{Cov}_{u\sim\nu_U^{d\lambda}}\left(w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_1(u^k)<F_1(u^1)\}}^m, [u^1]_j \mathbb{1}\{-[u^*]_j \geqslant [u^1]_j \geqslant [u^*]_j\}\right) < 0 \ .$$

Let j be the index such that $[\mathbf{\Sigma}_0^{-1/2} z_0]_j = \|\mathbf{\Sigma}_0^{-1/2} z_0\|_\infty$. We define the function $\omega_{\mathbf{w}_m,\nu_U^1}$ as

$$\omega_{\mathbf{w}_m,\nu_U^1}(s) = \lambda \sup\left\{W_{z,\mathbf{\Sigma}}^s \mid (z,\mathbf{\Sigma}) \in \mathsf{K}_s\right\}$$

with

$$W_{z,\mathbf{\Sigma}}^s = \mathrm{Cov}\left(w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_{z,\mathbf{\Sigma}}(u^k)<F_{z,\mathbf{\Sigma}}(u^1)\}}^m, [u^1]_j \mathbb{1}\{s \geqslant [u^1]_j \geqslant -s\}\right)$$

and $\mathsf{K}_s = \left\{(z,\mathbf{\Sigma}) \in \mathbb{R}^d \times \mathcal{S}_{++}^d \mid \lambda_1(\mathbf{\Sigma}) = 1, [\mathbf{\Sigma}^{-1/2} z]_j = \|\mathbf{\Sigma}^{-1/2} z\|_\infty \geqslant s\right\}$. Note that $W_{z,\mathbf{\Sigma}}^s$ is finite since $w_i^m \leqslant 1$ for $i = 1,\ldots,\lambda$ and by integrability of ν_U^d by **N**5. Then $s \mapsto \omega_{\mathbf{w}_m,\nu_U^1}(s)$ is decreasing. Indeed, let $s' > s > 0$. Then, for $(z,\mathbf{\Sigma}) \in \mathsf{K}_{s'}$, by bilinearity of the covariance since $[u^1]_j \mathbb{1}\{-s' \leqslant [u^1]_j \leqslant s'\} - [u^1]_j \mathbb{1}\{-s \leqslant [u^1]_j \leqslant s\} = [u^1]_j \mathbb{1}\{s \leqslant |[u^1]_j| \leqslant s'\}$:

$$W_{z,\mathbf{\Sigma}}^{s'} - W_{z,\mathbf{\Sigma}}^s = \mathrm{Cov}\left(w_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_{z,\mathbf{\Sigma}}(u^k)<F_{z,\mathbf{\Sigma}}(u^1)\}}^m, [u^1]_j \mathbb{1}\{s \leqslant |[u^1]_j| \leqslant s'\}\right)$$

which is negative by applying Theorem 4.6 again, similarly to before. Indeed, $[u^1]_j \mapsto F_1(u^1)$ is increasing on $[-[\boldsymbol{\Sigma}^{-1/2}z]_j, +\infty) \supset [-s', s']$ and thus by composition and by **W**1, the function $[u_1]_j \mapsto w^m_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_{z,\boldsymbol{\Sigma}}(u^k)<F_{z,\boldsymbol{\Sigma}}(u^1)\}}$ decreases on $[-s', s']$ so we can apply Theorem 4.6 to $X = [u_1]_j$ following the distribution ν_U^1 restricted to $[-s', -s] \cup [s, s']$, $Y = ([u_i]_k)_{(i,k)\neq(1,j)}$, $g(X) = X$, and $h(X, Y) = w^m_{1+\sum_{k\geqslant 2}\mathbb{1}\{F_{z,\boldsymbol{\Sigma}}(u^k)<F_{z,\boldsymbol{\Sigma}}(u^1)\}}$.
Then, since $\mathsf{K}_s \supset \mathsf{K}_{s'}$, we have $W^{s'}_{z,\boldsymbol{\Sigma}} < W^s_{z,\boldsymbol{\Sigma}}$ and therefore $\omega_{\mathbf{w}_m,\nu_U^1}$ is decreasing as well, ending the proof. □

Proposition 4.16 can be extended to the case of an ellipsoidal function, as stated below.

Proposition 4.17. Suppose that the objective function f is ellipsoidal and satisfies **F**3 with quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}^d_{++}$, the weights $\mathbf{w}_m$ satisfy **W**1, and that the distribution ν_U^d is a standard multivariate normal distribution, i.e., satisfies **N**5. Assume that $\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}$ is diagonal. For $j \in \{1, \ldots, d\}$, if $[(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})^{-1/2}\mathbf{H}^{1/2}z_0]_j > 0$, then

$$\mathbb{E}_0\left[\sum_{i=1}^{\mu} w_i^m \left[\hat{U}_1^{s_1(i)}\right]_j\right] < 0 \tag{4.116}$$

where $\hat{U}_1^i = (\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})^{-1/2}\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0^{1/2}U_1^i$ for $i = 1, \ldots, \lambda$. Furthermore, there exists a decreasing function $\omega_{\mathbf{w}_m,\nu_U^1} \colon \mathbb{R}_{++} \to \mathbb{R}_{--}$, such that for $j \in \{1, \ldots, d\}$, if $[(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})^{-1/2}\mathbf{H}^{1/2}z_0]_j = \|(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})^{-1/2}\mathbf{H}^{1/2}z_0\|_\infty > 0$, then

$$\mathbb{E}_0\left[\sum_{i=1}^{\mu} w_i^m \left[\hat{U}_1^{s_1(i)}\right]_j\right] \leqslant \omega_{\mathbf{w}_m,\nu_U^1}([(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})^{-1/2}\mathbf{H}^{1/2}z_0]_j) < 0 \ . \tag{4.117}$$

Proof. We simply apply Proposition 4.16 and Proposition 4.6 to obtain the desired result. □

The next proposition provides upper bounds on the expectation of $\|\mathbf{H}^{1/2}z_t\|^2$ in different initial configurations. Case (i) gives an upper bound that applies to any initial condition. Case (ii) deduces a decrease for particular initial conditions for which the expected inverse stepsize change is less than 1. The last case (iii) gives the initial conditions for a decrease in one or two steps which requires, in particular, to bound the initial normalized path on the rank-one update.

Statements (i)–(iii) are used later to derive a drift condition for the geometric ergodicity of the normalized Markov chain Θ underlying CMA-ES. However, they require conditions on the initial state of the chain that are more often satisfied for some hyperparameter settings. For instance, the assumption $(1 - c_\sigma)\mathbb{E}\|p_0\| \leqslant \delta_p\mathbb{E}\|\nu_U^d\|$ in (ii) always holds without cumulation on the stepsize, i.e., when $c_\sigma = 1$. Likewise in (iii), we need $r_0^{-1/2}(1 - c_c)\|q_0\| \leqslant \bar{q}\sqrt{\lambda_1(\boldsymbol{\Sigma}_0)}$ which is always true only when $c_c = 1$.

Proposition 4.18. Consider that the hyperparameters of CMA-ES satisfy **H**1. Suppose that the objective function f is ellipsoidal (satisfying **F**3) with quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}^d_{++}$, that the normalization function R satisfies **R**3 with the matrix $\mathbf{H}$, that the sampling distribution ν_U^d is multivariate normal (satisfies **N**5), that the weights $\mathbf{w}_m$ and $\mathbf{w}_c$ satisfy **W**1, and that the stepsize change Γ_{d_σ} satisfy **Γ**1, **Γ**2 and **Γ**4.

(i) Let $\epsilon > 0$ with $\epsilon < 1$. Let $d_\sigma > \bar{d}_\sigma$ (large enough) to satisfy **Γ**2. There exists $\kappa > 0$ such

that for all $c_1, c_\mu, d_\sigma > \bar{d_\sigma}$ such that $c_1 + c_\mu \leqslant 1 - \varepsilon$ and $d_\sigma \geqslant \varepsilon$, for all c_m, c_c, c_σ, for all initial conditions $\theta_0 = (z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \in \mathsf{X} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$, the Markov chain $\Theta = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) satisfies

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_1\|^2\right] \leqslant \frac{1}{1-c_1-c_\mu}\left(1+\frac{2}{d_\sigma}\right)^2 \times \|\mathbf{H}^{1/2}z_0\|^2 + \kappa c_m^2 \lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) \ . \tag{4.118}$$

(ii) Let $M_y > 0$. Suppose moreover that $\mathbf{\Gamma}3$ holds for $k = 4$ and some values of $\delta_\sigma > 0$ and $\delta_p > 0$ and $\bar{d_\sigma}$. Let $d_\sigma \geqslant \bar{d_\sigma}$. Assume that $\mu \leqslant \lambda/2$ and that μ_{eff} is sufficiently large to satisfy (4.71), that c_1, c_μ, $c_m^{3/2}$ are sufficiently smaller than d_σ^{-1}, and that d_σ^{-1} is sufficiently smaller than c_m. Then, for all (random) initial conditions $(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \in \mathsf{X} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$ such that p_0 has finite moments (i.e., $\mathbb{E}\|p_0\|^n < +\infty$ for all $n \in \mathbb{N}$), $\|\mathbf{\Sigma}_0^{1/2} z_0\|_\infty \geqslant M_y \lambda_1(\mathbf{\Sigma}_0)$, and either $(1 - c_\sigma)\mathbb{E}\|p_0\| \leqslant \delta_p \mathbb{E}\|\nu_U^d\|$ or $\mathbb{E}\|p_1\| \geqslant (1 + \delta_p)\mathbb{E}\|\nu_U^d\|$, we have that the Markov chain $\Theta = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) satisfies

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_1\|^2\right] \leqslant \left(1 - \frac{\delta_\sigma}{2d_\sigma}\right) \times \|\mathbf{H}^{1/2}z_0\|^2 \ . \tag{4.119}$$

(iii) Let $M_y > 0$ and $\bar{q} > 0$. Suppose moreover that $\mathbf{\Gamma}3$ holds for $k = 4$ and some values of $\delta_\sigma > 8$ and $\delta_p > 0$. Assume that $\mu \leqslant \lambda/2$ and that μ_{eff} is sufficiently large to have (4.71), that $c_1 d_\sigma$, $c_\mu d_\sigma$, $c_m^{3/2} d_\sigma$ are sufficiently small, that $c_m d_\sigma$ is sufficiently large, and that $1 - c_\sigma \leqslant \delta_p(1 + \delta_p)^{-1}$. Then, for all initial conditions $(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \in \mathsf{X} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$ such that p_0 has finite moments, $\|\mathbf{\Sigma}_0^{1/2} z_0\|_\infty \geqslant M_y \lambda_1(\mathbf{\Sigma}_0)$ and $r_0^{-1/2}(1 - c_c)\|q_0\| \leqslant \bar{q}\sqrt{\lambda_1(\mathbf{\Sigma}_0)}$, the Markov chain $\Theta = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) satisfies either

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_1\|^2\right] \leqslant \left(1 - \frac{\delta_\sigma}{2d_\sigma}\right) \times \|\mathbf{H}^{1/2}z_0\|^2 \tag{4.120}$$

or

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_2\|^2\right] \leqslant \left(1 - \frac{\delta_\sigma - 8}{8d_\sigma}\right) \times \|\mathbf{H}^{1/2}z_0\|^2 + \kappa c_m^2 \lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) \ . \tag{4.121}$$

Proof. Without loss of generality, assume that $\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}$ is diagonal (otherwise we use the spectral theorem to place into an appropriate coordinate system) with decreasingly ordered diagonal elements, and that $[\mathbf{H}^{1/2}z_0]_j \geqslant 0$ for $j = 1, \ldots, d$. By (4.17), we have

$$\begin{aligned}\mathbf{H}^{1/2}z_1 &= r_1^{-1/2}\Gamma_{d_\sigma}(p_1)^{-1} \times \left(\mathbf{H}^{1/2}z_0 + c_m \mathbf{H}^{1/2}\sqrt{\mathbf{\Sigma}_0}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right) \\ &= r_1^{-1/2}\Gamma_{d_\sigma}(p_1)^{-1} \times \left(\mathbf{H}^{1/2}z_0 + c_m \sqrt{\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}}\sum_{i=1}^{\mu} w_i^m \hat{U}_1^{s_1(i)}\right)\end{aligned}$$

where $\hat{U}_1^i = (\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2})^{-1/2}\mathbf{H}^{1/2}\mathbf{\Sigma}_0^{1/2}U_1^i$ follows the distribution ν_U^d by Proposition 4.6.

Then, as $r_1 \geqslant 1 - c_1 - c_\mu$ by Proposition 4.3, we obtain

$$\|\mathbf{H}^{1/2} z_1\| \leqslant (1 - c_1 - c_\mu)^{-1/2} \Gamma_{d_\sigma}(p_1)^{-1} \times \left\| \mathbf{H}^{1/2} z_0 + c_m \sqrt{\mathbf{H}^{1/2} \boldsymbol{\Sigma}_0 \mathbf{H}^{1/2}} \sum_{i=1}^{\mu} w_i^m \hat{U}_1^{s_1(i)} \right\|$$

and thus

$$\|\mathbf{H}^{1/2} z_1\|^2 \leqslant (1 - c_1 - c_\mu)^{-1} \Gamma_{d_\sigma}(p_1)^{-2} \times \left[\|\mathbf{H}^{1/2} z_0\|^2 + 2 c_m \sum_{j=1}^{d} \sqrt{\lambda_j \left(\mathbf{H}^{1/2} \boldsymbol{\Sigma}_0 \mathbf{H}^{1/2}\right)} [\mathbf{H}^{1/2} z_0]_j \left[\sum_{i=1}^{\mu} w_i^m \hat{U}_1^{s_1(i)} \right]_j + c_m^2 \lambda_1 \left(\mathbf{H}^{1/2} \boldsymbol{\Sigma}_0 \mathbf{H}^{1/2} \right) V_1 \right] \tag{4.122}$$

where $V_1 = \max_{i=1,\dots,\lambda} \|\hat{U}_1^i\|^2$. We prove now (i). By Proposition 4.17, we have that

$$\mathbb{E}_0 \left[\sum_{i=1}^{\mu} w_i^m \hat{U}_1^{s_1(i)} \right]_j \leqslant 0$$

and since the stepsize change satisfies $\boldsymbol{\Gamma}2$, we have $\Gamma_{d_\sigma}(p_1)^{-2} \leqslant (1 + 2 d_\sigma^{-1})^2$ for d_σ large enough. By using these inequalities in (4.122), we obtain that the conditional expectation of $\|\mathbf{H}^{1/2} z_1\|^2$ is upper-bounded by

$$\mathbb{E}_0 \left[\|\mathbf{H}^{1/2} z_1\|^2 \right] \leqslant (1 - c_1 - c_\mu)^{-1} \left(1 + \frac{2}{d_\sigma} \right)^2 \times \left(\|\mathbf{H}^{1/2} z_0\|^2 + c_m^2 \lambda_1 \left(\mathbf{H}^{1/2} \boldsymbol{\Sigma}_0 \mathbf{H}^{1/2} \right) \mathbb{E}\left[V_1 \right] \right) \quad ,$$

which proves (i) by taking $\kappa = \varepsilon^{-1}(1 + 2\varepsilon^{-1})^2 \mathbb{E} V_1 \geqslant (1 - c_1 - c_\mu)^{-1}(1 + 2 d_\sigma^{-1})^2 \mathbb{E} V_1 > 0$. Next, we prove (ii). We assume in a first time that $\mathbf{H} = \mathbf{I}_d$ so that $\boldsymbol{\Sigma}_0$ is diagonal. Define the following quantity:

$$\delta_y \coloneqq -\omega_{\mathbf{w}_m, \nu_U^1}(M_y) > 0 \tag{4.123}$$

where $\omega_{\mathbf{w}_m, \nu_U^1}$ is a function satisfying (4.117). Assume that $\|\boldsymbol{\Sigma}_0^{1/2} z_0\|_\infty \geqslant M_y \lambda_1 (\boldsymbol{\Sigma}_0)$—and thus since $\boldsymbol{\Sigma}_0$ is diagonal $\|\boldsymbol{\Sigma}_0^{-1/2} z_0\|_\infty \geqslant \|\lambda_1 (\boldsymbol{\Sigma}_0)^{-1} \boldsymbol{\Sigma}_0^{1/2} z_0\|_\infty \geqslant M_y$—and either $(1 - c_\sigma)\|p_0\| \leqslant \delta_p \mathbb{E}\|\nu_U^d\|$ or $\mathbb{E}_0\|p_1\| \geqslant (1 + \delta_p)\mathbb{E}\|\nu_U^d\|$. Then, let $j^* \in \{1, \dots, d\}$ such that

$$[\boldsymbol{\Sigma}_0^{1/2} z_0]_{j^*} = \|\boldsymbol{\Sigma}_0^{1/2} z_0\|_\infty \geqslant M_y \lambda_1 (\boldsymbol{\Sigma}_0) \tag{4.124}$$

and thus

$$[\boldsymbol{\Sigma}_0^{-1/2} z_0]_{j^*} = \frac{[\boldsymbol{\Sigma}_0^{1/2} z_0]_{j^*}}{\lambda_{j^*} (\boldsymbol{\Sigma}_0)} \geqslant \frac{[\boldsymbol{\Sigma}_0^{1/2} z_0]_{j^*}}{\lambda_1 (\boldsymbol{\Sigma}_0)} \geqslant M_y \quad . \tag{4.125}$$

By Proposition 4.16, we have, by decreaseness of $\omega_{\mathbf{w}_m, \nu_U^1}$:

$$\mathbb{E}_0 \left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)} \right]_{j^*} \leqslant \omega_{\mathbf{w}_m, \nu_U^1}([\boldsymbol{\Sigma}_0^{-1/2} z_0]_{j^*}) \leqslant -\delta_y \quad .$$

Suppose then that $\boldsymbol{\Gamma}3$ holds for some $\delta_p > 0$ and $\delta_\sigma > 0$, and let $M_\sigma > 0$ which satisfies Proposition 4.10(i). We recall that we assume without loss of generality that $[z_0]_{j^*} > 0$. Suppose first that (a) $\|\boldsymbol{\Sigma}_0^{1/2} z_0\| \leqslant M_\sigma \lambda_1 (\boldsymbol{\Sigma}_0)$. Then, for $j = 1, \dots, d$, by the matrix inequality $\boldsymbol{\Sigma}_0 \preceq \lambda_1 (\boldsymbol{\Sigma}_0) \mathbf{I}_d$, $[z_0]_j^2 \leqslant \|z_0\|^2 = \|\boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{\Sigma}_0^{1/2} z_0\|^2 \leqslant \lambda_1 (\boldsymbol{\Sigma}_0)^{-1} \|\boldsymbol{\Sigma}_0^{1/2} z_0\|^2 \leqslant M_\sigma^2 \lambda_1 (\boldsymbol{\Sigma}_0)$.

Thus, by sum, $\lambda_1(\mathbf{\Sigma}_0) \geqslant M_\sigma^{-2} d^{-1} \|z_0\|^2$. However, if $c_m > 0$ is sufficiently small, then, by (4.122) with $\mathbf{H}^{1/2} = \mathbf{I}_d$ and (4.123), and using $\mathbf{\Gamma}2$, we obtain, since $\mathbb{E}_0[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}]_j \leqslant 0$ for $j = 1, \ldots, d$ and in addition, using $\lambda_1(\mathbf{\Sigma}_0) \leqslant M_z^{-2}\|z_0\|^2$ since $M_z \lambda_1(\mathbf{\Sigma}_0) \leqslant \|\mathbf{\Sigma}_0^{1/2} z_0\| \leqslant \sqrt{\lambda_1(\mathbf{\Sigma}_0)}\|z_0\|$ and $\sqrt{\lambda_{j^*}(\mathbf{\Sigma}_0)}[z_0]_{j^*} \geqslant M_z \lambda_1(\mathbf{\Sigma}_0)$ by (4.124):

$$\mathbb{E}_0\left[\|z_1\|^2\right] \leqslant (1 - c_1 - c_\mu)^{-1} \left(1 + \frac{2}{d_\sigma}\right)^2 \tag{4.126}$$

$$\times \left(\|z_0\|^2 + 2c_m \underbrace{\sqrt{\lambda_{j^*}(\mathbf{\Sigma}_0)}[z_0]_{j^*}}_{\geqslant M_y \lambda_1(\mathbf{\Sigma}_0) \geqslant M_y M_\sigma^{-2} d^{-1} \|z_0\|^2} \underbrace{\mathbb{E}_0\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_{j^*}}_{\leqslant -\delta_y} + c_m^2 \underbrace{\lambda_1(\mathbf{\Sigma}_0)}_{\leqslant M_y^{-2}\|z_0\|^2} \mathbb{E}_0[V_1] \right) \tag{4.127}$$

$$\leqslant (1 - c_1 - c_\mu)^{-1} \left(1 + \frac{2}{d_\sigma}\right)^2 \times \left(\|z_0\|^2 - 2\delta_y M_\sigma^{-2} d^{-1} M_y c_m \|z_0\|^2 + c_m^2 M_y^{-2} \mathbb{E}_0[V_1] \|z_0\|^2 \right) \tag{4.128}$$

$$\leqslant \underbrace{\frac{1}{1 - c_1 - c_\mu}}_{>1} \underbrace{\left(1 + \frac{2}{d_\sigma}\right)^2}_{>1} \underbrace{(1 - \delta_y M_\sigma^{-2} d^{-1} M_y c_m)}_{<1} \|z_0\|^2 \ . \tag{4.129}$$

In particular, for the last inequality we choose $c_m > 0$ to be sufficiently small to have that

$$c_m^2 \lambda_1(\mathbf{\Sigma}_0)\, \mathbb{E}_0[V_1] \leqslant c_m^2 M_y^{-2} \|z_0\|^2 \mathbb{E}_0[V_1] \leqslant \delta_y M_\sigma^{-2} d^{-1} M_y c_m \|z_0\|^2 \ .$$

However, if c_1, c_μ and d_σ^{-1} are sufficiently smaller than c_m, then we get

$$\mathbb{E}_0\left[\|z_1\|^2\right] \leqslant \left(1 - \frac{1}{2}\delta_y M_\sigma^{-1} M_y c_m\right) \times \|z_0\|^2 \ , \tag{4.130}$$

and taking d_σ^{-1} sufficiently smaller than c_m we can set $1 - \frac{1}{2}\delta_y M_\sigma^{-1} M_y c_m \leqslant 1 - \frac{\delta_\sigma}{2d_\sigma}$ (where δ_σ satisfies $\mathbf{\Gamma}3$). We hence obtain

$$\mathbb{E}_0\left[\|z_1\|^2\right] \leqslant \left(1 - \frac{\delta_\sigma}{2d_\sigma}\right) \times \|z_0\|^2 \ .$$

Now, suppose that (b) $\|\mathbf{\Sigma}_0^{1/2} z_0\| > M_\sigma \lambda_1(\mathbf{\Sigma}_0)$. Then

$$[\mathbf{\Sigma}_0^{-1/2} z_0]_{j^*} = \frac{[\mathbf{\Sigma}_0^{1/2} z_0]_{j^*}}{\lambda_{j^*}(\mathbf{\Sigma}_0)} \geqslant \frac{\|\mathbf{\Sigma}_0^{1/2} z_0\|}{\lambda_1(\mathbf{\Sigma}_0)} > M_\sigma \ .$$

If $(1 - c_\sigma)\|p_0\| \leqslant \delta_p \mathbb{E}\|\nu_U^d\|$, then by Proposition 4.10(i) (applied with $\alpha = 0$), we have $\mathbb{E}_0[\|p_1\|] \geqslant (1 + \delta_p)\mathbb{E}\|\nu_U^d\|$, hence if $d_\sigma \geqslant 1$ is sufficiently large, then by $\mathbf{\Gamma}3$ (that we have assumed to hold for $k = 4$) we have $\mathbb{E}_0[\Gamma_{d_\sigma}(p_1)^{-4}] \leqslant 1 - 4d_\sigma^{-1}\delta_\sigma$.
However, since $\|\mathbf{\Sigma}_0^{1/2} z_0\|_\infty \geqslant M_y \lambda_1(\mathbf{\Sigma}_0)$, then

$$\lambda_1(\mathbf{\Sigma}_0) \leqslant 1/M_y \|\mathbf{\Sigma}_0^{1/2} z_0\|_\infty \leqslant 1/M_y \sqrt{\lambda_1(\mathbf{\Sigma}_0)}\|z_0\|_\infty \leqslant 1/M_y \sqrt{\lambda_1(\mathbf{\Sigma}_0)}\|z_0\| \tag{4.131}$$

and thus $\sqrt{\lambda_1(\boldsymbol{\Sigma}_0)} \leqslant 1/M_y\|z_0\|$. Therefore, for $\varepsilon \in (0,1)$, the following hold:

$$\begin{aligned}
&\varepsilon\|z_0\|^2 + 2c_m \sum_{j=1}^{d} \sqrt{\lambda_j(\boldsymbol{\Sigma}_0)}[z_0]_j \left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_j < 0 \\
&\qquad \Rightarrow \frac{\varepsilon\|z_0\|^2}{\|z_0\|} \leqslant 2c_m \sum_{j=1}^{d} \sqrt{\lambda_1(\boldsymbol{\Sigma}_0)} \left|\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_j\right| \\
&\qquad\qquad \Rightarrow \frac{\varepsilon\|z_0\|}{\sqrt{\lambda_1(\boldsymbol{\Sigma}_0)}} \leqslant 2c_m \sum_{j=1}^{d} \|U_1\|_\infty \Rightarrow \varepsilon M_z < 2c_m d\|U_1\|_\infty \quad (4.132)
\end{aligned}$$

and, if we denote Z the event

$$\varepsilon\|z_0\|^2 + 2c_m \sum_{j=1}^{d} \sqrt{\lambda_j(\boldsymbol{\Sigma}_0)}[z_0]_j \left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_j < 0$$

we have then $\mathbb{P}[\mathsf{Z}] \leqslant \mathbb{P}[\varepsilon M_y < 2c_m d\|U_1\|_\infty]$. Moreover, since the random variables $[U^i]_j$ follow ν_U^1 for $j = 1, \ldots, d$ and $i = 1, \ldots, \lambda$:

$$\mathbb{P}[\varepsilon M_y < 2c_m d\|U_1\|_\infty] = \mathbb{P}\left[\cup_{i=1}^{\lambda} \cup_{j=1}^{d} |[U_1^i]_j| > \frac{\varepsilon M_z}{2c_m d}\right] \leqslant d\lambda \times \mathbb{P}\left[|[U_1^1]_1| > \frac{\varepsilon M_z}{2c_m d}\right]$$

and by Markov's inequality, and since ν_U^1 is the standard normal distribution:

$$\mathbb{P}\left[|[U_1^1]_1| > \frac{\varepsilon M_z}{2c_m d}\right] \leqslant \frac{2c_m d\mathbb{E}|[U_1^1]_1|}{\varepsilon M_z} = \frac{4c_m d}{\varepsilon M_z\sqrt{2\pi}} \ .$$

All in all, $\mathbb{P}[\mathsf{Z}] \leqslant \frac{4d^2\lambda}{\varepsilon M_z\sqrt{2\pi}} c_m$ tends to 0 when c_m to 0. Therefore, if $c_m > 0$ is sufficiently small, using $\boldsymbol{\Gamma}2$, we have:

$$\begin{aligned}
&\mathbb{E}_0\left[\Gamma_{d_\sigma}^{-2}(p_1) \times \left(\varepsilon\|z_0\|^2 + 2c_m \sum_{j=1}^{d} \sqrt{\lambda_j(\boldsymbol{\Sigma}_0)}[z_0]_j \left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_j\right)\right] \\
&\qquad \leqslant (1 + 2/d_\sigma)^2 \times \mathbb{E}_0\left[\varepsilon\|z_0\|^2 + 2c_m \sum_{j=1}^{d} \sqrt{\lambda_j(\boldsymbol{\Sigma}_0)}[z_0]_j \left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_j\right]^+ \quad (4.133)
\end{aligned}$$

In addition, we have

$$\begin{aligned}
&\mathbb{E}_0\left[\varepsilon\|z_0\|^2 + 2c_m \sum_{j=1}^{d} \sqrt{\lambda_j(\boldsymbol{\Sigma}_0)}[z_0]_j \left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_j\right]^+ \\
&\qquad \leqslant \mathbb{E}_0\left[\varepsilon\|z_0\|^2 + 2c_m \sum_{j=1}^{d} \sqrt{\lambda_j(\boldsymbol{\Sigma}_0)}[z_0]_j \left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_j \mathbb{1}_{\mathsf{Z}}\right] \\
&\leqslant \varepsilon\|z_0\|^2 + 2c_m \sum_{j=1}^{d} \sqrt{\lambda_j(\boldsymbol{\Sigma}_0)}[z_0]_j \left(\underbrace{\mathbb{E}_0\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_j}_{\leqslant 0} + \underbrace{\mathbb{E}_0\left|\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_j\right| \mathbb{1}_{\mathsf{Z}}}_{\leqslant \sqrt{\mathbb{E}_0\left|\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_j\right|^2 \mathbb{P}[\mathsf{Z}]}}\right)
\end{aligned}$$

and since $|\sqrt{\lambda_j(\mathbf{\Sigma}_0)}[z_0]_j| \leqslant \sqrt{\lambda_1(\mathbf{\Sigma}_0)}\|z_0\| \leqslant \|z_0\|^2/M_z$, see (4.131), and $\mathbb{E}_0\left|\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]_j\right|^2 \leqslant \mathbb{E}_0[V_1]$ by definition of V_1, we can upper bound the RHS of the previous equation by

$$A = (1+2/d_\sigma)^2 \times \left(\varepsilon\|z_0\|^2 + \left(\frac{4d^2\lambda}{\varepsilon M_y\sqrt{2\pi}}c_m\right)^{1/2} \times 2c_m\frac{1}{M_y}\|z_0\|^2\sqrt{\mathbb{E}_0[V_1]}\right) \quad . \tag{4.134}$$

Moreover, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}_0\left[\Gamma_{d_\sigma}(p_1)^{-2} \times \left((1-\varepsilon)\|z_0\|^2 + c_m^2\lambda_1(\mathbf{\Sigma}_0)\, V_1\right)\right] \\ \leqslant \sqrt{\mathbb{E}_0\left[\Gamma_{d_\sigma}(p_1)^{-4}\right]} \times \left((1-\varepsilon)\|z_0\|^2 + c_m^2\lambda_1\left(\mathbf{\Sigma}_0\right)\sqrt{\mathbb{E}_0 V_1^2}\right)\end{aligned} \tag{4.135}$$

In addition, using assumption $\mathbf{\Gamma}3$ we can bound $\mathbb{E}_0\left[\Gamma_{d_\sigma}(p_1)^{-4}\right]$ by $1-4\frac{\delta_\sigma}{d_\sigma}$ and thus bound the RHS of the previous equation by

$$\begin{aligned}B = \sqrt{1-4\frac{\delta_\sigma}{d_\sigma}} \times \left((1-\varepsilon)\|z_0\|^2 + c_m^2\lambda_1\left(\mathbf{\Sigma}_0\right)\sqrt{\mathbb{E}_0 V_1^2}\right) \\ \leqslant \left(1-2\frac{\delta_\sigma}{d_\sigma}\right) \times \left((1-\varepsilon)\|z_0\|^2 + c_m^2 M_y^{-2}\|z_0\|^2\sqrt{\mathbb{E}_0 V_1^2}\right) \quad .\end{aligned} \tag{4.136}$$

where we have bounded $\sqrt{1-4\frac{\delta_\sigma}{d_\sigma}}$ by $\left(1-2\frac{\delta_\sigma}{d_\sigma}\right)$ which is true whenever $\frac{\delta_\sigma}{d_\sigma} \leqslant 1/4$. From (4.122) (with $\mathbf{H} = \mathbf{I}_d$), we can bound $\mathbb{E}_0\|z_1\|^2$ by the sum of (4.134) and (4.136). Set $\varepsilon > 0$ such that $\varepsilon \times (1+2/d_\sigma)^2 + (1-\varepsilon) \times (1-\delta_\sigma/2d_\sigma) \leqslant 1-\delta_\sigma/d_\sigma$. Moreover, we choose $c_m^{3/2}$ to be small enough compared to $1/d_\sigma$ so that in (4.134) we have $(\frac{4d^2\lambda}{\varepsilon M_y\sqrt{2\pi}}c_m)^{1/2} \times 2c_m\frac{1}{M_y}\mathbb{E}_0[V_1]^{1/2} \leqslant (1+2/d_\sigma)^{-2}\delta_\sigma/2d_\sigma$ and in (4.136) $c_m^2 M_y^{-2}\sqrt{\mathbb{E}_0 V_1^2} \leqslant \varepsilon$. By sum:

$$\mathbb{E}_0[\|z_1\|^2] \leqslant (1-c_1-c_\mu)^{-1} \times (A+B) \leqslant (1-c_1-c_\mu)^{-1}\left(1-\frac{\delta_\sigma}{d_\sigma}\right) \times \|z_0\|^2 \quad .$$

This gives the desired result if c_1 and c_μ are sufficiently smaller than $1/d_\sigma$. In both cases (a) and (b), we have proven (ii). If $\mathbf{H} \neq \mathbf{I}_d$, we apply Proposition 4.6. Denote $(\hat{z}_0, \hat{p}_0, \hat{q}_0, \hat{\mathbf{\Sigma}}_0, \hat{r}_0) = (\mathbf{H}^{1/2}z_0, p_0, \mathbf{H}^{1/2}q_0, \mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}, r_0)$ and $\{\hat{\theta}_t\}_{t\in\mathbb{N}} = \{(\hat{z}_t, \hat{p}_t, \hat{q}_t, \hat{\mathbf{\Sigma}}_t, \hat{r}_t)\}_{t\in\mathbb{N}}$ the Markov chain obeying (4.17) with an objective function $\hat{f}$ satisfying $\mathbf{F}2$ and a normalization function $\hat{R}$ satisfying $\mathbf{R}2$. Assume that $\|\mathbf{\Sigma}_0^{1/2}z_0\|_\infty \geqslant M_y\lambda_1(\mathbf{\Sigma}_0)$. Then,

$$\left\|(\mathbf{H}^{-1/2}\hat{\mathbf{\Sigma}}_0\mathbf{H}^{-1/2})^{1/2}\mathbf{H}^{-1/2}\hat{z}_0\right\|_\infty \geqslant M_y\lambda_1\left(\mathbf{H}^{-1/2}\hat{\mathbf{\Sigma}}_0\mathbf{H}^{-1/2}\right)$$

and since $\mathbf{I}_d \preceq \mathbf{H} \preceq \lambda_1(\mathbf{H})\mathbf{I}_d$, we have

$$\|\hat{\mathbf{\Sigma}}_0^{1/2}\hat{z}_0\|_\infty \geqslant \left\|(\mathbf{H}^{-1/2}\hat{\mathbf{\Sigma}}_0\mathbf{H}^{-1/2})^{1/2}\mathbf{H}^{-1/2}\hat{z}_0\right\|_\infty \geqslant M_y\lambda_1\left(\mathbf{H}^{-1/2}\hat{\mathbf{\Sigma}}_0\mathbf{H}^{-1/2}\right) \geqslant M_y\lambda_1(\mathbf{H})^{-1}\lambda_1(\mathbf{\Sigma}_0) \quad .$$

Moreover, if $(1-c_\sigma)\mathbb{E}\|p_0\| \leqslant \delta_p\mathbb{E}\|\nu_U^d\|$, then since $\hat{p}_0 = p_0$, we have $(1-c_\sigma)\mathbb{E}\|\hat{p}_0\| \leqslant \delta_p\mathbb{E}\|\nu_U^d\|$. Alternatively, if $\mathbb{E}\|p_1\| \geqslant (1+\delta_p)\mathbb{E}\|\nu_U^d\|$, then, since $\|\hat{p}_1\|$ is distributed as $\|p_1\|$

by Proposition 4.6, we have $\mathbb{E}\|\hat{p}_1\| \geqslant (1+\delta_p)\mathbb{E}\|\nu_U^d\|$. Then, by applying (ii) to $\{\hat{\theta}_t\}_{t\in\mathbb{N}}$ with $\hat{M}_y = M_y\lambda_1(\mathbf{H})^{-1}$, we find that, if $c_1 d_\sigma$, $c_\mu d_\sigma$, $c_m^2 d_\sigma$, $(c_m d_\sigma)^{-1}$ are sufficiently small, then

$$\mathbb{E}_0\left[\|\hat{z}_1\|^2\right] \leqslant \left(1 - \frac{\delta_\sigma}{2d_\sigma}\right) \times \|\hat{z}_0\|^2$$

and thus, by Proposition 4.6:

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_1\|^2\right] \leqslant \left(1 - \frac{\delta_\sigma}{2d_\sigma}\right) \times \|\mathbf{H}^{1/2}z_0\|^2 \ .$$

For (iii), let $M_\sigma > 0$ be a positive constant which satisfies Proposition 4.10(ii). Suppose first that $\|\mathbf{\Sigma}_0^{1/2}z_0\| \leqslant M_\sigma\lambda_1(\mathbf{\Sigma}_0)$. Then, as for (ii) and more precisely (4.129), we find that there exists $\delta_y > 0$ such that, if $c_m > 0$ is sufficiently small, then

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_1\|^2\right] \leqslant \frac{1}{1-c_1-c_\mu}\left(1+\frac{2}{d_\sigma}\right)^2 \times \left(\|\mathbf{H}^{1/2}z_0\|^2 - c_m\delta_y M_\sigma^{-2}d^{-1}M_y\|\mathbf{H}^{1/2}z_0\|^2\right) \ .$$

Hence, if c_m is sufficiently larger than $c_1, c_\mu, d_\sigma^{-1}$, then

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_1\|^2\right] \leqslant \left(1 - \frac{\delta_\sigma}{2d_\sigma}\right) \times \|\mathbf{H}^{1/2}z_0\|^2 \ .$$

Now, suppose that $\|\mathbf{\Sigma}_0^{1/2}z_0\| > M_\sigma\lambda_1(\mathbf{\Sigma}_0)$. Assume moreover that

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_1\|^2\right] > \left(1 - \frac{\delta_\sigma}{2d_\sigma}\right) \times \|\mathbf{H}^{1/2}z_0\|^2 \ .$$

Then, by contraposition of Proposition 4.18(ii), we find that $\mathbb{E}_0\|p_1\| < (1+\delta_p)\mathbb{E}\|\nu_U^d\|$ (and also $(1-c_\sigma)\mathbb{E}\|p_0\| > \delta_p\mathbb{E}\|\nu_U^d\|$). If as assumed in the proposition $1 - c_\sigma \leqslant \delta_p(1+\delta_p)^{-1}$, then, $(1-c_\sigma)\mathbb{E}_0\|p_1\| \leqslant \delta_p\mathbb{E}\|\nu_U^d\|$. Therefore, by Proposition 4.10(ii), we have $\mathbb{E}\|p_2\| \geqslant (1+\delta_p)\mathbb{E}\|\nu_U^d\|$.
Moreover, we have, almost surely, that

$$\begin{aligned}
\|\mathbf{H}^{1/2}z_2\|^2 &= r_2^{-1}\Gamma_{d_\sigma}(p_2)^{-2}\left\|\mathbf{H}^{1/2}z_1 + c_m\sqrt{\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}}\sum_{i=1}^{\mu} w_i^m \hat{U}_2^{s_2(i)}\right\|^2 \\
&\leqslant 2(1-c_1-c_\mu)^{-1}\left(1+\frac{2}{d_\sigma}\right)^2 \times \left[\left\|\mathbf{H}^{1/2}z_1\right\|^2 + \left\|c_m\sqrt{\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}}\sum_{i=1}^{\mu} w_i^m\hat{U}_2^{s_2(i)}\right\|^2\right] \\
&\leqslant 2(1-c_1-c_\mu)^{-1}\left(1+\frac{2}{d_\sigma}\right)^2 \times \left[\left\|\mathbf{H}^{1/2}z_1\right\|^2 + c_m^2\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2})\|\hat{U}_2\|_\infty^2\right] \\
&\leqslant 4(1-c_1-c_\mu)^{-2}\left(1+\frac{2}{d_\sigma}\right)^4 \times \left[\left\|\mathbf{H}^{1/2}z_0\right\|^2 + c_m^2\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2})\|\hat{U}_1\|_\infty^2\right] \\
&\quad + 2(1-c_1-c_\mu)^{-1}\left(1+\frac{2}{d_\sigma}\right)^2 c_m^2\|\hat{U}_2\|_\infty^2(1-c_1-c_\mu)^{-1} \\
&\qquad \times \left[\left(1-c_1-c_\mu+(2c_1\mu_{\text{eff}}+c_\mu)d\mu V_1^2\right)\lambda_k(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}) + 2c_1 r_0^{-1}\|\mathbf{H}^{1/2}q_0\|\right] \\
&\leqslant \left[\kappa_z + c_m^2 M_y^{-2}\kappa_{\mathbf{\Sigma}}(\|\hat{U}_1\|_\infty^2 + \|\hat{U}_2\|_\infty^2) + c_m^2\bar{q}^2(\|\hat{U}_1\|_\infty^2 + \|\hat{U}_2\|_\infty^2)\right] \times \|\mathbf{H}^{1/2}z_0\|^2 \ .
\end{aligned}$$

Let $\alpha > 0$. Up to choosing $M_\sigma > 0$ larger, we can assume that (we refer to the proof of Proposition 4.10(ii) and leave the details to the reader)

$$\mathbb{P}_0[\underbrace{\|\mathbf{\Sigma}_1^{1/2} z_1\| \geqslant M_y \lambda_1(\mathbf{\Sigma}_1)}_{=:\mathsf{A}_1}] \geqslant 1-\alpha \ .$$

By Cauchy-Schwarz inequality:

$$\begin{aligned}&\mathbb{E}_0\left[\|\mathbf{H}^{1/2} z_2\|^2 \mathbb{1}\{\mathsf{A}_1^c\}\right] \\ &\leqslant \sqrt{\underbrace{\mathbb{E}_0\left[\left[\kappa_z + M_y^{-2}\kappa_{\mathbf{\Sigma}}(\|\hat{U}_1\|_\infty^2 + \|\hat{U}_2\|_\infty^2) + \bar{q}^2(\|\hat{U}_1\|_\infty^2 + \|\hat{U}_2\|_\infty^2)\right]^2\right]}_{=:\zeta} \underbrace{\mathbb{E}_0[\mathbb{1}\{\mathsf{A}_1^c\}]}_{\leqslant \alpha}} \times \|\mathbf{H}^{1/2} z_0\|^2 \ .\end{aligned}$$

Then, by (ii), since p_1 has finite moments by Proposition 4.11:

$$\begin{aligned}\mathbb{E}_0\left[\|\mathbf{H}^{1/2} z_2\|^2\right] &= \mathbb{E}_0\left[\|\mathbf{H}^{1/2} z_2\|^2 \mathbb{1}_{\mathsf{A}_1}\right] + \mathbb{E}_0\left[\|\mathbf{H}^{1/2} z_2\|^2 \mathbb{1}_{\mathsf{A}_1^c}\right] \\ &\leqslant \left(1 - \frac{\delta_\sigma}{2d_\sigma}\right) \mathbb{E}_0[\|\mathbf{H}^{1/2} z_1\|^2] + \sqrt{\alpha\zeta} \times \|\mathbf{H}^{1/2} z_0\|^2 \ .\end{aligned}$$

Then, by (i) and by Proposition 4.8, there exists $\kappa > 0$ such that

$$\mathbb{E}_0\|\mathbf{H}^{1/2} z_2\|^2 \leqslant \left(1 - \frac{\delta_\sigma}{2d_\sigma}\right)\left(1 + \frac{2}{d_\sigma}\right)^2 \|\mathbf{H}^{1/2} z_0\|^2 + c_m^2 \kappa \lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) + \sqrt{\alpha\zeta} \times \|\mathbf{H}^{1/2} z_0\|^2 \ ,$$

which ends the proof by taking $d_\sigma > 0$ sufficiently large, since $\delta_\sigma > 8$, and $\alpha > 0$ sufficiently small. □

3.6 *Proof of Proposition 4.4*

We have now established various technical bricks to prove the state-dependent drift condition presented in Proposition 4.4. However, the proof relies on two lemmas that depend on the technical components from the previous section, which will be stated and proven in this section. Proposition 4.4 states that given the potential function $V : \mathsf{X} \to \mathbb{R}_{++}$ defined in (4.25) as

$$V(z, p, q, \mathbf{\Sigma}, r) = \|\mathbf{H}^{1/2} z\|^2 + \beta\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}) + \gamma_p\|p\| + \gamma_q\|\mathbf{H}^{1/2} q\|^2 + \gamma_r r$$

where β, γ_p, γ_q, γ_r are positive constants, there exists $\mathsf{K} \subset \mathsf{X}$ and $\varepsilon > 0$ such that for every initial state $\theta_0 = (z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0)$ outside K, the normalized Markov chain associated to CMA-ES satisfies the geometric drift condition

$$\mathbb{E}_0\left[V(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\right] \leqslant (1 - \varepsilon c_\mu) V(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \quad \text{for } t = 1 \text{ or } 2. \tag{D}$$

The compact set K is built using five subregions of the state space $\mathsf{X} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times [1 - c_1 - c_\mu, +\infty)$ (see Proposition 4.3 for why we admit this state space) defined as

$$\mathsf{A}_{z,\mathbf{\Sigma}} = \{(z, p, q, \mathbf{\Sigma}, r) \in \mathsf{X}, \lambda_1(\mathbf{\Sigma}) > M_{\mathbf{\Sigma}} \text{ and } \|\mathbf{\Sigma}^{1/2} z\| \leqslant \lambda_1(\mathbf{\Sigma}) \times M_y\} \tag{4.137}$$

$$\mathsf{B}_{z,\mathbf{\Sigma}} = \{(z, p, q, \mathbf{\Sigma}, r) \in \mathsf{X}, \|\mathbf{\Sigma}^{1/2} z\| > \lambda_1(\mathbf{\Sigma}) \times M_y\} \tag{4.138}$$

$$\mathsf{C}_{z,\mathbf{\Sigma},p} = \{(z, p, q, \mathbf{\Sigma}, r) \in \mathsf{X}, \lambda_1(\mathbf{\Sigma}) \leqslant M_{\mathbf{\Sigma}} \text{ and } \|\mathbf{\Sigma}^{1/2} z\| \leqslant \lambda_1(\mathbf{\Sigma}) \times M_y \text{ and } \|p\| > M_p\} \tag{4.139}$$

$$\begin{aligned}\mathsf{D}_{z,\boldsymbol{\Sigma},p,q} = \{&(z,p,q,\boldsymbol{\Sigma},r) \in \mathsf{X}, \\ &\lambda_1(\boldsymbol{\Sigma}) \leqslant M_{\boldsymbol{\Sigma}} \text{ and } \|\boldsymbol{\Sigma}^{1/2}z\| \leqslant \lambda_1(\boldsymbol{\Sigma}) \times M_y \text{ and } \|p\| \leqslant M_p \text{ and } \|q\| > M_q\}\end{aligned} \tag{4.140}$$

$$\begin{aligned}\mathsf{E}_{z,\boldsymbol{\Sigma},p,q,r} = \{&(z,p,q,\boldsymbol{\Sigma},r) \in \mathsf{X}, \\ &\lambda_1(\boldsymbol{\Sigma}) \leqslant M_{\boldsymbol{\Sigma}} \text{ and } \|\boldsymbol{\Sigma}^{1/2}z\| \leqslant \lambda_1(\boldsymbol{\Sigma}) \times M_y \text{ and } \|p\| \leqslant M_p \text{ and } \|q\| \leqslant M_q \text{ and } r > M_r\}\end{aligned} \tag{4.141}$$

where the positive constants $M_{\boldsymbol{\Sigma}}$, M_y, M_p, M_q and M_r will be fixed in Lemmas 4.7 and 4.8. We then define K as

$$\mathsf{K} = \mathsf{X} \setminus \{\mathsf{A}_{z,\boldsymbol{\Sigma}} \cup \mathsf{B}_{z,\boldsymbol{\Sigma}} \cup \mathsf{C}_{z,\boldsymbol{\Sigma},p} \cup \mathsf{D}_{z,\boldsymbol{\Sigma},p,q} \cup \mathsf{E}_{z,\boldsymbol{\Sigma},p,q,r}\}. \tag{4.142}$$

As shown in the next lemma, K is a compact of X.

Lemma 4.6. Consider the set K defined as (4.142), then K equals

$$\mathsf{K} = \left\{(z,p,q,\boldsymbol{\Sigma},r) \in \mathsf{X} \mid \lambda_1(\boldsymbol{\Sigma}) \leqslant M_{\boldsymbol{\Sigma}}, \|\boldsymbol{\Sigma}^{1/2}z\| \leqslant \lambda_1(\boldsymbol{\Sigma}) \times M_y, \|p\| \leqslant M_p, \|q\| \leqslant M_q, r \leqslant M_r\right\}$$

and thus it is a compact of X as it is closed and bounded.

Proof. Let $\theta = (z,p,q,\boldsymbol{\Sigma},r) \in \mathsf{X}$. We have

$$\begin{aligned}\theta \in \mathsf{K} &\Leftrightarrow \theta \notin \mathsf{A}_{z,\boldsymbol{\Sigma}} \cup \mathsf{B}_{z,\boldsymbol{\Sigma}} \cup \mathsf{C}_{z,\boldsymbol{\Sigma},p} \cup \mathsf{D}_{z,\boldsymbol{\Sigma},p,q} \cup \mathsf{E}_{z,\boldsymbol{\Sigma},p,q,r} \\ &\Leftrightarrow \theta \notin \mathsf{A}_{z,\boldsymbol{\Sigma}} \text{ and } \theta \notin \mathsf{B}_{z,\boldsymbol{\Sigma}} \text{ and } \theta \notin \mathsf{C}_{z,\boldsymbol{\Sigma},p} \text{ and } \theta \notin \mathsf{D}_{z,\boldsymbol{\Sigma},p,q} \text{ and } \theta \notin \mathsf{E}_{z,\boldsymbol{\Sigma},p,q,r} \\ &\Leftrightarrow \left[\lambda_1(\boldsymbol{\Sigma}) \leqslant M_{\boldsymbol{\Sigma}} \text{ or } \|\boldsymbol{\Sigma}^{1/2}z\| > \lambda_1(\boldsymbol{\Sigma}) \times M_y\right] \text{ and } \|\boldsymbol{\Sigma}^{1/2}z\| \leqslant \lambda_1(\boldsymbol{\Sigma}) \times M_y \\ &\qquad \text{and } \left[\lambda_1(\boldsymbol{\Sigma}) > M_{\boldsymbol{\Sigma}} \text{ or } \|\boldsymbol{\Sigma}^{1/2}z\| > \lambda_1(\boldsymbol{\Sigma}) \times M_y \text{ or } \|p\| \leqslant M_p\right] \\ &\qquad \text{and } \left[\lambda_1(\boldsymbol{\Sigma}) > M_{\boldsymbol{\Sigma}} \text{ or } \|\boldsymbol{\Sigma}^{1/2}z\| > \lambda_1(\boldsymbol{\Sigma}) \times M_y \text{ or } \|p\| > M_p \text{ or } \|q\| \leqslant M_q\right] \\ &\qquad \text{and } \left[\lambda_1(\boldsymbol{\Sigma}) > M_{\boldsymbol{\Sigma}} \text{ or } \|\boldsymbol{\Sigma}^{1/2}z\| > \lambda_1(\boldsymbol{\Sigma}) \times M_y \text{ or } \|p\| > M_p \text{ or } \|q\| > M_q \text{ or } r \leqslant M_r\right] \\ &\Leftrightarrow \lambda_1(\boldsymbol{\Sigma}) \leqslant M_{\boldsymbol{\Sigma}} \text{ and } \|\boldsymbol{\Sigma}^{1/2}z\| \leqslant \lambda_1(\boldsymbol{\Sigma}) \times M_y \text{ and } \|p\| \leqslant M_p \text{ and } \|q\| \leqslant M_q \text{ and } r \leqslant M_r.\end{aligned}$$

□

Given the definition of K as $\mathsf{X} \setminus \{\mathsf{A}_{z,\boldsymbol{\Sigma}} \cup \mathsf{B}_{z,\boldsymbol{\Sigma}} \cup \mathsf{C}_{z,\boldsymbol{\Sigma},p} \cup \mathsf{D}_{z,\boldsymbol{\Sigma},p,q} \cup \mathsf{E}_{z,\boldsymbol{\Sigma},p,q,r}\}$, we prove a geometric drift condition outside K by proving the geometric drift condition (D) for $\theta_0 \in \mathsf{A}_{z,\boldsymbol{\Sigma}}$ (in Lemma 4.7), for $\theta_0 \in \mathsf{B}_{z,\boldsymbol{\Sigma}}$ (in Lemma 4.8(i)), for $\theta_0 \in \mathsf{C}_{z,\boldsymbol{\Sigma},p}$ (in Lemma 4.8(ii)), for $\theta_0 \in \mathsf{D}_{z,\boldsymbol{\Sigma},p,q}$ (in Lemma 4.8(iii)) and for $\theta_0 \in \mathsf{E}_{z,\boldsymbol{\Sigma},p,q,r}$ (in Lemma 4.8(iv)). This in turn shows that the geometric drift condition is satisfied for all θ_0 outside K and thus concludes the proof of Proposition 4.4.

We are thus left with stating and proving Lemma 4.7 and Lemma 4.8. It is done in the next two sections.

3.6.1 *Control of V when the expected covariance matrix decreases*

In this section we complete the proof of Proposition 4.4 by proving in Lemma 4.7 the decrease condition (D) when the initial state belongs to $\mathsf{A}_{z,\boldsymbol{\Sigma}}$ defined in (4.137) with the constants $M_{\boldsymbol{\Sigma}} > 0$ and $M_y > 0$ that are fixed within the lemma. The proof of the lemma is based on the upper bound of the expected largest eigenvalue of the updated covariance matrix in Corollary 4.5.

Lemma 4.7. Consider the function V defined in (4.25) as

$$V(z,p,q,\mathbf{\Sigma},r) = \|\mathbf{H}^{1/2}z\|^2 + \beta\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}) + \gamma_p\|p\| + \gamma_q\|\mathbf{H}^{1/2}q\|^2 + \gamma_r r$$

where $\beta \geqslant 1$, $\gamma_p \geqslant 0$ and $\gamma_q, \gamma_r \in (0,1)$. Consider that the hyperparameters of CMA-ES satisfy **H**1, consider an ellipsoidal objective function f satisfying **F**3 with quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}^d_{++}$ and consider the normalization function $R(\cdot)$ verifying **R**3 with $\mathbf{H}$ for defining the normalized Markov chain satisfying (4.17). Suppose moreover that the sampling distribution ν_U^d is the standard normal distribution (i.e., such that **N**5 holds), and that the weights $\mathbf{w}_m$ and $\mathbf{w}_c$ and the stepsize change function $\Gamma_{d_\sigma}(\cdot)$ satisfy **W**1 and **Γ**1-**Γ**4, respectively.
There exist some constants $C \geqslant 1$, $\varepsilon \in (0,1)$, $M_{\mathbf{\Sigma}} > 0$ and $M_y > 0$, such that for any $c_1 \geqslant 0$, $c_\mu > 0$, $c_m > 0$, $c_c, c_\sigma \in (0,1]$ and $d_\sigma \geqslant 1$ that satisfy: $c_1 + c_\mu \in (0,1)$; d_σ^{-1} is sufficiently larger than c_μ, sufficiently larger than $c_m^{3/2}$, sufficiently larger than γ_q and sufficiently smaller than c_m; $(1-c_\sigma)$ and $2(1-c_1-c_\mu)^{-1}(1-c_c)^2$ are both smaller than $1-c_\mu$; c_1 is sufficiently small; we find that if $\beta = C/(d_\sigma c_\mu)$, then for every initial condition $\theta_0 \in \mathsf{A}_{z,\mathbf{\Sigma}} \subset \mathsf{X}$, where $\theta_0 = (z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \in \mathsf{X} = \mathbb{R}^{3d} \times R^{-1} \times [1-c_1-c_\mu, +\infty)$, and $\mathsf{A}_{z,\mathbf{\Sigma}}$ is defined in (4.137) as

$$\mathsf{A}_{z,\mathbf{\Sigma}} = \{(z,p,q,\mathbf{\Sigma},r) \in \mathsf{X} \text{ such that } \lambda_1(\mathbf{\Sigma}) > M_{\mathbf{\Sigma}} \text{ and } \|\mathbf{\Sigma}^{1/2}z\| \leqslant \lambda_1(\mathbf{\Sigma}) \times M_y\},$$

the Markov chain $\Theta = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17) satisfies the decrease condition (D).

Proof. We fix $C > 0$.
By Corollary 4.5, there exist constants $M_{\mathbf{\Sigma}} > 0$, $M_y > 0$ and $\delta \in (0,1)$, such that if c_1 is sufficiently smaller than c_μ and if c_μ is small enough, then $\|\mathbf{\Sigma}_0^{1/2}z_0\| \leqslant M_y\lambda_1(\mathbf{\Sigma}_0), \lambda_1(\mathbf{\Sigma}_0) \geqslant M_{\mathbf{\Sigma}}$ implies that $\mathbb{E}_0\left[\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right)\right] \leqslant (1-\delta c_\mu)\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) + 3(1-c_1-c_\mu)^{-1}c_1\|\mathbf{H}^{1/2}q_0\|^2$ where we used $r_0 \geqslant 1-c_1-c_\mu$. Moreover, by Proposition 4.18 (i), there exists $\kappa > 0$, such that

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_1\|^2\right] \leqslant (1-c_1-c_\mu)^{-1}(1+2d_\sigma^{-1})^2\|\mathbf{H}^{1/2}z_0\|^2 + \kappa c_m^2\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) \ .$$

Besides, by definition of p_1 in (4.17) and the triangular inequality

$$\begin{aligned}\mathbb{E}_0\left[\|p_1\|\right] &= \mathbb{E}_0\left[\left\|(1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right\|\right] \\ &\leqslant (1-c_\sigma)\|p_0\| + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbb{E}\|U_1\|_\infty \ ,\end{aligned}$$

where $\|U_1\|_\infty = \max_{i=1,\dots,\lambda}\|U_1^i\|$. When $\mathbf{H} = \mathbf{I}_d$, by definition of q_1 in (4.17), we have

$$\mathbb{E}_0\left[\|q_1\|^2\right] = \mathbb{E}_0\left[\left\|r_0^{-1/2}(1-c_c)q_0 + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\sqrt{\mathbf{\Sigma}_0}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right\|^2\right].$$

Using the inequalities $\|x+y\|^2 \leqslant 2\|x\|^2 + 2\|y\|^2$ and $\|\sqrt{\mathbf{\Sigma}_0}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\|^2 \leqslant \lambda_1(\mathbf{\Sigma}_0)\|U_1\|_\infty^2$ we obtain $\mathbb{E}_0\left[\|q_1\|^2\right] \leqslant 2\mathbb{E}_0\left[r_0^{-1}(1-c_c)^2\|q_0\|^2\right] + 2c_c(2-c_c)\mu_{\text{eff}}\mathbb{E}_0\left[\lambda_1(\mathbf{\Sigma}_0)\|U_1\|_\infty^2\right]$. Since $r_0 \in [1-c_1-c_\mu, +\infty)$, we bound $r_0^{-1} \leqslant (1-c_1-c_\mu)^{-1}$

and obtain

$$\mathbb{E}_0\left[\|q_1\|^2\right] \leqslant 2(1-c_1-c_\mu)^{-1}(1-c_c)^2\|q_0\|^2 + 2c_c(2-c_c)\mu_{\text{eff}}\lambda_1\left(\mathbf{\Sigma}_0\right)\mathbb{E}\|U_1\|_\infty^2. \quad (4.143)$$

Using the change of variable property in Proposition 4.6, we can deduce from the previous inequality that when $\mathbf{H} \neq \mathbf{I}_d$

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}q_1\|^2\right] \leqslant 2(1-c_1-c_\mu)^{-1}(1-c_c)^2\|\mathbf{H}^{1/2}q_0\|^2 + 2c_c(2-c_c)\mu_{\text{eff}}\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)\mathbb{E}\|U_1\|_\infty^2 \ . \quad (4.144)$$

Finally, by definition of r_1 in (4.17), when $\mathbf{H} = \mathbf{I}_d$:

$$\mathbb{E}_0[r_1] = \mathbb{E}_0\left[\lambda_d\left((1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_1 q_1 q_1^\top + c_\mu\sqrt{\mathbf{\Sigma}_0}\sum_{i=1}^{\mu} w_i^c U_1^{s_1(i)}[U_1^{s_1(i)}]^\top\sqrt{\mathbf{\Sigma}_0}\right)\right] \ .$$

Using the property $\lambda_d(\mathbf{A}+\mathbf{B}) \leqslant \lambda_d(\mathbf{A}) + \lambda_1(\mathbf{B})$ (see (4.41)) for $\mathbf{A}, \mathbf{B} \in \mathcal{S}^d$, we obtain

$$\mathbb{E}_0[r_1] \leqslant \mathbb{E}_0\left[\lambda_d\left(\sqrt{\mathbf{\Sigma}_0}\times\left((1-c_1-c_\mu)\mathbf{I}_d + c_\mu\sum_{i=1}^{\mu} w_i^c U_1^{s_1(i)}[U_1^{s_1(i)}]^\top\right)\times\sqrt{\mathbf{\Sigma}_0}\right) + \lambda_1\left(c_1 q_1 q_1^\top\right)\right].$$

Using that $\lambda_d(\mathbf{A}) \leqslant e_d\left(\mathbf{\Sigma}_0\right)^\top \mathbf{A} e_d\left(\mathbf{\Sigma}_0\right)$ (see (4.43)) and that $\sqrt{\mathbf{\Sigma}_0}e_d\left(\mathbf{\Sigma}_0\right) = \sqrt{\lambda_d\left(\mathbf{\Sigma}_0\right)}e_d\left(\mathbf{\Sigma}_0\right) = e_d\left(\mathbf{\Sigma}_0\right)$ (since $\lambda_d\left(\mathbf{\Sigma}_0\right) = \lambda_{\min}(\mathbf{\Sigma}_0) = 1$) we obtain

$$\mathbb{E}_0[r_1] \leqslant (1-c_1-c_\mu) + c_\mu\mathbb{E}_0\left[e_d\left(\mathbf{\Sigma}_0\right)^\top\sum_{i=1}^{\mu} w_i^c U_1^{s_1(i)}[U_1^{s_1(i)}]^\top e_d\left(\mathbf{\Sigma}_0\right)\right] + c_1\mathbb{E}_0\|q_1\|^2.$$

When $\mathbf{H} \neq \mathbf{I}_d$, using Proposition 4.6 we obtain that $\mathbb{E}_0[r_1] = \mathbb{E}_0[\hat{r}_1]$ where $\hat{r}_1$ is the variable on the ellipsoid function with normalization $\lambda_{\min}(\mathbf{H}^{1/2}\times\cdot\times\mathbf{H}^{1/2})$. Since we additionally have $e_d\left(\mathbf{\Sigma}_0\right)^\top\sum_{i=1}^{\mu} w_i^c U_1^{s_1(i)}[U_1^{s_1(i)}]^\top e_d\left(\mathbf{\Sigma}_0\right) \leqslant \|U_1\|_\infty^2$, we get (using with a slight abuse a notation without the hat for the Markov chain on the ellipsoid)

$$\mathbb{E}_0[r_1] \leqslant 1 - c_1 - c_\mu + c_\mu\mathbb{E}\|U_1\|_\infty^2 + c_1\mathbb{E}_0\|\mathbf{H}^{1/2}q_1\|^2.$$

All in all, using the formula for the potential function $V(\cdot)$ in (4.25), we have

$$\begin{aligned}
\mathbb{E}_0\left[V(z_1,p_1,q_1,\mathbf{\Sigma}_1,r_1)\right] &= \mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_1\|^2 + \beta\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right) + \gamma_p\|p_1\| + \gamma_q\|\mathbf{H}^{1/2}q_1\|^2 + \gamma_r r_1\right] \\
&\leqslant \underbrace{(1-c_1-c_\mu)^{-1}(1+2d_\sigma^{-1})^2}_{\leqslant 1-\delta c_\mu + \delta/8\frac{1}{\kappa_{\mathbf{H},M_y}}\beta c_\mu}\|\mathbf{H}^{1/2}z_0\|^2 \\
&\quad + \left(\beta(1-\delta c_\mu) + \underbrace{\kappa c_m^2}_{\leqslant\beta c_\mu\delta/8} + \underbrace{2(\gamma_q+c_1\gamma_r)c_c(2-c_c)\mu_{\text{eff}}\mathbb{E}\|U_1\|_\infty^2}_{\leqslant\beta c_\mu\delta/8}\right)\times\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) \\
&\quad + \gamma_p(1-c_\sigma)\|p_0\| \\
&\quad + \left(\underbrace{(\gamma_q+c_1\gamma_r)}_{\leqslant(1-c_\mu)^{-1}(1-\delta c_\mu)\gamma_q}\underbrace{2(1-c_1-c_\mu)^{-1}(1-c_c)^2}_{\leqslant 1-c_\mu} + \underbrace{3\beta(1-c_1-c_\mu)^{-1}c_1}_{\leqslant\delta c_\mu\gamma_q/2}\right)\times\|\mathbf{H}^{1/2}q_0\|^2 \\
&\quad + \underbrace{\gamma_p\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbb{E}\|U_1\|_\infty + \gamma_r\left(1-c_1-c_\mu+c_\mu\mathbb{E}\|U_1\|_\infty^2\right)}_{\leqslant\beta\delta c_\mu/8\times\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)} \ .
\end{aligned}$$

Define $\beta = C/(d_\sigma c_\mu)$ and let $\varepsilon_m > 0$ be such that $\kappa c_m^2 = \varepsilon_m \beta c_\mu$. Moreover, assume that c_m^2 is sufficiently smaller than d_σ^{-1} to have $\varepsilon_m = \kappa c_m^2 d_\sigma c_\mu / C \leqslant \delta/8$.
Likewise, let $\varepsilon_\gamma > 0$ be such that $2(\gamma_q + c_1\gamma_r)c_c(2-c_c)\mu_{\text{eff}}\mathbb{E}\|U_1\|_\infty^2 = \varepsilon_\gamma \beta c_\mu$. If γ_q and $c_1\gamma_r$ are sufficiently smaller than d_σ^{-1}, then we can assume $\varepsilon_\gamma \leqslant \delta/8$.
Next, let $\varepsilon_\sigma > 0$ be such that $(1-c_1-c_\mu)^{-1}(1+2d_\sigma^{-1})^2 = 1 - \delta c_\mu + \varepsilon_\sigma/d_\sigma$. Furthermore as explained below the following inequality holds

$$\|\mathbf{H}^{1/2} z_0\|^2 \leqslant \kappa_{\mathbf{H},M_y} \lambda_1\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}\right) \tag{4.145}$$

where $\kappa_{\mathbf{H},M_y} = M_y^2 \text{Cond}(\mathbf{H}) > 0$ is a constant that depends only on M_y and $\mathbf{H}$. Indeed, $\|\mathbf{H}^{1/2}z_0\|^2 \leqslant \lambda_1(\mathbf{H})\|z_0\|^2 = \lambda_1(\mathbf{H})\|\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_0^{1/2}z_0\|^2 \leqslant \lambda_1(\mathbf{H})\frac{1}{\lambda_1(\boldsymbol{\Sigma}_0)}\|\boldsymbol{\Sigma}_0^{1/2}z_0\|^2 \leqslant \lambda_1(\mathbf{H})M_z^2\lambda_1(\boldsymbol{\Sigma}_0)$. However since $\lambda_1(\boldsymbol{\Sigma}_0) = \lambda_1(\mathbf{H}^{-1/2}\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}\mathbf{H}^{-1/2}) = \|\mathbf{H}^{-1/2}\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}\mathbf{H}^{-1/2}\|$, using the property that the subordinate norm of a product of two matrices is lower or equal than the product of the norm of the matrices, we find that

$$\lambda_1(\boldsymbol{\Sigma}_0) \leqslant \|\mathbf{H}^{-1/2}\|\|\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}\|\|\mathbf{H}^{-1/2}\| = \lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})/\lambda_{\min}(\mathbf{H}) = \lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) \tag{4.146}$$

since by assumption $\lambda_{\min}(\mathbf{H}) = 1$. Overall we find that

$$\|\mathbf{H}^{1/2}z_0\|^2 \leqslant \text{Cond}(\mathbf{H})M_z^2\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}). \tag{4.147}$$

Up to taking $C \geqslant 1$ sufficiently larger than M_y^2, we can assume that $\varepsilon_\sigma \leqslant (\delta C)/(8M_y^2\text{Cond}(\mathbf{H}))$ and that $\beta \geqslant 1$, and thus we obtain: $(1-c_1-c_\mu)^{-1}(1+2d_\sigma^{-1})^2 = (1-\delta c_\mu + \varepsilon_\sigma/d_\sigma) \leqslant 1 - \delta c_\mu + \frac{\delta}{8}\frac{C}{M_y^2\text{Cond}(\mathbf{H})}\frac{c_\mu}{c_\mu d_\sigma} = 1 - \delta c_\mu + \frac{\delta}{8}\frac{1}{\kappa_{\mathbf{H},M_y}}\beta c_\mu$ and thus

$$\begin{aligned}(1-c_1-c_\mu)^{-1}(1+2d_\sigma^{-1})^2\|\mathbf{H}^{1/2}z_0\|^2 &\leqslant \left(1 - \delta c_\mu + \frac{\delta}{8}\frac{1}{\kappa_{\mathbf{H},M_y}}\beta c_\mu\right)\|\mathbf{H}^{1/2}z_0\|^2 \\ &\leqslant (1-\delta c_\mu)\|\mathbf{H}^{1/2}z_0\|^2 + \beta\frac{\delta}{8}c_\mu\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})\end{aligned}$$

where we have used (4.145) for the last inequality. Moreover, up to taking $M_{\boldsymbol{\Sigma}}$ larger, we can assume that $\gamma_p\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbb{E}\|U_1\|_\infty + \gamma_r(1-c_1-c_\mu+c_\mu\mathbb{E}\|U_1\|_\infty^2) \leqslant M_{\boldsymbol{\Sigma}}\beta\delta c_\mu/8 \leqslant \beta\delta c_\mu/8 \times \lambda_1(\boldsymbol{\Sigma}_0) \leqslant \beta\delta c_\mu/8 \times \lambda_1\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}\right)$ where we have used (4.146) for the last inequality.
Finally, assume that $(1+c_\sigma) \leqslant 1 - c_\mu \leqslant 1 - \delta c_\mu$ and that c_1 is sufficiently small to have $\gamma_q + c_1\gamma_r \leqslant (1-c_\mu)^{-1}(1-\delta c_\mu)\gamma_q$ and $3\beta(1-c_1-c_\mu)^{-1}c_1 \leqslant \delta c_\mu\gamma_q/2$. Then, we obtain using also the assumption that $2(1-c_1-c_\mu)^{-1}(1-c_c)^2 \leqslant 1 - c_\mu$

$$\begin{aligned}\mathbb{E}_0\left[V(z_1,p_1,q_1,\boldsymbol{\Sigma}_1,r_1)\right] &\leqslant (1-\delta c_\mu)\|\mathbf{H}^{1/2}z_0\|^2 + \beta\left(1 - \delta c_\mu + 4\frac{\delta}{8}c_\mu\right)\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) \\ &\quad + \gamma_p(1-\delta c_\mu)\|p_0\| + \gamma_q\left((1-c_\mu)^{-1}(1-\delta c_\mu)(1-c_\mu) + \frac{\delta}{2}c_\mu\right)\|\mathbf{H}^{1/2}q_0\|^2 \\ &\leqslant \left(1 - \frac{\delta}{2}c_\mu\right) \times \left(\|\mathbf{H}^{1/2}z_0\|^2 + \beta\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) + \gamma_p\|p_0\| + \gamma_q\|\mathbf{H}^{1/2}q_0\|^2\right) \\ &\leqslant \left(1 - \frac{\delta}{2}c_\mu\right) \times V(z_0,p_0,q_0,\boldsymbol{\Sigma}_0,r_0)\ . \end{aligned} \tag{4.148}$$

□

3.6.2 *Control of V when the expected mean decreases*

In this last lemma needed for the proof of Proposition 4.4, we analyze cases where the expected drift decreases and when Corollary 4.5 does not apply to obtain a decrease of the expected largest eigenvalue of the covariance matrix, i.e., when the initial state belongs to $\mathsf{B}_{z,\mathbf{\Sigma}}$ (i), $\mathsf{C}_{z,\mathbf{\Sigma},p}$ (ii), $\mathsf{D}_{z,\mathbf{\Sigma},p,q}$ (iii), or $\mathsf{E}_{z,\mathbf{\Sigma},p,q,r}$ (iv), see (4.138), (4.139), (4.140), (4.141), respectively.

Lemma 4.8. Consider the CMA-ES algorithm with hyperparameters satisfying **H**1 optimizing an ellipsoidal objective function f satisfying **F**3 with quasi-Hessian matrix $\mathbf{H}$ and a normalization function R satisfying **R**3 with the matrix $\mathbf{H}$. Suppose also that the sampling distribution ν_U^d is standard normal (satisfies **N**5).
Let $C \geqslant 1$, $M_{\mathbf{\Sigma}} > 0$, and $M_y > 0$, and let $\delta_\sigma > 8$ be sufficiently larger than CM_z^{-2}. Moreover, let $\delta_p > 0$ be such that the stepsize change function Γ_{d_σ} satisfies **Γ**1-**Γ**4 (where **Γ**3 holds with δ_σ, δ_p and $k = 4$). Moreover, suppose that $\mu \leqslant \lambda/2$, that the weights $\mathbf{w}_m$ and $\mathbf{w}_c$ are such that **W**1 holds and that μ_{eff} is sufficiently large to have (4.71). Consider the potential function V defined in (4.25) as

$$V(z,p,q,\mathbf{\Sigma},r) = \|\mathbf{H}^{1/2}z\|^2 + \beta\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}) + \gamma_p\|p\| + \gamma_q\|\mathbf{H}^{1/2}q\|^2 + \gamma_r r$$

where $\beta, \gamma_p, \gamma_q, \gamma_r > 0$. Then, there exists some constant $\varepsilon > 0 > 0$ such that, for any $c_1 \geqslant 0$, $c_\mu > 0$, $c_m > 0$, $c_c, c_\sigma \in (0,1]$ and $d_\sigma \geqslant 1$ satisfying

- $c_1 + c_\mu < 1$,
- d_σ^{-1} is sufficiently
  - larger than $c_\mu, c_m^{3/2}$
  - smaller than c_m
- $(1-c_c)^2 \leqslant (1-c_\mu)(1-c_1-c_\mu)/2$ and $1-c_\sigma \leqslant 1-c_\mu$,
- c_1 is sufficiently small,

and for any values of $\beta \geqslant 1, \gamma_p, \gamma_q, \gamma_r > 0$ with

- $\gamma_p, \gamma_q, \gamma_r$ sufficiently smaller than d_σ^{-1},
- $c_1\gamma_r$ sufficiently smaller than $c_\mu\gamma_q$,
- $\beta = C/(c_\mu d_\sigma)$,

and for some well-chosen values of $M_p, M_q, M_r > 0$, then the normalized Markov chain of CMA-ES, $\Theta = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying (4.17), satisfies the decrease condition (D) for any initial condition $\theta_0 = (z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0)$ if either

(i) $\theta_0 \in \mathsf{B}_{z,\mathbf{\Sigma}}$, see (4.138), and if either $c_c = 1$ or $c_\sigma = 1$; or

(ii) $\theta_0 \in \mathsf{C}_{z,\mathbf{\Sigma},p}$, see (4.139); or

(iii) $\theta_0 \in \mathsf{D}_{z,\mathbf{\Sigma},p,q}$, see (4.140); or

(iv) $\theta_0 \in \mathsf{E}_{z,\mathbf{\Sigma},p,q,r}$, see (4.141).

Proof.
In all cases, observe that, if $c_1\mu_{\text{eff}} \leqslant c_\mu$ and that c_1 and c_μ are small enough so that $(1-c_1-c_\mu)^{-1} \leqslant \sqrt{3/2}$, then by Corollary 4.4 and Proposition 4.3, since r_0 and r_1 are greater than or

equal to $1 - c_1 - c_\mu \geqslant \sqrt{2/3}$ and thus $1/(r_1 r_0) \leqslant 3/2$:

$$\mathbb{E}_0\left[\lambda_1\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_1\mathbf{H}^{1/2}\right)\right] \leqslant (1 + \underbrace{3(1-c_1-c_\mu)^{-1}c_\mu d_\mu \mathbb{E}\|U_1\|_\infty^2}_{\leqslant\sqrt{6}c_\mu d_\mu\mathbb{E}\|U_1\|_\infty^2 := \rho c_\mu})\lambda_1\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}\right) + \underbrace{\frac{2c_1}{r_1 r_0}}_{\leqslant 3c_1}(1-c_c)^2\|\mathbf{H}^{1/2}q_0\|^2 \ . \quad (4.149)$$

We assume throughout the proof that the above holds. Note that we also have, by exploiting (4.17) and the triangular inequality, since, by definition of ρ in (4.149), $\rho \geqslant \mathbb{E}\|U_1\|_\infty^2$:

$$\mathbb{E}_0\|p_1\| = \mathbb{E}_0\left\|(1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right\| \leqslant (1-c_\sigma)\|p_0\| + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}\rho} \ . \quad (4.150)$$

Likewise using (4.144), we find

$$\mathbb{E}_0\|\mathbf{H}^{1/2}q_1\|^2 \leqslant 2(1-c_1-c_\mu)^{-1}(1-c_c)^2\|\mathbf{H}^{1/2}q_0\|^2 + 2c_c(2-c_c)\mu_{\text{eff}}\rho\lambda_1\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}\right) \ . \quad (4.151)$$

Moreover, by Corollary 4.4,

$$r_1\lambda_d(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_1\mathbf{H}^{1/2}) \leqslant \left(1-c_1-c_\mu+(2c_1\mu_{\text{eff}}+c_\mu)d_\mu\|U_1\|_\infty^2\right)\lambda_d(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) + 2c_1 r_0^{-1}(1-c_c)^2\|\mathbf{H}^{1/2}q_0\|^2 \ .$$

However, $\lambda_d(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_1\mathbf{H}^{1/2}) = \lambda_d(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) = 1$ by assumption **R**3. Then, as for (4.149), we finally obtain

$$\mathbb{E}_0[r_1] \leqslant (1+\rho c_\mu) + 3c_1(1-c_c)^2\|\mathbf{H}^{1/2}q_0\|^2 \ . \quad (4.152)$$

We start now the proof of (i). Assume $\|\boldsymbol{\Sigma}_0^{1/2}z_0\| \geqslant M_y\lambda_1(\boldsymbol{\Sigma}_0)$. Since $\|\boldsymbol{\Sigma}_0^{1/2}z_0\| \leqslant \sqrt{\lambda_1(\boldsymbol{\Sigma}_0)}\|z_0\|$, this implies in particular that $\|z_0\| \geqslant M_y\sqrt{\lambda_1(\boldsymbol{\Sigma}_0)}$. First suppose that $c_c = 1$ and thus that $r_0^{-1/2}(1-c_c)\|q_0\| = 0 \leqslant \bar{q}\sqrt{\lambda_1(\boldsymbol{\Sigma}_0)}$ for any $\bar{q} > 0$. By (4.149) we have that

$$\mathbb{E}_0\left[\lambda_1\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_1\mathbf{H}^{1/2}\right)\right] \leqslant (1+\rho c_\mu)\,\lambda_1\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}\right) \ . \quad (4.153)$$

and by (4.151)

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}q_1\|^2\right] \leqslant 2\mu_{\text{eff}}\rho\lambda_1\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}\right) \quad (4.154)$$

and by (4.152)

$$\mathbb{E}_0[r_1] \leqslant 1 + c_\mu\rho \ .$$

Suppose then that c_1, c_μ and $c_m^{3/2}$ are sufficiently smaller than d_σ^{-1} and that d_σ^{-1} is sufficiently smaller than c_m to apply Proposition 4.18(iii). Then, we have

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_1\|^2\right] \leqslant \left(1 - \frac{\delta_\sigma}{2d_\sigma}\right) \times \|\mathbf{H}^{1/2}z_0\|^2 \ , \quad \text{(a)}$$

or

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_2\|^2\right] \leqslant \left(1 - \frac{\delta_\sigma - 8}{8d_\sigma}\right) \times \|\mathbf{H}^{1/2}z_0\|^2 + \kappa c_m^2\lambda_1\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}\right) \ . \quad \text{(b)}$$

Suppose first that (a) holds. All in all, we obtain, since $\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}) \geqslant \lambda_1(\mathbf{\Sigma}_0) \geqslant 1$ (see (4.146)):

$$\begin{aligned}
\mathbb{E}_0\left[V(z_1,p_1,q_1,\mathbf{\Sigma}_1,r_1)\right] &= \mathbb{E}_0\|\mathbf{H}^{1/2}z_1\|^2 + \beta\mathbb{E}[\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2})] + \gamma_p\mathbb{E}_0\|p_1\| \\
&\quad + \gamma_q\mathbb{E}_0[\|\mathbf{H}^{1/2}q_1\|^2] + \gamma_r\mathbb{E}_0[r_1] \\
&\leqslant \left(1-\frac{\delta_\sigma}{2d_\sigma}\right)\times\|\mathbf{H}^{1/2}z_0\|^2 + \beta(1+\rho c_\mu)\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}) \\
&\quad + \gamma_p\left((1-c_\sigma)\|p_0\| + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}\rho}\right) \\
&\quad + \gamma_q\times 2\mu_{\text{eff}}\rho\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}) + \gamma_r(1+\rho c_\mu) \\
&\leqslant \left(1-\frac{\delta_\sigma}{2d_\sigma}\right)\times\|\mathbf{H}^{1/2}z_0\|^2 \\
&\quad + \left[\begin{array}{l}\beta\times(1+\rho c_\mu) + \gamma_p\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}\rho} \\ +\gamma_q 2\mu_{\text{eff}}\rho + \gamma_r\,(1+c_\mu\rho)\end{array}\right]\times\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) \\
&\quad + \gamma_p(1-c_\sigma)\|p_0\| \ .
\end{aligned}$$

Then, if $\gamma_p\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}} \leqslant \sqrt{\rho}\frac{C}{d_\sigma} = \beta c_\mu\sqrt{\rho}$, $\gamma_q\mu_{\text{eff}} \leqslant \frac{C}{d_\sigma} = \beta c_\mu$, $\gamma_r \leqslant \min\{1,\rho\}\frac{C}{d_\sigma} = \beta c_\mu\min\{1,\rho\}$, since $\beta = \frac{C}{c_\mu d_\sigma}$, we obtain then

$$\begin{aligned}
\mathbb{E}_0\left[V(z_1,p_1,q_1,\mathbf{\Sigma}_1,r_1)\right] &\leqslant \left(1-\frac{\delta_\sigma}{2d_\sigma}\right)\times\|\mathbf{H}^{1/2}z_0\|^2 + [\beta\times(1+6\rho c_\mu)]\times\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) \\
&\quad + \gamma_p(1-c_\sigma)\|p_0\| \ .
\end{aligned}$$

Furthermore, suppose that $7\rho\beta c_\mu = \frac{7\rho C}{d_\sigma} \leqslant \frac{\delta_\sigma}{4d_\sigma}\times M_y^2$ (this is encompassed in the assumption that δ_σ is sufficiently larger than CM_y^{-2}). Then,

$$\begin{aligned}
[\beta\times(1+6\rho c_\mu)]\times\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) &= [\beta\times(1-\rho c_\mu)]\times\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) + 7\rho\beta c_\mu\times\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) \\
&\leqslant [\beta\times(1-\rho c_\mu)]\times\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right) + \frac{\delta_\sigma}{4d_\sigma}\times\|\mathbf{H}^{1/2}z_0\|^2
\end{aligned}$$

and thus

$$\mathbb{E}_0\left[V(z_1,,p_1,q_1,\mathbf{\Sigma}_1,r_1)\right] \leqslant \max\left\{1-\frac{\delta_\sigma}{4d_\sigma}, 1-\rho c_\mu, 1-c_\sigma\right\}\times V(z_0,p_0,q_0,\mathbf{\Sigma}_0,r_0) \ ,$$

which proves (i) since we assume d_σ^{-1} to be sufficiently larger than c_μ and that $1-c_\sigma \leqslant 1-c_\mu$ in the case (a). Now suppose (b). As in (4.153) we have

$$\mathbb{E}_0\left[\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_2\mathbf{H}^{1/2})\right] \leqslant (1+\rho c_\mu)\mathbb{E}_0\left[\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2})\right] \leqslant (1+\rho c_\mu)^2\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}) \ . \tag{4.155}$$

Like in (4.150)

$$\mathbb{E}_0\|p_2\| \leqslant (1-c_\sigma)\mathbb{E}_0\|p_1\| + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}\rho} \leqslant (1-c_\sigma)^2\|p_0\|^2 + (2-c_\sigma)\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}\rho} \tag{4.156}$$

and like in (4.154)

$$\mathbb{E}_0\|\mathbf{H}^{1/2}q_2\|^2 \leqslant 2\mu_{\text{eff}}\rho\mathbb{E}_0[\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})] \leqslant 2\mu_{\text{eff}}\rho(1+\rho c_\mu)\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) \ . \quad (4.157)$$

Finally, from (4.152):

$$\mathbb{E}_0[r_2] \leqslant 1+\rho c_\mu \ . \quad (4.158)$$

All in all using $\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) \geqslant 1$:

$$\begin{aligned}\mathbb{E}_0\left[V(z_2,p_2,q_2,\boldsymbol{\Sigma}_2,r_2)\right] &\leqslant \left(1-\frac{\delta_\sigma-8}{8d_\sigma}\right)\times\|\mathbf{H}^{1/2}z_0\|^2 \\ &+\left[\begin{array}{l}\kappa c_m^2+\beta\times(1+\rho c_\mu)^2+\gamma_p(2-c_\sigma)\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}\rho} \\ +2\gamma_q\mu_{\text{eff}}\rho(1+\rho c_\mu) \\ +\gamma_r\times(1+c_\mu\rho)\end{array}\right]\times\lambda_1\left(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}\right) \\ &+\gamma_p(1-c_\sigma)^2\|p_0\| \ .\end{aligned}$$

Then, as for case (a), given that $\gamma_p\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}\rho} \leqslant \beta c_\mu\rho$, $\gamma_q\mu_{\text{eff}} \leqslant \beta c_\mu$ and $\gamma_r \leqslant \beta c_\mu\min\{1,\rho\}$, and if $\kappa c_m^2 \leqslant \rho C/d_\sigma = \beta\rho c_\mu$, we have

$$\begin{aligned}\mathbb{E}_0\left[V(z_2,p_2,q_2,\boldsymbol{\Sigma}_2,r_2)\right] &\leqslant \left(1-\frac{\delta_\sigma-8}{8d_\sigma}\right)\times\|\mathbf{H}^{1/2}z_0\|^2 \\ &+\beta\left(1+7\rho c_\mu+3\rho^2c_\mu\right)\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})+\gamma_p(1-c_\sigma)^2\|p_0\|\end{aligned}$$

Then, if $(8\rho+3\rho^2)\beta c_\mu = (8\rho+3\rho^2)C/d_\sigma \leqslant (\delta-8)/(16d_\sigma)M_z^2$, we obtain:

$$\begin{aligned}\mathbb{E}_0\left[V(z_2,p_2,q_2,\boldsymbol{\Sigma}_2,r_2)\right] &\leqslant \left(1-\frac{\delta_\sigma-8}{16d_\sigma}\right)\times\|\mathbf{H}^{1/2}z_0\|^2 \\ &+\beta\left(1-\rho c_\mu\right)\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})+\gamma_p(1-c_\sigma)^2\|p_0\| \\ &\leqslant \max\{1-\frac{\delta_\sigma-8}{16d_\sigma},1-\rho c_\mu,(1-c_\sigma)^2\}V(z_0,p_0,q_0,\boldsymbol{\Sigma}_0,r_0) \ ,\end{aligned}$$

ending the proof of (i) in case (b), since d_σ^{-1} is supposed to be sufficiently larger than c_μ and $(1-c_\sigma)^2 \leqslant 1-c_\sigma \leqslant 1-c_\mu$.
Now, instead of $c_c = 1$ we assume $c_\sigma = 1$ so that $(1-c_\sigma)\|p_0\| \leqslant \delta_p\mathbb{E}\|\nu_U^d\|$. By Proposition 4.18(ii), we have

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}z_1\|^2\right] \leqslant \left(1-\frac{\delta_\sigma}{2d_\sigma}\right)\|\mathbf{H}^{1/2}z_0\|^2 \ . \quad (4.159)$$

Besides, from (4.149) and using that $\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) \leqslant \text{Cond}(\mathbf{H})^2M_y^{-2}\|\mathbf{H}^{1/2}z_0\|^2$, since

$$\begin{aligned}\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) \leqslant \lambda_1(\mathbf{H})\lambda_1(\boldsymbol{\Sigma}_0) &\leqslant \lambda_1(\mathbf{H})M_z^{-1}\|\sqrt{\boldsymbol{\Sigma}_0}z_0\| \leqslant \lambda_1(\mathbf{H})M_z^{-1}\sqrt{\lambda_1(\boldsymbol{\Sigma}_0)}\|z_0\| \\ &\leqslant \text{Cond}(\mathbf{H})M_z^{-1}\sqrt{\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})}\|\mathbf{H}^{1/2}z_0\| \ ,\end{aligned}$$

we have

$$\begin{aligned}&\mathbb{E}_0\left[\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_1\mathbf{H}^{1/2})\right] \leqslant (1+\rho c_\mu)\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})+3c_1\|\mathbf{H}^{1/2}q_0\|^2 \\ &\leqslant (1-\rho c_\mu)\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})+2\rho c_\mu\text{Cond}(\mathbf{H})^2M_y^{-2}\|\mathbf{H}^{1/2}z_0\|^2+3c_1\|\mathbf{H}^{1/2}q_0\|^2 \ . \quad (4.160)\end{aligned}$$

Since $c_\sigma = 1$, by (4.150), we have

$$\mathbb{E}_0\|p_1\| \leqslant \sqrt{\mu_{\text{eff}}\rho} \ . \tag{4.161}$$

We assume that $2(1-c_1-c_\mu)^{-1}(1-c_c)^2 \leqslant 1-c_\mu$. Then, by (4.151) we have

$$\mathbb{E}_0\left[\|\mathbf{H}^{1/2}q_1\|^2\right] \leqslant (1-c_\mu)\|\mathbf{H}^{1/2}q_0\|^2 + 2\mu_{\text{eff}}\rho\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) \ . \tag{4.162}$$

Then, combining (4.159), (4.160), (4.161), (4.162) with the upper bound of $\mathbb{E}_0 r_1$ in (4.152) and the expression of the potential function V in (4.25), we have (using one more time that $\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) \geqslant 1$)

$$\begin{aligned}
\mathbb{E}_0\left[V(z_1,p_1,q_1,\boldsymbol{\Sigma}_1,r_1)\right] &\leqslant \left(1-\frac{\delta_\sigma}{2d_\sigma}\right)\|\mathbf{H}^{1/2}z_0\|^2 + \beta\times 2\rho c_\mu \text{Cond}(\mathbf{H})^2 M_y^{-2}\|\mathbf{H}^{1/2}z_0\|^2 \\
&\quad + \beta\times(1-\rho c_\mu)\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) + \beta\times 3c_1\|\mathbf{H}^{1/2}q_0\|^2 \\
&\quad + \gamma_p\sqrt{\mu_{\text{eff}}\rho} + \gamma_q\times(1-c_\mu)\|\mathbf{H}^{1/2}q_0\|^2 \\
&\quad + \gamma_q\times 2\mu_{\text{eff}}\rho\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) \\
&\quad + \gamma_r\times(1+\rho c_\mu) + \gamma_r\times 3c_1\|\mathbf{H}^{1/2}q_0\|^2 \\
&\leqslant \left(1-\frac{\delta_\sigma}{2d_\sigma}+2\beta\rho c_\mu\text{Cond}(\mathbf{H})^2M_y^{-2}\right)\|\mathbf{H}^{1/2}z_0\|^2 \\
&\quad + \left(\beta(1-\rho c_\mu)+\gamma_p\sqrt{\mu_{\text{eff}}\rho}+2\gamma_q\mu_{\text{eff}}\rho+\gamma_r(1+\rho c_\mu)\right)\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2}) \\
&\quad + \left(\gamma_q(1-c_\mu)+(\beta+\gamma_r)\times 3c_1\right)\|\mathbf{H}^{1/2}q_0\|^2 \ .
\end{aligned} \tag{4.163}$$

We set $\beta = C/(c_\mu d_\sigma)$ where we assume that $C > 0$ is smaller than $\delta_\sigma M_z^2/(8\text{Cond}(\mathbf{H})^2\rho)$ (see the assumption that δ_σ is sufficiently larger than CM_z^{-2}), so that in front of $\|\mathbf{H}^{1/2}z_0\|^2$ in (4.163) we have

$$1-\frac{\delta_\sigma}{2d_\sigma}+2\beta\rho c_\mu\text{Cond}(\mathbf{H})^2M_y^{-2} \leqslant 1-\frac{\delta_\sigma}{2d_\sigma}+2\frac{\delta_\sigma M_y^2\times\rho c_\mu\text{Cond}(\mathbf{H})^2M_y^{-2}}{8\text{Cond}(\mathbf{H})\rho\times c_\mu d_\sigma} = 1-\frac{\delta_\sigma}{2d_\sigma}+\frac{\delta_\sigma}{4d_\sigma} = 1-\frac{\delta_\sigma}{4d_\sigma} \ . \tag{4.164}$$

We also suppose that c_1 is sufficiently small so that $c_1 \leqslant \gamma_q c_\mu^2 d_\sigma/(6C)$ and therefore $\beta\times 3c_1 = 3Cc_1/(c_\mu d_\sigma) \leqslant \gamma_q c_\mu/2$. Moreover, we assume that $c_1\gamma_r \leqslant \gamma_q c_\mu/12$, which implies that the term in front of $\|\mathbf{H}^{1/2}q_0\|^2$ in (4.163) satisfies

$$\gamma_q(1-c_\mu)+(\beta+\gamma_r)\times 3c_1 \leqslant \gamma_q\times\left(1-c_\mu+\frac{c_\mu}{2}+\frac{c_\mu}{4}\right) = \gamma_q\times\left(1-\frac{c_\mu}{4}\right) \ . \tag{4.165}$$

Finally, we choose $\gamma_p > 0$, $\gamma_q > 0$, $\gamma_r > 0$ sufficiently smaller than $1/d_\sigma$ to have that $\beta\rho c_\mu/4 = C\rho/(4d_\sigma)$ is larger than $\gamma_p\sqrt{\mu_{\text{eff}}\rho}$, $\gamma_q\times 2\mu_{\text{eff}}\rho$ and $\gamma_r(1+\rho c_\mu)$. Therefore, the term in front of $\lambda_1(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_0\mathbf{H}^{1/2})$ is such that

$$\beta(1-\rho c_\mu)+\gamma_p\sqrt{\mu_{\text{eff}}\rho}+2\gamma_q\mu_{\text{eff}}\rho+\gamma_r(1+\rho c_\mu) \leqslant \beta\left(1-\rho c_\mu+3\frac{\rho c_\mu}{4}\right) = \beta\left(1-\frac{\rho c_\mu}{4}\right) \ . \tag{4.166}$$

Using the bounds (4.164), (4.165) and (4.166) in (4.163) gives

$$\begin{aligned}\mathbb{E}_0\left[V(z_1,p_1,q_1,\mathbf{\Sigma}_1,r_1)\right] &\leqslant \left(1-\frac{\delta_\sigma}{4d_\sigma}\right)\|\mathbf{H}^{1/2}z_0\|^2+\beta\left(1-\frac{\rho c_\mu}{4}\right)\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2})\\ &\quad+\gamma_q\times\left(1-\frac{c_\mu}{4}\right)\|\mathbf{H}^{1/2}q_0\|^2\\ &\leqslant \max\left\{1-\frac{\delta_\sigma}{4d_\sigma},1-\frac{\rho c_\mu}{4},1-\frac{c_\mu}{4}\right\}V(z_0,p_0,q_0,\mathbf{\Sigma}_0,r_0)\end{aligned}$$

which ends the proof of (i) since we have assumed that d_σ^{-1} is sufficiently larger than c_μ..
For cases (ii)-(iv), observe that since we have assume $c_1+c_\mu<1$ as well as $1/d_\sigma$ small enough, then we find $\epsilon>0$ such that $c_1+c_\mu\leqslant 1-\epsilon$ and $d_\sigma\geqslant\epsilon$. For this ϵ, using Proposition 4.18(i), we have the existence of $\kappa>0$ such that

$$\mathbb{E}_0\|\mathbf{H}^{1/2}z_1\|^2\leqslant(1-c_1-c_\mu)^{-1}\left(1+\frac{2}{d_\sigma}\right)^2\|\mathbf{H}^{1/2}z_0\|^2+\kappa c_m^2\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)\ . \tag{4.167}$$

Moreover, from (4.149) (4.150), (4.151) and (4.152) we have

$$\mathbb{E}_0[\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2}\right)]\leqslant(1+\rho c_\mu)\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)+3c_1\|\mathbf{H}^{1/2}q_0\|^2\ . \tag{4.168}$$

$$\mathbb{E}_0\|p_1\|\leqslant(1-c_\sigma)\|p_0\|+\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}\rho}\ , \tag{4.169}$$

$$\mathbb{E}_0\|\mathbf{H}^{1/2}q_1\|^2\leqslant\underbrace{2(1-c_1-c_\mu)^{-1}(1-c_c)^2}_{\leqslant 1-c_\mu}\|\mathbf{H}^{1/2}q_0\|^2+2c_c(2-c_c)\mu_{\text{eff}}\rho\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)\ , \tag{4.170}$$

and

$$\mathbb{E}_0[r_1]\leqslant 1+c_\mu\rho+3c_1\|\mathbf{H}^{1/2}q_0\|^2\ . \tag{4.171}$$

Hence, we obtain, from the expression of the potential function (4.25):

$$\begin{aligned}\mathbb{E}_0\left[V(z_1,p_1,q_1,\mathbf{\Sigma}_1,r_1)\right] &= \mathbb{E}_0[\|\mathbf{H}^{1/2}z_1\|^2]+\beta\mathbb{E}_0[\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_1\mathbf{H}^{1/2})]+\gamma_p\mathbb{E}_0[\|p_1\|]\\ &\quad+\gamma_q\mathbb{E}_0[\|\mathbf{H}^{1/2}q_1\|^2]+\gamma_r\mathbb{E}_0[r_1]\\ &\leqslant(1-c_1-c_\mu)^{-1}\left(1+\frac{2}{d_\sigma}\right)^2\|\mathbf{H}^{1/2}z_0\|^2+\kappa c_m^2\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)\\ &\quad+\beta(1+\rho c_\mu)\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)+3\beta c_1\|\mathbf{H}^{1/2}q_0\|^2\\ &\quad+\gamma_p(1-c_\sigma)\|p_0\|+\gamma_p\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}\rho}\\ &\quad+\gamma_q(1-c_\mu)\|\mathbf{H}^{1/2}q_0\|^2+2\gamma_q c_c(2-c_c)\mu_{\text{eff}}\rho\lambda_1\left(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}\right)\\ &\quad+\gamma_r\left(1+c_\mu\rho+3c_1\|\mathbf{H}^{1/2}q_0\|^2\right)\ .\end{aligned} \tag{4.172}$$

For (ii), assume now that $\|\mathbf{\Sigma}_0^{1/2}z_0\|\leqslant M_y\lambda_1(\mathbf{\Sigma}_0)$, $\lambda_1(\mathbf{\Sigma}_0)\leqslant M_{\mathbf{\Sigma}}$ and $\|p_0\|\geqslant M_p$. Hence, using that $\|\mathbf{H}^{1/2}z_0\|^2\leqslant\lambda_1(\mathbf{H})\|z_0\|^2\leqslant\lambda_1(\mathbf{H})\|\mathbf{\Sigma}_0^{1/2}z_0\|^2\leqslant\lambda_1(\mathbf{H})M_y^2\lambda_1(\mathbf{\Sigma}_0)^2\leqslant$

$\lambda_1(\mathbf{H})M_y^2M_{\mathbf{\Sigma}}^2$ and that $\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}_0\mathbf{H}^{1/2}) \leqslant \lambda_1(\mathbf{H})\lambda_1(\mathbf{\Sigma}_0) \leqslant \lambda_1(\mathbf{H})M_{\mathbf{\Sigma}}$, we get:

$$\begin{aligned}\mathbb{E}_0\left[V(z_1,p_1,q_1,\mathbf{\Sigma}_1,r_1)\right] \leqslant & \underbrace{(1-c_1-c_\mu)^{-1}\left(1+\frac{2}{d_\sigma}\right)^2\lambda_1(\mathbf{H})M_y^2M_{\mathbf{\Sigma}}^2}_{\leqslant 2\times 3^2\lambda_1(\mathbf{H})M_z^2M_{\mathbf{\Sigma}}^2 := A_{M_z,M_{\mathbf{\Sigma}},\mathbf{H}}} \\ & + \underbrace{\left(\beta(1+\rho c_\mu)+\kappa c_m^2+2\gamma_q c_c(2-c_c)\mu_{\text{eff}}\rho\right)\lambda_1(\mathbf{H})M_{\mathbf{\Sigma}}}_{\leqslant(\beta(1+\rho)+\kappa+2\gamma_q\mu_{\text{eff}}\rho)\lambda_1(\mathbf{H})M_{\mathbf{\Sigma}} := B^{\beta,\gamma_q}_{M_{\mathbf{\Sigma}},\mathbf{H}}} \\ & + \gamma_p(1-c_\sigma)\|p_0\| + \underbrace{\gamma_p\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}\rho}+\gamma_r(1+c_\mu\rho)}_{\leqslant\gamma_p\sqrt{\mu_{\text{eff}}\rho}+\gamma_r(1+\rho) := C^{\gamma_p,\gamma_r}} \\ & + \left(\gamma_q(1-c_\mu)+3c_1(\beta+\gamma_r)\right)\|\mathbf{H}^{1/2}q_0\|^2 \ .\end{aligned}$$

Then, if we set $M_p > 0$ to be sufficiently large to have that $A_{M_z,M_{\mathbf{\Sigma}},\mathbf{H}} + B^{\beta,\gamma_q}_{M_{\mathbf{\Sigma}},\mathbf{H}} + C^{\gamma_p,\gamma_r} \leqslant \gamma_p c_\sigma M_p/2 \leqslant \gamma_p c_\sigma\|p_0\|/2$, and by using the upper bound (4.165) of the term in front of $\|\mathbf{H}^{1/2}q_0\|^2$ in the previous equation, we obtain:

$$\begin{aligned}\mathbb{E}_0\left[V(z_1,p_1,q_1,\mathbf{\Sigma}_1,r_1)\right] &\leqslant \gamma_p\left(1-\frac{c_\sigma}{2}\right)\|p_0\| + \gamma_q\left(1-\frac{c_\mu}{4}\right)\|\mathbf{H}^{1/2}q_0\|^2 \\ &\leqslant \max\left\{1-\frac{c_\sigma}{2}, 1-\frac{c_\mu}{4}\right\}\times V(z_0,p_0,q_0,\mathbf{\Sigma}_0,r_0)\end{aligned}$$

which proves (ii).
For (iii), suppose $\|\mathbf{\Sigma}_0^{1/2}z_0\| \leqslant M_y\lambda_1(\mathbf{\Sigma}_0)$, $\lambda_1(\mathbf{\Sigma}_0) \leqslant M_{\mathbf{\Sigma}}$, $\|p_0\| \leqslant M_p$ and $\|q_0\| \geqslant M_q$. Then, similarly to case (ii), using (4.172), we have that

$$\begin{aligned}\mathbb{E}_0\left[V(z_1,p_1,q_1,\mathbf{\Sigma}_1,r_1)\right] \leqslant & \underbrace{(1-c_1-c_\mu)^{-1}\left(1+\frac{2}{d_\sigma}\right)^2\lambda_1(\mathbf{H})M_y^2M_{\mathbf{\Sigma}}^2}_{\leqslant A_{M_z,M_{\mathbf{\Sigma}},\mathbf{H}}} \\ & + \underbrace{\left(\beta(1+\rho c_\mu)+\kappa c_m^2+2\gamma_q c_c(2-c_c)\mu_{\text{eff}}\rho\right)\lambda_1(\mathbf{H})M_{\mathbf{\Sigma}}}_{\leqslant B^{\beta,\gamma_q}_{M_{\mathbf{\Sigma}},\mathbf{H}}} \\ & + \underbrace{\gamma_p(1-c_\sigma)M_p}_{\leqslant\gamma_pM_p := C^{\gamma_p}_{M_p}} + \underbrace{\gamma_p\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}\rho}+\gamma_r(1+c_\mu\rho)}_{\leqslant C^{\gamma_p,\gamma_r}} \\ & + \underbrace{\left(\gamma_q(1-c_\mu)+3c_1(\beta+\gamma_r)\right)}_{\leqslant(1-c_\mu/4)\gamma_q}\|\mathbf{H}^{1/2}q_0\|^2 \ .\end{aligned}$$

Then, set $M_q > 0$ such that $A_{M_z,M_{\mathbf{\Sigma}},\mathbf{H}} + B^{\beta,\gamma_q}_{M_{\mathbf{\Sigma}},\mathbf{H}} + C^{\gamma_p}_{M_p} + C^{\gamma_p,\gamma_r} \leqslant \gamma_q c_\mu M_q^2/8$. Since $\gamma_q c_\mu M_q^2/8 \leqslant \gamma_q c_\mu\|q_0\|^2/8 \leqslant \gamma_q c_\mu\|\mathbf{H}^{1/2}q_0\|^2/8$ (since $\|q_0\| = \|\mathbf{H}^{-1/2}\mathbf{H}^{1/2}q_0\| \leqslant \lambda_1(\mathbf{H}^{-1/2})\|\mathbf{H}^{1/2}q_0\| = \|\mathbf{H}^{1/2}q_0\|$) we find

$$\mathbb{E}_0\left[V(z_1,p_1,q_1,\mathbf{\Sigma}_1,r_1)\right] \leqslant \gamma_q\left(1-\frac{c_\mu}{8}\right)\|\mathbf{H}^{1/2}q_0\|^2 \leqslant \left(1-\frac{c_\mu}{8}\right)V(z_0,p_0,q_0,\mathbf{\Sigma}_0,r_0)$$

which proves (iii).

Finally for (iv), assume that $\|\mathbf{\Sigma}_0^{1/2} z_0\| \leqslant M_y \lambda_1(\mathbf{\Sigma}_0)$, $\lambda_1(\mathbf{\Sigma}_0) \leqslant M_{\mathbf{\Sigma}}$, $\|p_0\| \leqslant M_p$, $\|q_0\| \leqslant M_q$, and $r_0 \geqslant M_r$. Then, if we adopt the same notations than for case (iii), we have

$$\mathbb{E}_0\left[V(z_1, p_1, q_1, \mathbf{\Sigma}_1, r_1)\right] \leqslant A_{M_z, M_{\mathbf{\Sigma}}, \mathbf{H}} + B^{\beta,\gamma_q}_{M_{\mathbf{\Sigma}}, \mathbf{H}} + C^{\gamma_p}_{M_p} + C^{\gamma_p,\gamma_r} + \gamma_q \left(1 - \frac{c_\mu}{4}\right) \|\mathbf{H}^{1/2} q_0\|^2 \ .$$

Therefore, we suppose that $M_r > 0$ is chosen sufficiently large so that $A_{M_z, M_{\mathbf{\Sigma}}, \mathbf{H}} + B^{\beta,\gamma_q}_{M_{\mathbf{\Sigma}}, \mathbf{H}} + C^{\gamma_p}_{M_p} + C^{\gamma_p,\gamma_r} \leqslant \gamma_r M_r/2 \leqslant \gamma_r r_0/2$, which gives

$$\begin{aligned}\mathbb{E}_0\left[V(z_1, p_1, q_1, \mathbf{\Sigma}_1, r_1)\right] &\leqslant \gamma_q \left(1 - \frac{c_\mu}{4}\right) \|\mathbf{H}^{1/2} q_0\|^2 + \gamma_r \frac{1}{2} r_0 \\ &\leqslant \max\left\{1 - \frac{c_\mu}{4}, \frac{1}{2}\right\} V(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0) \quad (4.173)\end{aligned}$$

ending the proof. □

4 Discussion

The main contribution of this paper is a proof of geometric ergodicity of Markov chains obtained by normalization of state variables of the covariance matrix adaptation evolution strategy (CMA-ES). Mathematical guarantees of convergence of CMA-ES have been an open question for years and this work constitutes a great advance towards them. While CMA-ES is a complex algorithm, the stochastic processes studied in this paper encompass several variants of the algorithm that are suitable in practice. Specifically, we rely on the rank-mu update of the covariance matrix for the proof, while we include the rank-one update and (only) one of the two evolution paths that are used to update stepsize and covariance matrix in the algorithm. Moreover, hyperparameter settings covered by our results include variants without the rank-one update of the covariance matrix and/or without the evolution paths for stepsize or covariance matrix. Having only a single evolution path motivates a future analysis of variants which use the same path to update stepsize and covariance matrix. Previous theoretical analyses of such variants showed invariance to affine transformations of the search space [15], while empirical studies found no decisive deterioration in their performance [49, 19]. However, these variants preclude the standard fix of a shortcoming when the initial stepsize is too small [70, h_σ in Eq. (46)]. We believe that a convergence analysis similar to the one in this paper is possible for such algorithms.

In our analysis of geometric ergodicity, we use a state-dependent (or multi-step) drift criterion. As the update of the stepsize relies on a momentum term in the evolution path, in practice we have to wait about $1/c_\sigma \approx \sqrt{d}$ iterations until the path becomes large and increases the stepsize. However, the analysis requires only two steps, and the multi-step approach is not needed without cumulation for the stepsize ($c_\sigma = 1$).

In our analysis, we rely on several assumptions and algorithm properties. First, the objective function is unimodal and has ellipsoidal, Lebesgue-negligible level sets. This is essential to obtain explicit and simple forms of the selection function, see Proposition 4.5. Second, the sampling distribution is normal. A standard normal vector has independent coordinates with finite moments and is invariant under rotation. The latter allows a change of variables by affine transformations of the search space as in Proposition 4.6. The former simplifies the computation of several expected quantities. Most notably, the expected inverse stepsize change must be smaller than 1 when the length of the path is sufficiently large, in order bound the normalized mean, see Proposition 4.18 (ii) and (iii).

Our analysis relies on well-chosen hyperparameters. The learning rates c_1 and c_μ for the covariance matrix must be small, as we use a Taylor expansion of its updated eigenvectors in Proposition 4.14. We chose the learning rate c_m of the mean sufficiently small to prevent the mean to jump too far over the optimum, see in particular Proposition 4.18 (ii). Moreover, the variance effective mass μ_{eff} must be large when c_σ is small such that the expected length of the path grows sufficiently fast, see (4.71) and Proposition 4.10. Finally, we set either c_c or c_σ to 1, thereby dropping at least one of the two evolution paths from the original algorithm and we have not included negative weights for the rank-mu update of the covariance matrix in our analysis.

The above conditions are for the most part in line with the working principles of CMA-ES. Some of them have been empirically tested revealing only a modest influence on the algorithm performance [49]. However, some assumptions still could affect the performance of the algorithm. Specifically, the effective selection mass μ_{eff} is assumed to be at least of the magnitude of d/c_σ, see (4.71), and this affects the performance when μ_{eff} is larger than the dimension d. Furthermore, imposing a small learning rate on the mean, $c_m \ll 1$, decreases the convergence speed of the algorithm [50].

A Technical results

Lemma 4.9. Let ξ be a random vector valued in $\mathbb{R}^d$ and $\Lambda \in \mathbb{R}^d$ be a fixed nonzero vector. Suppose that the probability distribution of ξ is invariant to orthogonal rotation. Then, the random variable $\langle \Lambda, \xi \rangle$ follows the same distribution than $\|\Lambda\|[\xi]_1$.

Proof. Let $\mathbf{R}_\Lambda$ be an orthogonal rotation such that $\mathbf{R}_\Lambda e_1 = \Lambda/\|\Lambda\|$, where e_1 is the vector of $\mathbb{R}^d$ with first coordinate equal to one, and the other coordinates to zero. Then $\langle \Lambda, \xi \rangle = \|\Lambda\| \langle \mathbf{R}_\Lambda e_1, \xi \rangle = \|\Lambda\| \langle e_1, \mathbf{R}_\Lambda^\top \xi \rangle = \|\Lambda\| [\mathbf{R}_\Lambda^\top \xi]_1$. However, $\mathbf{R}_\Lambda^\top \xi$ follows the same distribution than ξ such that $\langle \Lambda, \xi \rangle$ follows the same distribution than $\|\Lambda\|[\xi]_1$ □

B Proof of Proposition 4.1

Proof of Proposition 4.1. This is immediate for **Γ**1 and **Γ**4. For **Γ**2, since $\exp(1/d_\sigma) = 1 + 1/d_\sigma + o(1/d_\sigma)$ when $d_\sigma \to +\infty$, suppose that $d_\sigma > 0$ is sufficiently large to have $\exp(1/d_\sigma) \leqslant 1 + 2/d_\sigma$ (this is true for $d_\sigma \geqslant 1$). Then since $c_\sigma \leqslant 1$, we have, for any $u \in \mathbb{R}^d$:

$$\Gamma^1_{\text{CSA}}(u)^{-1} = \exp\left(\frac{c_\sigma}{d_\sigma}\left(1 - \frac{\|u\|}{\mathbb{E}\|\nu_U^d\|}\right)\right) \leqslant \exp\left(\frac{1}{d_\sigma}\right) \leqslant 1 + \frac{2}{d_\sigma} \ ,$$

and

$$\Gamma^2_{\text{CSA}}(u)^{-1} = \exp\left(\frac{c_\sigma}{2d_\sigma}\left(1 - \frac{\|u\|^2}{d}\right)\right) \leqslant \exp\left(\frac{1}{2d_\sigma}\right) \leqslant 1 + \frac{2}{d_\sigma} \ ,$$

so that $\|1/\Gamma^1_{\text{CSA}}\|_\infty = \sup_u |\Gamma^1_{\text{CSA}}(u)^{-1}| \leqslant 1 + \frac{2}{d_\sigma}$, similarly $\|1/\Gamma^2_{\text{CSA}}\|_\infty \leqslant 1 + \frac{2}{d_\sigma}$ and thus **Γ**2 is satisfied. We prove now that Γ^1_{CSA} satisfies **Γ**3. Let $k \in \mathbb{N}$ and $\delta_\sigma > 0$. Let p be a random variable valued in $\mathbb{R}^d$ with finite moments $\mathbb{E}\|p\|^k$ such that $\mathbb{E}\|p\| \geqslant (1 + \delta_p)\mathbb{E}\|\nu_U^d\|$ for some $\delta_p > 0$. Then:

$$\Gamma^1_{\text{CSA}}(p)^{-k} = \exp\left(\frac{kc_\sigma}{d_\sigma}\left(1 - \frac{\|p\|}{\mathbb{E}\|\nu_U^d\|}\right)\right) = 1 + \frac{kc_\sigma}{d_\sigma}\left(1 - \frac{\|p\|}{\mathbb{E}\|\nu_U^d\|}\right) + o\left(\frac{1}{d_\sigma}\right)$$

and for $d_\sigma \geqslant kc_\sigma$, we have, since $1 + x - x^2 \leqslant \exp(x) \leqslant 1 + x + x^2$ for $x \leqslant 1$, that:

$$\gamma^1_{d_\sigma}(p) \coloneqq d_\sigma \left| \Gamma^1_{\text{CSA}}(p)^{-k} - 1 - \frac{kc_\sigma}{d_\sigma}\left(1 - \frac{\|p\|}{\mathbb{E}\|\nu_U^d\|}\right)\right| \leqslant \frac{k^2c_\sigma^2}{d_\sigma}\left(1 - \frac{\|p\|}{\mathbb{E}\|\nu_U^d\|}\right)^2$$

defines a sequence of functions which tend pointwise to zero when d_σ goes to $+\infty$ and which are p-integrable. Yet, for $d_\sigma \geqslant kc_\sigma$, the following domination relation holds: $\left|\gamma^1_{d_\sigma}(p)\right| \leqslant kc_\sigma\left(1 - \frac{\|p\|}{\mathbb{E}\|\nu_U^d\|}\right)^2$ where the RHS is integrable (since p has finite moments) and is independent of d_σ. Therefore, by the dominated convergence theorem, we obtain $\lim_{d_\sigma\to\infty}\mathbb{E}\left[\gamma^1_{d_\sigma}(p)\right] = \mathbb{E}\left[\lim_{d_\sigma\to\infty}\gamma^1_{d_\sigma}(p)\right] = 0$ which gives the following estimation when $d_\sigma \to \infty$

$$\mathbb{E}\left[\Gamma^1_{\text{CSA}}(p)^{-k}\right] = 1 + \frac{kc_\sigma}{d_\sigma}\left(1 - \frac{\mathbb{E}\|p\|}{\mathbb{E}\|\nu_U^d\|}\right) + o(d_\sigma^{-1}) \leqslant 1 - \delta_p\frac{kc_\sigma}{d_\sigma} + o\left(\frac{1}{d_\sigma}\right) \quad .$$

Hence, when $\delta_p > 2c_\sigma^{-1}\delta_\sigma$ we obtain that $\mathbb{E}\left[\Gamma^1_{\text{CSA}}(p)^{-k}\right] \leqslant 1 - 2k\frac{\delta_\sigma}{d_\sigma} + o\left(\frac{1}{d_\sigma}\right)$ and if d_σ is sufficiently large, we obtain that $\mathbb{E}\left[\Gamma^1_{\text{CSA}}(p)^{-k}\right] \leqslant 1 - k\frac{\delta_\sigma}{d_\sigma}$,which proves **Γ3**. For Γ^2_{CSA}, as above let p be a random variable valued in $\mathbb{R}^d$ with finite moments $\mathbb{E}\|p\|^k$ such that $\mathbb{E}\|p\| \geqslant (1+\delta_p)\mathbb{E}\|\nu_U^d\|$ for some $\delta_p > 0$ and let $\xi \sim \nu_U^d$ be independent of p. Then:

$$\begin{aligned}\mathbb{E}[\|p\|^2 - \|\xi\|^2] &= \mathbb{E}\left[(\|p\| - \|\xi\|)(\|p\| + \|\xi\|)\right] \\ &= \mathbb{E}\left[(\|p\| - \|\xi\|) \times \|p\|\right] + \mathbb{E}\left[(\|p\| - \|\xi\|) \times \|\xi\|\right] \quad .\end{aligned}$$

However, by definition of p and ξ, we have that

$$\mathbb{E}\left[(\|p\| - \|\xi\|) \times \|\xi\|\right] \geqslant (1+\delta_p)(\mathbb{E}\|\nu_U^d\|)^2 - \mathbb{E}\|\nu_U^d\|^2$$

is positive if δ_p is sufficiently large. Hence

$$\begin{aligned}\mathbb{E}[\|p\|^2 - \|\xi\|^2] &> \mathbb{E}\left[(\|p\| - \|\xi\|) \times \|p\|\right] = \mathbb{E}\left[(\|p\| - \|\xi\|)\right]\mathbb{E}\|p\| + \operatorname{Cov}\left(\|p\| - \|\xi\|, \|p\|\right) \\ &= \mathbb{E}\left[(\|p\| - \|\xi\|)\right]\mathbb{E}\|p\| + \operatorname{Var}(\|p\|) \geqslant \left(\mathbb{E}\left[\|p\|\right] - \mathbb{E}\left[\|\xi\|\right]\right)\mathbb{E}\|p\| \\ &\geqslant \left((1+\delta_p)\mathbb{E}\|\nu_U^d\| - \mathbb{E}\|\nu_U^d\|\right)(1+\delta_p)\mathbb{E}\|\nu_U^d\| = \delta_p(1+\delta_p)(\mathbb{E}\|\nu_U^d\|)^2 \quad ,\end{aligned}$$

where we used $\operatorname{Cov}\left(\|p\| - \|\xi\|, \|p\|\right) = \operatorname{Var}(\|p\|)$ since p and ξ are independent. Hence we have shown that $\mathbb{E}\|p\|^2 > \mathbb{E}(\|\xi\|^2) + \delta_p(1+\delta_p)(\mathbb{E}\|\nu_U^d\|)^2$. In addition by **N**4 which implies that $\mathbb{E}[\|\xi\|^2] = d$, $\mathbb{E}\|p\|^2 > d + \delta_p(1+\delta_p)(\mathbb{E}\|\nu_U^d\|)^2$. Moreover, by Taylor expansion:

$$\Gamma^2_{\text{CSA}}(p)^{-k} = \exp\left(\frac{kc_\sigma}{2d_\sigma}\left(1 - \frac{\|p\|^2}{d}\right)\right) = 1 + \frac{kc_\sigma}{2d_\sigma}\left(1 - \frac{\|p\|^2}{d}\right) + o\left(\frac{1}{d_\sigma}\right)$$

and for $d_\sigma \geqslant kc_\sigma/2$, using the inequality $|\exp(x) - 1 - x| \leqslant x^2$ for $x \leqslant 1$:

$$\gamma^2_{d_\sigma}(p) \coloneqq d_\sigma\left|\Gamma^2_{\text{CSA}}(p)^{-k} - 1 - \frac{kc_\sigma}{2d_\sigma}\left(1 - \frac{\|p\|^2}{d}\right)\right| \leqslant \frac{k^2c_\sigma^2}{4d_\sigma}\left(1 - \frac{\|p\|^2}{d}\right)^2 \leqslant \frac{kc_\sigma}{2}\left(1 - \frac{\|p\|^2}{d}\right)^2$$

where the LHS defines a sequence of p-integrable functions which tend pointwise to zero for d_σ to $+\infty$, and the RHS a random variable independent of d_σ and which is integrable since p has finite moments. Therefore, by dominated convergence, we find the following estimation:

$$\mathbb{E}\left[\Gamma^2_{\text{CSA}}(p)^{-k}\right] = 1 + \frac{kc_\sigma}{2d_\sigma}\left(1 - \frac{\mathbb{E}\|p\|^2}{d}\right) + o(d_\sigma^{-1}) \leqslant 1 - \frac{\delta_p(1+\delta_p)(\mathbb{E}\|\nu_U^d\|)^2}{d}\frac{kc_\sigma}{d_\sigma} + o(d_\sigma^{-1}) \quad .$$

Thus, when $\delta_p > 2d(\mathbb{E}\|\nu_U^d\|)^{-2}c_\sigma^{-1}\delta_\sigma$ and when d_σ is sufficiently large, we have $\mathbb{E}\left[\Gamma^2_{\mathrm{CSA}}(p)^{-k}\right] \leqslant 1 - k\frac{\delta_\sigma}{d_\sigma}$ which proves that Γ^2_{CSA} satisfies $\mathbf{\Gamma}3$. □

C Proofs in Section 3.3

C.1 *Proof of Theorem 4.6*

Proof. We use the same ideas as in [139, Chapter 1, Section 2]. Let X' be an independent copy of X which is independent of Y and observe that, by bilinearity of the covariance operator:

$$\begin{aligned}\mathrm{Cov}(g(X) - g(X'), h(X,Y) - h(X',Y)) &= \mathrm{Cov}(g(X), h(X,Y)) + \mathrm{Cov}(g(X'), h(X',Y)) \\ &\quad - \mathrm{Cov}(g(X), h(X',Y)) - \mathrm{Cov}(g(X'), h(X,Y)) \ . \quad (4.174)\end{aligned}$$

Since (X',Y) is a copy of (X,Y), we have $\mathrm{Cov}(g(X), h(X,Y)) = \mathrm{Cov}(g(X'), h(X',Y))$. Moreover, since X is independent of (X',Y) and since X' is independent of (X,Y), the last two terms in the RHS of (4.174) are equal to zero. Thus

$$\begin{aligned}\mathrm{Cov}(g(X), h(X,Y)) &= \frac{1}{2}\mathrm{Cov}(g(X) - g(X'), h(X,Y) - h(X',Y)) \\ &= \frac{1}{2}\mathbb{E}\left[(g(X) - g(X'))(h(X,Y) - h(X',Y))\right] \ .\end{aligned}$$

Since g is nondecreasing and $x \mapsto h(x,Y)$ is nonincreasing, then almost surely we obtain $(g(X) - g(X'))(h(X,Y) - h(X',Y)) \leqslant 0$ and thus $\mathbb{E}\left[(g(X) - g(X'))(h(X,Y) - h(X',Y))\right] \leqslant 0$ and in turn $\mathrm{Cov}(g(X), h(X,Y)) \leqslant 0$. When g is increasing and X and $h(X,Y)$ are not almost surely constant, then with positive probability the above inequality is strict and thus $\mathrm{Cov}(g(X), h(X,Y)) < 0$. □

C.2 *Proof of Corollary 4.2*

Proof. Let $x_1, \dots, x_\lambda \in \mathbb{R}$ be such that $x_1 \leqslant \cdots \leqslant x_\lambda$. We have $\sum_{i=1}^\mu w_i x_i = \frac{1}{\mu}\sum_{i=1}^\mu x_i + \sum_{j=1}^\mu \left(w_j - \frac{1}{\mu}\right) x_j$. Let J be an uniform random variable valued in $\{1, \dots, \mu\}$. By assumption on the weight w_i, the function $j \mapsto w_j - 1/\mu$ is nonincreasing on $\{1, \dots, \mu\}$. Moreover, the function $j \mapsto x_j$ is nondecreasing on $\{1, \dots, \mu\}$. Therefore, by Theorem 4.6, $\mathrm{Cov}(w_J - 1/\mu, x_J) \leqslant 0$ and thus

$$\begin{aligned}\mathbb{E}\left[\left(w_J - \frac{1}{\mu}\right) x_J\right] = \frac{1}{\mu}\sum_{j=1}^\mu \left(w_j - \frac{1}{\mu}\right) x_j &\leqslant \mathbb{E}\left[w_J - \frac{1}{\mu}\right]\mathbb{E}\left[x_J\right] \\ &= \underbrace{\frac{1}{\mu}\sum_{j=1}^\mu \left(w_j - \frac{1}{\mu}\right)}_{=0} \frac{1}{\mu}\sum_{k=1}^\mu x_k = 0.\end{aligned}$$

Then, $\frac{1}{\mu}\sum_{j=1}^\mu \left(w_j - \frac{1}{\mu}\right) x_j \leqslant 0$ and thus $\sum_{i=1}^\mu w_i x_i \leqslant \frac{1}{\mu}\sum_{i=1}^\mu x_i$. Applying this inequality to $\xi^{s(1)} \leqslant \cdots \leqslant \xi^{s(\lambda)}$ (which holds almost surely) we obtain $\sum_{i=1}^\mu w_i \xi^{s(i)} \leqslant \frac{1}{\mu}\sum_{i=1}^\mu \xi^{s(i)}$. Taking the expectation we recover the first inequality of (4.37).

For the second inequality, we denote for $k = 1, \dots, \lambda$, $s_k \in \mathfrak{S}_k$ a permutation such that $\xi^{s_k(1)} \leqslant \dots \leqslant \xi^{s_k(k)}$. By property of the expectation of order statistics [35, Section 3.4], we have $(k-i)\mathbb{E}[\xi^{s_k(i)}] + i\mathbb{E}[\xi^{s_k(i+1)}] = k\mathbb{E}[\xi^{s_{k-1}(i)}]$ for $i = 1, \dots, k-1$. We set $\mu_k = \lfloor k/2 \rfloor$ for $k = 2, \dots, \lambda$, so that, if k is even $\mu_{k-1} = \mu_k - 1$, and if k is odd $\mu_{k-1} = \mu_k$.
When k is even and thus $\mu_k = k/2$, by summing for $i = 1, \dots, \mu_k - 1$ the equality $(k-i)\mathbb{E}[\xi^{s_k(i)}] + i\mathbb{E}[\xi^{s_k(i+1)}] = k\mathbb{E}[\xi^{s_{k-1}(i)}]$ and simplifying, we obtain: $(k-1)\sum_{i=1}^{\mu_k - 1}\mathbb{E}\left[\xi^{s_k(i)}\right] + (\mu_k - 1)\mathbb{E}\left[\xi^{s_k(\mu_k)}\right] = k\sum_{i=1}^{\mu_k - 1}\mathbb{E}\left[\xi^{s_{k-1}(i)}\right]$. Since $\mathbb{E}\left[\xi^{s_k(\mu_k)}\right] = 0$ [35, Section 3.4, Corollary 1B], we obtain $\sum_{i=1}^{\mu_k}\mathbb{E}\left[\xi^{s_k(i)}\right] = \frac{k}{k-1}\sum_{i=1}^{\mu_{k-1}}\mathbb{E}\left[\xi^{s_{k-1}(i)}\right]$.
When k is odd, by summing the equality $(k-i)\mathbb{E}[\xi^{s_k(i)}] + i\mathbb{E}[\xi^{s_k(i+1)}] = k\mathbb{E}[\xi^{s_{k-1}(i)}]$ for $i = 1, \dots, \mu_k$ and simplifying, we obtain $(k-1)\sum_{i=1}^{\mu_k}\mathbb{E}\left[\xi^{s_k(i)}\right] + \mu_k\mathbb{E}\left[\xi^{s_k(\mu_k+1)}\right] = k\sum_{i=1}^{\mu_k}\mathbb{E}\left[\xi^{s_{k-1}(i)}\right]$. Since $\mathbb{E}\left[\xi^{s_k(\mu_k+1)}\right] \geqslant 0$ [35, Section 3.4, Corollary 1B], we obtain the following inequality which is then true for k odd and even:

$$\sum_{i=1}^{\mu_k}\mathbb{E}\left[\xi^{s_k(i)}\right] \leqslant \frac{k}{k-1}\sum_{i=1}^{\mu_{k-1}}\mathbb{E}\left[\xi^{s_{k-1}(i)}\right] \ . \tag{4.175}$$

By induction $\sum_{i=1}^{\mu_\lambda}\mathbb{E}\left[\xi^{s_\lambda(i)}\right] \leqslant \frac{\lambda}{\lambda-1} \times \frac{\lambda-1}{\lambda-2}\sum_{i=1}^{\mu_{\lambda-2}}\mathbb{E}\left[\xi^{s_{\lambda-2}(i)}\right] \leqslant \frac{\lambda}{\lambda-1} \times \frac{\lambda-1}{\lambda-2} \times \dots \times \frac{3}{2}\sum_{i=1}^{\mu_2}\mathbb{E}\left[\xi^{s_2(i)}\right]$. Since $\mu_2 = 1$, we obtain then $\sum_{i=1}^{\mu_\lambda}\mathbb{E}\left[\xi^{s_\lambda(i)}\right] \leqslant \frac{\lambda}{2}\mathbb{E}\left[\xi^{s_2(1)}\right]$. On the one hand, since $\mu \leqslant \mu_\lambda \leqslant \lambda/2$, we have almost surely $\frac{1}{\mu}\sum_{i=1}^{\mu}\xi^{s_\lambda(i)} \leqslant \frac{1}{\mu_\lambda}\sum_{i=1}^{\mu_\lambda}\xi^{s_\lambda(i)}$ and on the other hand, $\xi^{s_2(1)} = \min\{\xi^1, \xi^2\}$. This gives the desired result (4.37). □

C.3 *Proof of Lemma 4.2*

Proof of Lemma 4.2. First assume that $G(x) = g \circ F(x)$ for ν_U^d-almost every $x \in \mathbb{R}^d$, where $g\colon F(\mathbb{R}^d) \to G(\mathbb{R}^d)$ is an increasing function. Then, almost surely, $G(U^i) = g(F(U^i))$ for $i = 1, \dots, \lambda$. By definition, $F(U^{s_{F;U}(1)}) \leqslant \dots \leqslant F(U^{s_{F;U}(\lambda)})$, and since g is increasing, $G(U^{s_{F;U}(1)}) \leqslant \dots \leqslant G(U^{s_{F;U}(\lambda)})$. The latter defines the permutation $s_{G;U}$ and therefore almost surely $s_{G;U} = s_{F;U}$.
Conversely, assume that $s_{G;U} = s_{F;U}$ almost surely. Therefore, given $i, j \in \{1, \dots, \lambda\}$ we have almost surely $F(U^i) \leqslant F(U^j) \Leftrightarrow G(U^i) \leqslant G(U^j)$. Since $U^1, \dots, U^\lambda$ are independent and follow ν_U^d, for ν_U^d-almost every $x, y \in \mathbb{R}^d$,

$$F(x) \leqslant F(y) \Leftrightarrow G(x) \leqslant G(y) \ . \tag{4.176}$$

If we denote S the support of ν_U^d, then (4.176) holds for every $x, y \in \mathsf{S}$. Let $x, y \in \mathsf{S}$ such that $F(x) = F(y)$, hence $F(x) \leqslant F(y)$ and thus $G(x) \leqslant G(y)$, and $F(y) \leqslant F(x)$ and thus $G(y) \leqslant G(x)$, so that $G(y) = G(x)$. Hence for $x, y \in \mathsf{S}$

$$F(x) = F(y) \Leftrightarrow G(x) = G(y) \ .$$

This implies that, for every $t \in F(\mathbb{R}^d)$, there exists a unique $g(t) \in G(\mathbb{R}^d)$ such that for $x \in \mathsf{S}$, $F(x) = t \Leftrightarrow G(x) = g(t)$. Moreover, if $t < t'$ are two elements of $F(\mathbb{R}^d)$, then by (4.176), we have $g(t) < g(t')$. This proves that $g\colon F(\mathbb{R}^d) \to G(\mathbb{R}^d)$ defines an increasing function such that, for every $x \in \mathsf{S}$, $G(x) = g \circ F(x)$. □

C.4 Proof of Proposition 4.6

Proof of Proposition 4.6. Since U_1 is distributed according to $(\nu_U^d)^{\otimes\lambda}$, which is invariant under rotation, and since as seen in Corollary 4.3, the matrix $R_0^{\mathbf{H}}$ is orthogonal, then $\hat{U}_1 = (\mathbf{R}_0^{\mathbf{H}})^{-1}U_1$ is distributed according to $(\nu_U^d)^{\otimes\lambda}$ as well. Moreover, since U_1 is independent of θ_0, then $\hat{U}_1$ is independent of $\hat{\theta}_0$. Therefore, in order to obtain the equality in distribution (4.56), we assume without loss of generality that the random input used to update the Markov chain $\{\hat{\theta}_t\}_{t\in\mathbb{N}}$ at the first iteration is $\hat{U}_1$ while we assumed it in the statement of Proposition 4.6 that it is equal to U_1, the random input used to update for θ_0.
Note first that for $i = 1, \dots, \lambda$,

$$\sqrt{\hat{\mathbf{\Sigma}}_0}\hat{U}_1^i = \sqrt{\mathbf{H}^{-1/2}\mathbf{\Sigma}_0\mathbf{H}^{-1/2}}\left(\mathbf{H}^{-1/2}\mathbf{\Sigma}_0\mathbf{H}^{-1/2}\right)^{-1/2}\mathbf{H}^{-1/2}\sqrt{\mathbf{\Sigma}_0}U_1^i = \mathbf{H}^{-1/2}\sqrt{\mathbf{\Sigma}_0}U_1^i \ . \tag{4.177}$$

Additionally, if the permutation $s_1 \in \mathfrak{S}_\lambda$ sorts the $F_1(U_1^i)$, $i = 1, \dots, \lambda$, and if the permutation $\hat{s}_1 \in \mathfrak{S}_\lambda$ sorts the $\hat{F}_1(\hat{U}_1^i)$, $i = 1, \dots, \lambda$, where F_1 is defined by (4.45), and $\hat{F}_1$ by (4.45) when replacing f by $\hat{f}$, then, almost surely, $s_1 = \hat{s}_1$. Indeed, similarly to the proof of Proposition 4.5 and Corollary 4.3, for $i, j \in \{1, \dots, \lambda\}$: $\hat{F}_1(\hat{U}_1^i) < \hat{F}_1(\hat{U}_1^j)$ if and only if $\hat{f}\left(\hat{z}_0 + \sqrt{\hat{\mathbf{\Sigma}}_0}\hat{U}_1^i\right) < \hat{f}\left(\hat{z}_0 + \sqrt{\hat{\mathbf{\Sigma}}_0}\hat{U}_1^j\right)$. By definition of $\hat{z}_0$ and using (4.177) this is equivalent to

$$\hat{f}\left(\mathbf{H}^{-1/2}\left(z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^i\right)\right) < \hat{f}\left(\mathbf{H}^{-1/2}\left(z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^j\right)\right)$$

and by definition of $\hat{f}$ equivalent to

$$\left[\mathbf{H}^{-\frac{1}{2}}\left(z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^i\right)\right]^\top \mathbf{H}\left[\mathbf{H}^{-\frac{1}{2}}\left(z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^i\right)\right] < \left[\mathbf{H}^{-\frac{1}{2}}\left(z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^j\right)\right]^\top \mathbf{H}\mathbf{H}^{-\frac{1}{2}}\left(z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^j\right).$$

which can be simplified as $\left[z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^i\right]^\top \left[z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^i\right] < \left[z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^j\right]^\top \left[z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^j\right]$ which is also equivalent to $f(z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^i) < f(z_0 + \sqrt{\mathbf{\Sigma}_0}U_1^j)$, i.e. $F_1(U_1^i) < F_1(U_1^j)$. Using (4.177), we have also

$$\begin{aligned}\hat{p}_1 &= (1-c_\sigma)\hat{p}_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\sum_{i=1}^{\mu} w_i^m \hat{U}_1^{\hat{s}_1(i)} \\ &= (1-c_\sigma)\left(\mathbf{R}_0^{\mathbf{H}}\right)^{-1} p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\sum_{i=1}^{\mu} w_i^m \left(\mathbf{R}_0^{\mathbf{H}}\right)^{-1} U_1^{s_1(i)} = \left(\mathbf{R}_0^{\mathbf{H}}\right)^{-1} p_1 \ .\end{aligned}$$

Since $\left(\mathbf{R}_0^{\mathbf{H}}\right)^{-1}$ is an orthogonal matrix, we obtain $\|\hat{p}_1\| = \|p_1\|$, and by $\mathbf{\Gamma}4$, $\Gamma_{d_\sigma}(\hat{p}_1) = \Gamma_{d_\sigma}(p_1)$. Moreover, by (4.177),

$$\begin{aligned}\hat{q}_1 &= \hat{r}_0^{-1/2}(1-c_c)\hat{q}_0 + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\sqrt{\hat{\mathbf{\Sigma}}_0}\sum_{i=1}^{\mu} w_i^m \hat{U}_1^{\hat{s}_1(i)} \\ &= r_0^{-1/2}(1-c_c)\mathbf{H}^{-1/2}q_0 + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{H}^{-1/2}\sqrt{\mathbf{\Sigma}_0}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)} = \mathbf{H}^{-1/2}q_1 \ .\end{aligned}$$

We also have, since $\hat{R}(\mathbf{H}^{-1/2}\mathbf{A}\mathbf{H}^{-1/2}) = R(\mathbf{A})$ for any $\mathbf{A} \in \mathcal{S}^d_{++}$, that

$$\begin{aligned}
\hat{r}_1 &= \hat{R}\left((1-c_1-c_\mu)\hat{\mathbf{\Sigma}}_0 + c_1\hat{q}_1\hat{q}_1^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \left[\sqrt{\hat{\mathbf{\Sigma}}_0}\hat{U}_1^{\hat{s}_1(i)}\right]\left[\sqrt{\hat{\mathbf{\Sigma}}_0}\hat{U}_1^{\hat{s}_1(i)}\right]^\top\right) \\
&= \hat{R}\Big((1-c_1-c_\mu)\mathbf{H}^{-1/2}\mathbf{\Sigma}_0\mathbf{H}^{-1/2} + c_1\mathbf{H}^{-1/2}q_1q_1^\top\mathbf{H}^{-1/2} \\
&\qquad\qquad + c_\mu \sum_{i=1}^{\mu} w_i^c \mathbf{H}^{-1/2}\left[\sqrt{\mathbf{\Sigma}_0}U_1^{s_1(i)}\right]\left[\sqrt{\mathbf{\Sigma}_0}U_1^{s_1(i)}\right]^\top \mathbf{H}^{-1/2}\Big) \\
&= R\left((1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_1q_1q_1^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \left[\sqrt{\mathbf{\Sigma}_0}U_1^{s_1(i)}\right]\left[\sqrt{\mathbf{\Sigma}_0}U_1^{s_1(i)}\right]^\top\right) = r_1 \ .
\end{aligned}$$

Likewise, we obtain

$$\begin{aligned}
\hat{\mathbf{\Sigma}}_1 &= \frac{(1-c_1-c_\mu)\hat{\mathbf{\Sigma}}_0 + c_1\hat{q}_1\hat{q}_1^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \left[\sqrt{\hat{\mathbf{\Sigma}}_0}\hat{U}_1^{\hat{s}_1(i)}\right]\left[\sqrt{\hat{\mathbf{\Sigma}}_0}\hat{U}_1^{\hat{s}_1(i)}\right]^\top}{\hat{r}_1} \\
&= \frac{\mathbf{H}^{-1/2}\left((1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_1q_1q_1^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \left[\sqrt{\mathbf{\Sigma}_0}U_1^{s_1(i)}\right]\left[\sqrt{\mathbf{\Sigma}_0}U_1^{s_1(i)}\right]^\top\right)\mathbf{H}^{-1/2}}{r_1} \\
&= \mathbf{H}^{-1/2}\mathbf{\Sigma}_1\mathbf{H}^{-1/2} \ .
\end{aligned}$$

Finally, we have,

$$\hat{z}_1 = \frac{\hat{z}_0 + c_m\sqrt{\hat{\mathbf{\Sigma}}_0}\sum_{i=1}^{\mu} w_i^m \hat{U}_1^{\hat{s}_1(i)}}{\hat{r}_1^{1/2}\Gamma_{d_\sigma}(\hat{p}_1)} = \frac{\mathbf{H}^{-1/2}z_0 + c_m\mathbf{H}^{-1/2}\sqrt{\mathbf{\Sigma}_0}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}}{r_1^{1/2}\Gamma_{d_\sigma}(p_1)} = \mathbf{H}^{-1/2}z_1 \ ,$$

which ends the proof. □

C.5 *Proof of Proposition 4.7*

Proof of Proposition 4.7. Let $\bar{x} \in \mathsf{X}$ and consider a realization $u = (u^1, \dots, u^\lambda) \in (\mathbb{R}^d)^\lambda$ of the random variable U. Assume for the sake of contradiction that the permutation $s_{F_x;u}$ does not tend to $s_{F_{\bar{x}};u}$ when $x \to \bar{x}$. Consider $\bar{s} \in \mathfrak{S}_\lambda \setminus \{s_{F_{\bar{x}};u}\}$ an accumulation point of $\{s_{F_x}\}_{x\in\mathsf{X}}$ at a neighborhood of $\bar{x}$. Indeed, an accumulation point different to $s_{F_{\bar{x}};u}$ exists since $\mathfrak{S}_\lambda$ is a finite set. Then, there exists a sequence $\{x_k\}_{k\in\mathbb{N}}$ of points of X which tends to $\bar{x}$ such that $s_{F_{x_k};u}$ tends to $\bar{s}$. Since $\mathfrak{S}_\lambda$ is a finite set, this implies that there exists a large enough $k_0 \in \mathbb{N}$ such that for every $k \geqslant k_0$, we have $s_{F_{x_k};u} = \bar{s}$. Besides, since $\bar{s} \neq s_{F_{\bar{x}};u}$, then there exist $i, j \in \{1, \dots, \lambda\}$ such that $\bar{s}^{-1}(i) < \bar{s}^{-1}(j)$ and $s^{-1}_{F_{\bar{x}};u}(i) > s^{-1}_{F_{\bar{x}};u}(j)$. Therefore, for $k \geqslant k_0$,

$$F_{x_k}(u^i) < F_{x_k}(u^j) \quad \text{and} \quad F_{\bar{x}}(u^i) > F_{\bar{x}}(u^j) \ . \tag{4.178}$$

Yet, this is a contradiction with the pointwise convergence of F_{x_k} to $F_{\bar{x}}$. Indeed if F_{x_k} converges pointwise to $F_{\bar{x}}$, $F_{x_k}(u^i)$ converges to $F_{\bar{x}}(u^i)$ and $F_{x_k}(u^j)$ converges to $F_{\bar{x}}(u^j)$. Hence, if $F_{\bar{x}}(u^i) > F_{\bar{x}}(u^j)$, then for k large enough $F_{x_k}(u^i) > F_{x_k}(u^j)$ which is in contradiction with (4.178). Therefore we have shown that the permutation $s_{F_x;u}$ tends to $s_{F_{\bar{x}};u}$ when x goes to $\bar{x}$. Hence, since $\mathfrak{S}_\lambda$ is finite, there exists a neighborhood N_u of $\bar{x}$, such that for all $x \in \mathsf{N}_u$ $s_{F_x;u} = s_{F_{\bar{x}};u}$. Let $\phi \in \mathsf{L}^1(\nu_U^{d\lambda})$. The equality of the permutations on N_u implies then that for

all $x \in \mathsf{N}_u$ $\phi\left((u^{s_{F_x;u}(i)})_{i=1,\dots,\lambda}\right) = \phi\left((u^{s_{F_{\bar{x}};u}(i)})_{i=1,\dots,\lambda}\right)$. Thus, for every realization u of the random variable U, the limit

$$\lim_{x\to\bar{x}} \phi\left((u^{s_{F_x;u}(i)})_{i=1,\dots,\lambda}\right) = \phi\left((u^{s_{F_{\bar{x}};u}(i)})_{i=1,\dots,\lambda}\right)$$

holds. Therefore, almost surely, we have $\lim_{x\to\bar{x}} \phi\left((U^{s_{F_x;U}(i)})_{i=1,\dots,\lambda}\right) = \phi\left((U^{s_{F_{\bar{x}};U}(i)})_{i=1,\dots,\lambda}\right)$. Moreover, the following uniform upper bound holds for every $x \in \mathsf{X}$

$$\left|\phi\left((U^{s_{F_x;U}(i)})_{i=1,\dots,\lambda}\right)\right| \leqslant \sum_{s\in\mathfrak{S}_\lambda} \left|\phi\left((U^{s(i)})_{i=1,\dots,\lambda}\right)\right| \ ,$$

and since $s \in \mathfrak{S}_\lambda$ is finite and $\phi \in \mathsf{L}^1(\nu_U^{d\lambda})$, by the dominated convergence theorem, we obtain $\lim_{x\to\bar{x}} \mathbb{E}\left[\phi\left((U^{s_{F_x;U}(i)})_{i=1,\dots,\lambda}\right)\right] = \mathbb{E}\left[\phi\left((U^{s_{F_{\bar{x}};U}(i)})_{i=1,\dots,\lambda}\right)\right]$, which proves the continuity of $x \in \mathsf{X} \mapsto \mathbb{E}[\phi((U^{s_{F_x;U}(i)})_{i=1,\dots,\lambda})]$. □

C.6 *Proof of Proposition 4.8*

Proof of Proposition 4.8. Let $k \in \{1,\dots,d\}$. By (4.17), we have

$$r_1\boldsymbol{\Sigma}_1 = (1-c_1-c_\mu)\boldsymbol{\Sigma}_0 + c_\mu\sqrt{\boldsymbol{\Sigma}_0}\sum_{i=1}^{\mu} w_i^c \left[U_1^{s_1(i)}\right]\left[U_1^{s_1(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_0} + c_1 q_1 q_1^\top \ . \tag{4.179}$$

Since the matrix $c_\mu\sqrt{\boldsymbol{\Sigma}_0}\sum_{i=1}^{\mu} w_i^c \left[U_1^{s_1(i)}\right]\left[U_1^{s_1(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_0} + c_1 q_1 q_1^\top \succeq 0$ by (4.39) we obtain

$$\lambda_k\left(r_1\boldsymbol{\Sigma}_1\right) \geqslant \lambda_k\left((1-c_1-c_\mu)\boldsymbol{\Sigma}_0\right) \ , \tag{4.180}$$

which proves (4.64). Moreover, since $q_1 = r_0^{-1/2}(1-c_c)q_0 + \sqrt{\mu_{\text{eff}}c_c(2-c_c)}\boldsymbol{\Sigma}_0^{1/2}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}$ by Lemma 4.1 applied to $u = r_0^{-1/2}(1-c_c)q_0$ and $v = \sqrt{\mu_{\text{eff}}c_c(2-c_c)}\boldsymbol{\Sigma}_0^{1/2}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}$, we obtain

$$q_1 q_1^\top \preceq 2r_0^{-1}(1-c_c)^2 q_0 q_0^\top + 2c_c(2-c_c)\mu_{\text{eff}}\sqrt{\boldsymbol{\Sigma}_0}\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_0} \ . \tag{4.181}$$

Therefore, since $r_1\boldsymbol{\Sigma}_1 = (1-c_1-c_\mu)\boldsymbol{\Sigma}_0 + c_1 q_1 q_1^\top + c_\mu\sqrt{\boldsymbol{\Sigma}_0}\sum_{i=1}^{\mu} w_i^c \left[U_1^{s_1(i)}\right]\left[U_1^{s_1(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_0}$, using the sum property (4.38), we have

$$\begin{aligned} r_1\boldsymbol{\Sigma}_1 \preceq\ & (1-c_1-c_\mu)\boldsymbol{\Sigma}_0 + c_\mu\sqrt{\boldsymbol{\Sigma}_0}\sum_{i=1}^{\mu} w_i^c \left[U_1^{s_1(i)}\right]\left[U_1^{s_1(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_0} \\ & + 2c_1c_c(2-c_c)\mu_{\text{eff}}\sqrt{\boldsymbol{\Sigma}_0}\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_0} + 2c_1 r_0^{-1}(1-c_c)^2 q_0 q_0^\top \ . \end{aligned} \tag{4.182}$$

Define

$$\mathbf{A}_1 = (1-c_1-c_\mu)\boldsymbol{\Sigma}_0 + c_\mu\sqrt{\boldsymbol{\Sigma}_0}\sum_{i=1}^{\mu} w_i^c \left[U_1^{s_1(i)}\right]\left[U_1^{s_1(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_0} + 2c_1c_c(2-c_c)\mu_{\text{eff}}\sqrt{\boldsymbol{\Sigma}_0}\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]\left[\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_0} \ . \quad (4.183)$$

then (4.182) reads $r_1\boldsymbol{\Sigma}_1 \preceq \mathbf{A}_1 + 2c_1r_0^{-1}(1-c_c)^2 q_0q_0^\top$ such that using (4.40) we find that

$$\lambda_k(r_1\boldsymbol{\Sigma}_1) \leqslant \lambda_k(\mathbf{A}_1 + 2c_1r_0^{-1}(1-c_c)^2 q_0q_0^\top)$$

and using (4.41) we obtain $\lambda_k(r_1\boldsymbol{\Sigma}_1) \leqslant \lambda_k(\mathbf{A}_1) + 2c_1r_0^{-1}(1-c_c)^2\lambda_1\left(q_0q_0^\top\right)$. However, $\lambda_1(q_0q_0^\top) = \|q_0\|^2$, and using (4.63) applied to (4.183) with $\mathbf{A} = (1-c_1-c_\mu)\boldsymbol{\Sigma}_0$, we have

$$\lambda_k(\mathbf{A}_1) \leqslant (1-c_1-c_\mu)\lambda_k(\boldsymbol{\Sigma}_0)(1+\mu d(c_\mu + 2c_1c_c(2-c_c)\mu_{\text{eff}}))\max_{i=1,\dots,\mu}\|U_1^i\|_\infty^2 \quad (4.184)$$

$$\leqslant \lambda_k(\boldsymbol{\Sigma}_0)\times\left(1-c_1-c_\mu+(2c_1\mu_{\text{eff}}+c_\mu)d\mu\max_{i=1,\dots,\mu}\|U_1^i\|_\infty^2\right). \quad (4.185)$$

Indeed, since $\sum_{i=1}^{\mu} w_i^m = 1$, we have $\|\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\|_\infty \leqslant \max_{i=1,\dots,\mu}\|U_1^i\|_\infty$. We also used that $c_c(2-c_c) \leqslant 1$ and $1-c_1-c_\mu \leqslant 1$. Overall we have shown that

$$\lambda_k(r_1\boldsymbol{\Sigma}_1) \leqslant \lambda_k(\boldsymbol{\Sigma}_0)\times\left(1-c_1-c_\mu+(2c_1\mu_{\text{eff}}+c_\mu)d\mu\max_{i=1,\dots,\mu}\|U_1^i\|_\infty^2\right)+2c_1r_0^{-1}(1-c_c)^2\|q_0\|^2.$$

□

C.7 *Proof of Proposition 4.9*

Proof of Proposition 4.9. By the spectral theorem, we suppose without loss of generality that $\boldsymbol{\Sigma}_0 = \text{diag}(\lambda_1,\dots,\lambda_d)$ with $\lambda_i \coloneqq \lambda_i(\boldsymbol{\Sigma}_0)$. We have, by (4.60)

$$\tilde{\boldsymbol{\Sigma}}_1 = (1-c_1-c_\mu)\boldsymbol{\Sigma}_0 + c_1\sqrt{\boldsymbol{\Sigma}_0}\tilde{\mathrm{M}}_1^1\sqrt{\boldsymbol{\Sigma}_0} + c_\mu\sqrt{\boldsymbol{\Sigma}_0}\tilde{\mathrm{M}}_1^\mu\sqrt{\boldsymbol{\Sigma}_0} \quad (4.186)$$

where

$$\tilde{\mathrm{M}}_1^1 \coloneqq \boldsymbol{\Sigma}_0^{-1/2}[q_1][q_1]^\top\boldsymbol{\Sigma}_0^{-1/2} = [\cdot][\cdot]^\top\left(r_0^{-1/2}(1-c_\sigma)\boldsymbol{\Sigma}_0^{-1/2}q_0^c + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right)$$

is the rank-one update, and $\tilde{\mathrm{M}}_1^\mu$ is the matrix defined by (4.61), so-called the rank-mu update. Moreover the notation $[\cdot][\cdot]^\top$ is understood here as $[\cdot][\cdot]^\top(v) = vv^\top$ for any vector $v \in \mathbb{R}^d$. Then we can write

$$c_1\tilde{\mathrm{M}}_1^1 + c_\mu\tilde{\mathrm{M}}_1^\mu = c_\mu\sum_{i=1}^{\mu+1} v_iv_i^\top \quad (4.187)$$

with $v_i = w_i^c U_1^{s_1(i)}$ for $i = 1,\dots,\mu$, and

$$v_{\mu+1} = \sqrt{c_1/c_\mu}\times\left(r_0^{-1/2}(1-c_\sigma)\boldsymbol{\Sigma}_0^{-1/2}q_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\sum_{i=1}^{\mu} w_i^m U_1^{s_1(i)}\right) \ . \quad (4.188)$$

Therefore $\|v_i\|_\infty \leqslant \|U_1\|_\infty$ for $i = 1, \dots, \mu$. Moreover, since $c_1 \leqslant c_\mu/\mu_{\text{eff}}$ by assumption, $r_0^{-1} \leqslant (1 - c_1 - c_\mu)^{-1}$ (see (4.20)), $\|\boldsymbol{\Sigma}_0^{-1/2} q_0\|_\infty \leqslant \|q_0\|_\infty \leqslant \bar{q}$ (since the assumption on the normalization function R ensures that $\lambda_i(\boldsymbol{\Sigma}_0) \geqslant 1$), and $c_\sigma(2 - c_\sigma) \leqslant 1$, we obtain

$$\|v_{\mu+1}\|_\infty \leqslant \sqrt{\mu_{\text{eff}}^{-1}} \times \left((1 - c_1 - c_\mu)^{-1/2} \bar{q} + \sqrt{\mu_{\text{eff}}} \|U_1\|_\infty\right) = (1 - c_1 - c_\mu)^{-1/2} \mu_{\text{eff}}^{-1/2} \bar{q} + \|U_1\|_\infty$$

and thus we have for $i = 1, \dots, \mu + 1$

$$\|v_i\|_\infty \leqslant V_1 \coloneqq (1 - c_1 - c_\mu)^{-1/2} \mu_{\text{eff}}^{-1/2} \bar{q} + \|U_1\|_\infty \ . \tag{4.189}$$

All in all, we have

$$r_1 \boldsymbol{\Sigma}_1 = \tilde{\boldsymbol{\Sigma}}_1 = \mathbf{A}_0 + (1 - c_1 - c_\mu)^{-1} c_\mu \sqrt{\mathbf{A}_0} \sum_{i=1}^{\mu+1} v_i v_i^\top \sqrt{\mathbf{A}_0} \tag{4.190}$$

where $\mathbf{A}_0 = (1 - c_1 - c_\mu)\boldsymbol{\Sigma}_0$. Then, $\boldsymbol{\Sigma}_1$ has the same eigenvectors than $\tilde{\boldsymbol{\Sigma}}_1$, and by (4.68), there exists a random variable $\tilde{\rho}_1 > 0$, which depends polynomially on V_1, hence with finite moments by **N**4, and independent of the initial condition $(z_0, p_0, q_0, \boldsymbol{\Sigma}_0, r_0)$, such that for $i, j = 1, \dots, d$,

$$|[e_i(\boldsymbol{\Sigma}_1)]_j| \leqslant \tilde{\rho}_1 \times \sqrt{\frac{\min\{\lambda_i, \lambda_j\}}{\max\{\lambda_i, \lambda_j\}}} \ . \tag{4.191}$$

Moreover, if we denote $\nu_i = \lambda_i(\tilde{\boldsymbol{\Sigma}}_1)$, then since $e_i(\boldsymbol{\Sigma}_1) = e_i(\tilde{\boldsymbol{\Sigma}}_1)$, we have $\nu_i e_i(\boldsymbol{\Sigma}_1) = \tilde{\boldsymbol{\Sigma}}_1 e_i(\boldsymbol{\Sigma}_1)$. We can moreover use (4.186) to compute $\tilde{\boldsymbol{\Sigma}}_1 e_i(\boldsymbol{\Sigma}_1)$ and then apply the j^{th} coordinate

$$\nu_i [e_i(\boldsymbol{\Sigma}_1)]_j = (1 - c_1 - c_\mu)\lambda_j [e_i(\boldsymbol{\Sigma}_1)]_j + c_1 \sum_{l=1}^d [\tilde{\mathrm{M}}_1^1]_{jl} \sqrt{\lambda_j \lambda_l} [e_i(\boldsymbol{\Sigma}_1)]_l + c_\mu \sum_{l=1}^d [\tilde{\mathrm{M}}_1^\mu]_{jl} \sqrt{\lambda_j \lambda_l} [e_i(\boldsymbol{\Sigma}_1)]_l \ . \tag{4.192}$$

Rearranging, we find, $(\nu_i - (1 - c_1 - c_\mu)\lambda_j)\,[e_i(\boldsymbol{\Sigma}_1)]_j = \sum_{l=1}^d \left(c_1 [\tilde{\mathrm{M}}_1^1]_{jl} + c_\mu [\tilde{\mathrm{M}}_1^\mu]_{jl}\right) \sqrt{\lambda_j \lambda_l} [e_i(\boldsymbol{\Sigma}_1)]_l$. Note that since $\tilde{\boldsymbol{\Sigma}}_1 \succeq (1 - c_1 - c_\mu)\boldsymbol{\Sigma}_0$, by (4.40), $\nu_i \geqslant (1 - c_1 - c_\mu)\lambda_i$.
Thus, when $i < j$ in $\{1, \dots, d\}$, since $\lambda_i \geqslant \lambda_j$, we have

$$\begin{aligned}(\nu_i - (1 - c_1 - c_\mu)\lambda_j)\,|[e_i(\boldsymbol{\Sigma}_1)]_j| &\geqslant (1 - c_1 - c_\mu)(\lambda_i - \lambda_j)\,|[e_i(\boldsymbol{\Sigma}_1)]_j| \\ &= (1 - c_1 - c_\mu)\,|\lambda_i - \lambda_j|\,|[e_i(\boldsymbol{\Sigma}_1)]_j|\end{aligned}$$

which gives $(1 - c_1 - c_\mu)\,|\lambda_i - \lambda_j|\,|[e_i(\boldsymbol{\Sigma}_1)]_j| \leqslant \left|\sum_{l=1}^d \left(c_1 [\tilde{\mathrm{M}}_1^1]_{jl} + c_\mu [\tilde{\mathrm{M}}_1^\mu]_{jl}\right) \sqrt{\lambda_j \lambda_l} [e_i(\boldsymbol{\Sigma}_1)]_l\right|$.
By (4.191), we have $|[e_i(\boldsymbol{\Sigma}_1)]_l| \leqslant \tilde{\rho}_1 \sqrt{\lambda_i/\lambda_l}$. Then, by replacing in the RHS of the above equation, we have then that

$$(1 - c_1 - c_\mu)\,(\lambda_i - \lambda_j)\,|[e_i(\boldsymbol{\Sigma}_1)]_j| \leqslant \sqrt{\lambda_j \lambda_i} \tilde{\rho}_1 \sum_{l=1}^d \left|c_1 [\tilde{\mathrm{M}}_1^1]_{jl} + c_\mu [\tilde{\mathrm{M}}_1^\mu]_{jl}\right| \ .$$

When $j < i$, observe that since $\tilde{\boldsymbol{\Sigma}}_1$ obeys (4.190), taking $\mathbf{A} = \mathbf{A}_0 = (1 - c_1 - c_\mu)\boldsymbol{\Sigma}_0$ with eigenvalues $\xi_i = (1 - c_1 - c_\mu)\lambda_i$, and $u_i = (\sqrt{c_\mu}/\sqrt{1 - c_1 - c_\mu})v_i$ into (4.69) we obtain

$$\nu_i \leqslant \xi_i \times (1 + \mu d c_\mu / (1 - c_1 - c_\mu) V_1^2) = \lambda_i \times (1 - c_1 - c_\mu + \mu d c_\mu V_1^2)$$

and by rearranging and taking the absolute value in (4.192), since $\lambda_j > \lambda_i$, we get

$$(1-c_1-c_\mu)\,(\lambda_j-\lambda_i)\,|[e_i\,(\boldsymbol{\Sigma}_1)]_j| \\ \leqslant c_\mu\mu d V_1^2\lambda_i\,|[e_i\,(\boldsymbol{\Sigma}_1)]_j| + \left|\sum_{l=1}^{d}\left(c_1[\tilde{\mathrm{M}}_1^1]_{jl}+c_\mu[\tilde{\mathrm{M}}_1^\mu]_{jl}\right)\sqrt{\lambda_j\lambda_l}[e_i\,(\boldsymbol{\Sigma}_1)]_l\right| \ .$$

In both cases ($i<j$ and $i>j$), using the expression $c_1\tilde{\mathrm{M}}_1^1+c_\mu\tilde{\mathrm{M}}_1^\mu = c_\mu\sum_{i=1}^{\mu+1}v_iv_i^\top$ from (4.187), we find

$$(1-c_1-c_\mu)\,|\lambda_i-\lambda_j|\,|[e_i\,(\boldsymbol{\Sigma}_1)]_j| \leqslant \sqrt{\lambda_j\lambda_i}\tilde{\rho}_1 c_\mu\mu d V_1^2 + \sqrt{\lambda_j\lambda_i}\tilde{\rho}_1\sum_{l=1}^{d}c_\mu\sum_{k=1}^{\mu+1}|[v^{(k)}]_j[v^{(k)}]_l| \ .$$

Since we have defined V_1 in (4.189) as the upper-bound on $\|v_i\|_\infty$ using $\sum_{l=1}^{d}c_\mu\sum_{k=1}^{\mu+1}|[v^{(k)}]_j[v^{(k)}]_l| \leqslant (\mu+1)d\|v_i\|_\infty^2$ in the previous equation, we have then

$$(1-c_1-c_\mu)\,|\lambda_i-\lambda_j|\,|[e_i\,(\boldsymbol{\Sigma}_1)]_j| \leqslant 2\sqrt{\lambda_j\lambda_i}\tilde{\rho}_1 c_\mu d(\mu+1)V_1^2 \ . \tag{4.193}$$

Moreover, since $c_1 \leqslant \mu_{\mathrm{eff}}^{-1}c_\mu \leqslant c_\mu \leqslant 1/4$, then, if we have $\min\{\lambda_i,\lambda_j\} \leqslant (1-\sqrt{c_\mu})\max\{\lambda_i,\lambda_j\}$, then $\max\{\lambda_i,\lambda_j\}-\max\{\lambda_i,\lambda_j\} \geqslant \sqrt{c_\mu}\max\{\lambda_i,\lambda_j\}$ and we get that

$$(1-c_1-c_\mu)\,|\lambda_i-\lambda_j| = (1-c_1-c_\mu)(\max\{\lambda_i,\lambda_j\}-\min\{\lambda_i,\lambda_j\}) \\ \geqslant (1-c_1-c_\mu)\sqrt{c_\mu}\max\{\lambda_i,\lambda_j\} \geqslant \frac{1}{2}\times\sqrt{c_\mu}\max\{\lambda_i,\lambda_j\} \tag{4.194}$$

and thus by dividing by $\sqrt{c_\mu}\max\{\lambda_i,\lambda_j\}/2$ in (4.193) we deduce that

$$|[e_i\,(\boldsymbol{\Sigma}_1)]_j| \leqslant \frac{2\sqrt{\lambda_j\lambda_i}\tilde{\rho}_1 c_\mu d(\mu+1)V_1^2}{\sqrt{c_\mu}\max\{\lambda_i,\lambda_j\}/2} = \frac{\sqrt{c_\mu}4d(\mu+1)V_1^2\tilde{\rho}_1\sqrt{\lambda_j\lambda_i}}{\max\{\lambda_i,\lambda_j\}}.$$

Since $\sqrt{\lambda_j\lambda_i} = \sqrt{\max\{\lambda_i,\lambda_j\}\min\{\lambda_i,\lambda_j\}}$ we obtain that $|[e_i\,(\boldsymbol{\Sigma}_1)]_j| \leqslant \sqrt{c_\mu}\times 4d(\mu+1)V_1^2\tilde{\rho}_1\times\sqrt{\frac{\min\{\lambda_i,\lambda_j\}}{\max\{\lambda_i,\lambda_j\}}}$. From this equation, we obtain (4.70) for $i<j$ and $i>j$ by taking $\rho_1 = 4d(\mu+1)V_1^2\tilde{\rho}_1$. Note that ρ_1 has finite moments since the random variables V_1 (by **N**4) and $\tilde{\rho}_1$ both have finite moments (as recalled earlier in the proof). □

C.8 *Proof of Lemma 4.3*

Proof of Lemma 4.3. Given $l\in\mathbb{R}^d$, consider an orthonormal basis $(l/\|l\|, e_2,\ldots,e_d)$, so that:

$$\mathbb{E}\left\|\sum_{i=1}^{\mu}w_i^m U^{s_{F_l;U}(i)}\right\| = \mathbb{E}\sqrt{\left|\sum_{i=1}^{\mu}w_i^m\left\langle\frac{l}{\|l\|},U^{s_{F_l;U}(i)}\right\rangle\right|^2+\left\|\sum_{i=1}^{\mu}w_i^m\left\langle e_j,U^{s_{F_l;U}(i)}\right\rangle_{j=2,\ldots,d}\right\|^2}$$

Using Jensen's inequality (since the square root function is concave) to obtain $\sqrt{|a|^2+\|b\|^2} = \sqrt{2}\sqrt{|a|^2/2+\|b\|^2/2} \geqslant \sqrt{2}/2\sqrt{|a|^2}+\sqrt{2}/2\sqrt{\|b\|^2} = \sqrt{2}/2|a|+\sqrt{2}/2\|b\|$ for $a\in\mathbb{R}$ and

$b \in \mathbb{R}^d$ we obtain

$$\mathbb{E}\left\|\sum_{i=1}^{\mu} w_i^m U^{s_{F_l;U}(i)}\right\| \geqslant \frac{\sqrt{2}}{2}\left|\mathbb{E}\sum_{i=1}^{\mu} w_i^m \left\langle \frac{l}{\|l\|}, U^{s_{F_l;U}(i)}\right\rangle\right| + \frac{\sqrt{2}}{2}\mathbb{E}\left\|\sum_{i=1}^{\mu} w_i^m \left\langle e_j, U^{s_{F_l;U}(i)}\right\rangle_{j=2,\ldots,d}\right\|. \tag{4.195}$$

Since each U_1^i for $i = 1, \ldots \lambda$, follows a standard multivariate normal distribution, its coordinates in any orthonormal basis are independent (i.e., **N**5 implies **N**2 and **N**3 in any basis) and follow a standard multivariate normal distribution. Thus, for $j = 2, \ldots, d$, the random vector $\langle e_j, U_1^i\rangle_{i=1,\ldots,\lambda}$ is independent of $\langle l, U_1^i\rangle_{i=1,\ldots,\lambda}$. Since the permuation $s_{F_l;U_1}$ is entirely determined by $\langle l, U_1^i\rangle_{i=1,\ldots,\lambda}$, $\langle e_j, U_1^i\rangle_{i=1,\ldots,\lambda}$ is independent of $s_{F_l;U_1}$, and then for each $i = 1, \ldots, \lambda$ and each $j = 2, \ldots, d$, $\langle e_j, U_1^{s_{F_l;U_1}(i)}\rangle$ follows a standard multivariate normal distribution and $\{\langle e_j, U_1^{s_{F_l;U_1}(i)}\rangle \mid i = 1, \ldots, \lambda, j = 2, \ldots, d\}$ are independent. Therefore we can compute the RHS expectation of (4.195):

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{\mu} w_i^m \underbrace{\left\langle e_j, U_1^{s_{F_l;U_1}(i)}\right\rangle_{j=2,\ldots,d}}_{\sim \nu_U^{d-1}}\right\| &= \int \|\mathbf{w}_m^\top \xi\|\, \nu_U^{\mu(d-1)}(\mathrm{d}\xi) \\ &= \|\mathbf{w}_m\| \int \|\xi\| \nu_U^{\mu(d-1)}(\mathrm{d}\xi) = \sqrt{\frac{1}{\mu_{\text{eff}}}}\mathbb{E}\|\nu_U^{d-1}\| \ . \end{aligned} \tag{4.196}$$

The random variables $\{\langle l/\|l\|, U_1^i\rangle, i = 1, \ldots, \lambda\}$ are independent and follow a standard multivariate normal distribution. Besides, the permutation $s_{F_l;U}$ is determined by ranking those random variables, i.e.,

$$\langle l/\|l\|, U_1^{s_{F_l;U}(1)}\rangle \leqslant \langle l/\|l\|, U_1^{s_{F_l;U}(2)}\rangle \leqslant \ldots \leqslant \langle l/\|l\|, U_1^{s_{F_l;U}(\lambda)}\rangle$$

such that $\langle l/\|l\|, U_1^{s_{F_l;U}(1)}\rangle$ represent order statistics of standard multivariate normal distributions. Hence, denoting $\xi^{s_\lambda(i)}$ the i^{th} out of λ order statistics of standard multivariate normal distribution and using Corollary 4.2

$$\mathbb{E}\left[\sum_{i=1}^{\mu} w_i^m \left\langle \frac{l}{\|l\|}, U_1^{s_{F_l;U_1}(i)}\right\rangle\right] = \mathbb{E}\left[\sum_{i=1}^{\mu} w_i^m \xi^{s_\lambda(i)}\right] \leqslant \frac{1}{\mu}\mathbb{E}\left[\xi^{s_\lambda(i)}\right] \leqslant \mathbb{E}\left[\xi^{s_2(1)}\right] = \frac{2}{\sqrt{2}} I_{\nu_U^1} < 0 \ .$$

We obtain (4.73) by applying the previous inequality as well as (4.196) to (4.195).

□

Chapter 5

Linear convergence of CMA-ES on ellipsoidal problems and learning of second-order information

Comments on Chapter 5: This chapter is the final step of our proof of linear convergence for CMA-ES. It is the sequel of the results in Chapters 2 and 4: as the normalized Markov chain is geometrically ergodic, we use limit theorems to finally prove the convergence. We will remind in this chapter the definition of the normalized Markov chain, however we refer to the previous chapters for the previously established results that are used in the proofs of this chapter.

Abstract

In this chapter, we prove the linear convergence of CMA-ES on ellipsoidal minimization problems. It is based on that the normalization of the states of CMA-ES may define an ergodic Markov chain. Moreover, we prove that, for some hyperparameter settings, the convergence rate is the same for every ellipsoidal objective functions and that the covariance matrix approximates the inverse Hessian. The latter heavily relies on the affine-invariance property of CMA-ES, that we will recall and formalize in this chapter.

1 Assumptions for the theoretical analysis of CMA-ES 192
1.1 Assumptions on the objective function 192
1.2 Assumption on the sampling distribution 192
1.3 Assumption on the weights 192
1.4 Assumptions on the stepsize change 192
2 Invariance to rotation and to affine transformation 193
3 Linear convergence of CMA-ES on ellipsoidal objective functions 194
3.1 Integrability with respect to the invariant probability measure 195
3.2 Linear behavior 200
3.3 Equal convergence rates for different ellipsoidal functions 202
3.4 Positivity of the convergence rate 202
4 Limit distribution of the covariance matrix 207
5 Learning of the inverse Hessian of convex-quadratic functions 209

1 Assumptions for the theoretical analysis of CMA-ES

We start by summarizing the different assumptions required to proceed with our analysis.

1.1 Assumptions on the objective function

We consider in this chapter objective functions that are spherical (**F**1) and ellipsoidal (**F**2). The matrix $\mathbf{H}$ in Assumption **F**2 is called the quasi-Hessian matrix of f.

F1. *The objective function f is spherical, i.e., it is an increasing transformation of $x \in \mathbb{R}^d \mapsto x^\top x$.*

F2. *The objective function f is ellipsoidal, i.e., it is an increasing transformation of $x \in \mathbb{R}^d \mapsto x^\top \mathbf{H} x$ where $\mathbf{H} \in \mathcal{S}^d_{++}$ is chosen without loss of generality such that $\lambda_{\min}(\mathbf{H}) = 1$.*

1.2 Assumption on the sampling distribution

From now on, we suppose that the sampling distribution ν_U^d is the standard multivariate normal distribution $\mathcal{N}(0, \mathbf{I}_d)$.

1.3 Assumption on the weights

We recall here that the weights $\mathbf{w}_m = (w_1^m, \dots, w_\mu^m)$ and $\mathbf{w}_c = (w_1^c, \dots, w_\mu^c)$ are nonincreasing and sum to one..

W1. *We have $w_1^\star \geqslant \cdots \geqslant w_\mu^\star > 0$ and $\sum_{i=1}^\mu w_i^\star = 1$.*

1.4 Assumptions on the stepsize change

We summarize now the assumptions on the stepsize change Γ that were used throughout the thesis and on which we rely in this final chapter.

Γ1. *The stepsize change $\Gamma \colon \mathbb{R}^d \to \mathbb{R}_{++}$ is a locally Lipschitz map, differentiable in every nonzero vectors of $\mathbb{R}^d$, such that $\Gamma(0) < 1$, and $\liminf \Gamma(p) \geqslant (1 - c_c)^{-1}$ when $\|p\| \to +\infty$.*

Γ2. *When $d_\sigma > 0$ is large enough, we have $\|1/\Gamma\|_\infty \leqslant 1 + 2/d_\sigma$.*

Γ3. *Given $\delta_\sigma > 0$ and $k \in \mathbb{N}$, if $\delta_p > 0$ and $d_\sigma \geqslant 1$ are large enough, then, for any r.v. $p \in \mathbb{R}^d$ such that $\mathbb{E}\|p\| \geqslant (1 + \delta_p)\mathbb{E}\|\nu_U^d\|$, we have*

$$\mathbb{E}\left[\Gamma(p)^{-k}\right] \leqslant 1 - k\frac{\delta_\sigma}{d_\sigma}.$$

Γ4. *The stepsize change Γ is invariant by orthogonal rotation, that is, for every $\mathbf{P} \in \mathsf{O}(d)$ and $p \in \mathbb{R}^d$, we have $\Gamma(\mathbf{P}p) = \Gamma(p)$.*

Γ5. *There exists a polynomial function $P \colon \mathbb{R}^d \to \mathbb{R}$ such that $\log \Gamma \leqslant P$.*

Assumption **Γ**1 was required to obtain the irreducibility of a normalized Markov chain underlying CMA-ES in Chapter 2, while **Γ**2 and **Γ**3 are used to prove its geometrical ergodicity in Chapter 4. Assumption **Γ**4 is a necessary condition for the affine invariance of CMA-ES, and assumption **Γ**5 gives useful integrability properties. Note that these assumptions are satisfied by the stepsize change function Γ^2_{CSA}, that we use to obtain a positive convergence rate. We recall here that

$$\Gamma^2_{\text{CSA}}(p) = \exp\left(\frac{c_\sigma}{2d_\sigma}\left(\frac{\|p\|^2}{d} - 1\right)\right) \quad . \tag{5.1}$$

2 Invariance to rotation and to affine transformation

One key property of CMA-ES that we use in this paper to describe the distribution of the covariance matrix (including the learning of second-order information) is the invariance to rotations of the search space.

Theorem 5.1 (Rotation-invariance of CMA-ES)**.** Suppose that the stepsize change Γ is invariant to rotation, i.e., $\mathbf{\Gamma}4$ is true. Let $\phi_0 = (m_0, p_0^\sigma, p_0^c, \sigma_0, \mathbf{C}_0) \in \mathbb{R}^{3d} \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d$, and $\mathbf{U} = \{U_{t+1}\}_{t\in\mathbb{N}}$ an i.i.d. process independent of ϕ_0, with $U_1 \sim \nu_U^{d\lambda}$, where $\nu_U^{d\lambda} = \mathcal{N}(0, \mathbf{I}_{d\lambda})$ is the standard multivariate normal distribution.
Denote $\{\phi_t\}_{t\in\mathbb{N}} = \{(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$ the iterates produced by CMA-ES, when minimizing an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ with initialization ϕ_0 and random inputs $\mathbf{U}$.
Let $\mathbf{R} \in \mathsf{O}_d(\mathbb{R})$ be an orthogonal matrix, and $b \in \mathbb{R}^d$, and define

$$\begin{aligned} \Phi_{\mathbf{R},b}\colon\ & \mathbb{R}^{3d} \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \to \mathbb{R}^{3d} \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \\ & (m, p^\sigma, p^c, \sigma, \mathbf{C}) \mapsto \left(\mathbf{R}^\top(m-b), \mathbf{R}^\top p^\sigma, \mathbf{R}^\top p^c, \sigma, \mathbf{R}^\top \mathbf{C}\mathbf{R}\right) \end{aligned} \tag{5.2}$$

and

$$\begin{aligned} \Psi_{\mathbf{R},b}\colon\ & \mathbb{R}^{d\lambda} \times \mathbb{R}^{3d} \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \to \mathbb{R}^{d\lambda} \\ & (u^1, \ldots, u^\lambda, m, p^\sigma, p^c, \sigma, \mathbf{C}) \mapsto \left(\mathbf{R}^\top u^i\right)_{i=1,\ldots,\lambda}. \end{aligned} \tag{5.3}$$

Define then $\hat{\phi}_0 = (\hat{m}_0, \hat{p}_0^\sigma, \hat{p}_0^c, \hat{\sigma}_0, \hat{\mathbf{C}}_0) = \Phi_{\mathbf{R},b}(\phi_0)$, and $\hat{\mathbf{U}} = \{\hat{U}_{t+1}\}_{t\in\mathbb{N}} = \{\Psi_{\mathbf{R},b}(U_{t+1}, \phi_t)\}_{t\in\mathbb{N}}$. Then, $\hat{\mathbf{U}}$ is an i.i.d. process, independent of $\hat{\phi}_0$, with $\hat{U}_1 \sim \nu_U^{d\lambda}$.
Moreover, if we denote $\{\hat{\phi}_t\}_{t\in\mathbb{N}} = \{(\hat{m}_t, \hat{p}_t^\sigma, \hat{p}_t^c, \hat{\sigma}_t, \hat{\mathbf{C}}_t)\}_{t\in\mathbb{N}}$ the iterates produced by CMA-ES, when minimizing the function $f(\mathbf{R}\cdot + b)$, with initialization $\hat{\phi}_0$, and random inputs $\hat{\mathbf{U}}$, then, for every $t \in \mathbb{N}$,

$$(\hat{m}_t, \hat{p}_t^\sigma, \hat{p}_t^c, \hat{\sigma}_t, \hat{\mathbf{C}}_t) = \Phi_{\mathbf{R},b}(\phi_t)\ . \tag{5.4}$$

Furthermore when $c_\sigma = 1$, CMA-ES is known to be affine-invariant. This property is central to prove that the convergence rate of CMA-ES is the same for every ellipsoidal objective function. Besides, we rely on the affine-invariance to show that the covariance matrix approximates the Hessian matrix $\mathbf{H}$ of f.

Theorem 5.2 (Affine-invariance of CMA-ES [70, 15])**.** Suppose that the stepsize change Γ is invariant to rotation ($\mathbf{\Gamma}4$) and that $c_\sigma = 1$. Let $\phi_0 = (m_0, p_0^c, \sigma_0, \mathbf{C}_0) \in \mathbb{R}^{2d} \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d$, and $\mathbf{U} = \{U_{t+1}\}_{t\in\mathbb{N}}$ an i.i.d. process independent of ϕ_0, with $U_1 \sim \nu_U^{d\lambda}$.
Denote $\{\phi_t\}_{t\in\mathbb{N}} = \{(m_t, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$ the iterates produced by CMA-ES, when minimizing an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ with initialization ϕ_0 and random inputs $\mathbf{U}$.
Let $\mathbf{B} \in \mathsf{GL}_d(\mathbb{R})$ be an orthogonal matrix, and $b \in \mathbb{R}^d$, and define

$$\begin{aligned} \Phi_{\mathbf{B},b}\colon\ & \mathbb{R}^{2d} \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \to \mathbb{R}^{2d} \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \\ & (m, p^c, \sigma, \mathbf{C}) \mapsto \left(\mathbf{B}^{-1}(m-b), \mathbf{B}^{-1} p^c, \sigma, \mathbf{B}^{-1}\mathbf{C}\mathbf{B}^{-\top}\right) \end{aligned} \tag{5.5}$$

and

$$\begin{aligned} \Psi_{\mathbf{B},b}\colon\ & \mathbb{R}^{d\lambda} \times \mathbb{R}^{2d} \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \to \mathbb{R}^{d\lambda} \\ & (u^1, \ldots, u^\lambda, m, p^c, \sigma, \mathbf{C}) \mapsto \left((\mathbf{B}^{-1}\mathbf{C}\mathbf{B}^{-\top})^{-1/2}\mathbf{B}^{-1}\mathbf{C}^{1/2}u^i\right)_{i=1,\ldots,\lambda}. \end{aligned} \tag{5.6}$$

Define then $\hat{\phi}_0 = (\hat{m}_0, \hat{p}_0^c, \hat{\sigma}_0, \hat{\mathbf{C}}_0) = \Phi_{\mathbf{B},b}(\phi_0)$, and $\hat{\mathbf{U}} = \{\hat{U}_{t+1}\}_{t\in\mathbb{N}} = \{\Psi_{\mathbf{B},b}(U_{t+1}, \phi_t)\}_{t\in\mathbb{N}}$. Then, $\hat{\mathbf{U}}$ is an i.i.d. process, independent of $\hat{\phi}_0$, with $\hat{U}_1 \sim \nu_U^{d\lambda}$.
Moreover, if we denote $\{\hat{\phi}_t\}_{t\in\mathbb{N}} = \{(\hat{m}_t, \hat{p}_t^c, \hat{\sigma}_t, \hat{\mathbf{C}}_t)\}_{t\in\mathbb{N}}$ the iterates produced by CMA-ES, when minimizing the function $f(\mathbf{B}\cdot + b)$, with initialization $\hat{\phi}_0$, and random inputs $\hat{\mathbf{U}}$, then, for every $t \in \mathbb{N}$,

$$\hat{\phi}_t = \Phi_{\mathbf{B},b}(\phi_t) \ . \tag{5.7}$$

3 Linear convergence of CMA-ES on ellipsoidal objective functions

In this section we prove that CMA-ES converges linearly when the objective function is ellipsoidal. More precisely, we show that the distance of the mean m_t to the minimum of the objective function behaves asymptotically geometrically, when the hyperparameters of the algorithm are well-chosen. Moreover, we underline, at least when the stepsize is updated without cumulation, that the geometric rate does not vary from one ellipsoidal objective function to an other. Finally, we prove that, when the stepsize is updated via (5.1), the mean indeed converges to the minimum. These properties are exposed in Theorem 5.3.

The proof of Theorem 5.3 is divided in several parts. It relies on the geometric ergodicity of a normalized Markov chain, see Theorem 4.3 in the previous chapter. We find in Section 3.1 several functions that are consequently integrable with respect to the invariant probability measure of this Markov chain. These results are useful in Section 3.2 for the proof of the linear behavior of the algorithm. In Section 3.3, we use the affine-invariance of CMA-ES variants (see Theorem 5.2) to deduce that the convergence rate does not depend on the choice of the objective function, as long as it is ellipsoidal. Finally, in Section 3.4, we prove the convergence of the algorithm for a specific stepsize change function.

Theorem 5.3. Consider an ellipsoidal objective function f via **F**2 with quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}_{++}^d$. Suppose that the stepsize change function Γ and the weights $\mathbf{w}_m$, $\mathbf{w}_c$ satisfy **Γ**1-**Γ**5 and **W**1, respectively. Suppose that $\mu \leqslant \lambda/2$ and μ_{eff} is sufficiently large, $c_1 + c_\mu < 1$, d_σ^{-1} is sufficiently larger than $c_\mu > 0$ and $c_m^{3/2} > 0$, and sufficiently smaller than c_m, and both $2c_c$ and c_σ are larger than c_μ. Besides assume that $c_c = 1$ or $c_\sigma = 1$.
Then, CMA-ES behaves linearly, i.e., there exists a real constant $\mathrm{CR} \in \mathbb{R}$, such that for every initialization $(m_0, \sigma_0, \mathbf{C}_0, p_0^\sigma, p_0^c) \in \mathbb{R}^d \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \times \mathbb{R}^d \times \mathbb{R}^d$ of CMA-ES, we have

$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty} \frac{1}{T} \left[\log \frac{\sigma_T}{\sigma_0} + \frac{1}{2} \log \frac{\lambda_d(\mathbf{H}^{-1/2}\mathbf{C}_T\mathbf{H}^{-1/2})}{\lambda_d(\mathbf{H}^{-1/2}\mathbf{C}_0\mathbf{H}^{-1/2})} \right] = -\mathrm{CR} \tag{5.8}$$

and

$$\lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|} \right] = \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2} \log \frac{\lambda_d(\mathbf{H}^{-1/2}\mathbf{C}_{t+1}\mathbf{H}^{-1/2})}{\lambda_d(\mathbf{H}^{-1/2}\mathbf{C}_t\mathbf{H}^{-1/2})} \right] = -\mathrm{CR}. \tag{5.9}$$

If $c_\sigma = 1$, CMA-ES is affine-invariant and the value of CR does not depend on the quasi-Hessian $\mathbf{H} \in \mathcal{S}_{++}^d$.
If $\Gamma = \Gamma_{\text{CSA}}^2$, we have $\mathrm{CR} > 0$ and thus for every $t \in \mathbb{N}$, $\|m_t - x^*\| \leqslant C \times \rho^t$ for some $C > 0$ and $\rho = \exp(-\mathrm{CR}) \in (0, 1)$.

The proof of Theorem 5.3 is decomposed into Propositions 5.2 to 5.4 stated and proven below.

3.1 *Integrability with respect to the invariant probability measure*

In this section, we establish the integrability of some functions with respect to the invariant probability measure π of the normalized Markov chain underlying CMA-ES. We remind that the normalized Markov chain is defined for $t \in \mathbb{N}$ by

$$z_t = \frac{m_t - x^*}{\sqrt{R(\mathbf{C}_t)}\sigma_t} \quad , \quad p_t = p_t^\sigma \quad , \quad q_t = \frac{p_t^c}{\sqrt{R(\mathbf{C}_{t-1})}} \quad , \quad \mathbf{\Sigma}_t = \frac{\mathbf{C}_t}{R(\mathbf{C}_t)} \quad , \quad r_t = \frac{R(\mathbf{C}_t)}{R(\mathbf{C}_{t-1})} \tag{5.10}$$

where the normalization function R satisfy one of the following

R1. *R is positively homogeneous with $R(\mathbf{I}_d) = 1$,*

R2. $R(\cdot) = \lambda_{\min}(\cdot)$.

When minimizing an ellipsoidal objective function with quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}_{++}^d$ with $\lambda_{\min}(\mathbf{H}) = 1$, we assume

R3. $R(\cdot) = \lambda_{\min}(\mathbf{H}^{1/2} \times \cdot \times \mathbf{H}^{1/2})$.

Equivalently, see Proposition 2.2 in Chapter 2, when the objective function f is scaling-invariant and the normalization function R is positively homogeneous, the Markov chain (5.10) follows the update equations

$$\begin{aligned}
z_{t+1} &= \frac{z_t + c_m\sqrt{\mathbf{\Sigma}_t}\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}}{\sqrt{r_{t+1}}\Gamma_{d_\sigma}(p_{t+1})} \\
p_{t+1} &= (1-c_\sigma)p_t + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \\
q_{t+1} &= r_t^{-1/2}(1-c_c)q_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{\Sigma}_t^{1/2}\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \\
\mathbf{\Sigma}_{t+1} &= \frac{\tilde{\mathbf{\Sigma}}_{t+1}}{r_{t+1}} \\
r_{t+1} &= R(\tilde{\mathbf{\Sigma}}_{t+1}) \ .
\end{aligned} \tag{5.11}$$

It was proven in the previous chapter (more precisely in Theorem 4.3) that (5.10) may define a geometrically ergodic Markov chain when the objective function is ellipsoidal and satisfies **F**2 for some quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}_{++}^d$. Consequently, the normalized Markov chain converges to an invariant probability measure π. Moreover, we proved that the potential function V given by

$$V(z, p, q, \mathbf{\Sigma}, r) = \|\mathbf{H}^{1/2}z\|^2 + \lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}) + \|p\| + \|\mathbf{H}^{1/2}q\|^2 + r \tag{5.12}$$

defines an integrable function with respect to the measure π. Further results of integrability with respect to π would then allow the application of the Law of Large Numbers to deduce the linear behavior of CMA-ES. This section is devoted to these results. First, we recall the following criterion for integrability with respect to π that we use in the sequel.

Proposition 5.1 ([110, Contraposition of Theorem 14.3.3(i)])**.** Suppose that P is a positive recurrent and aperiodic Markov chain on the measurable space $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ with unique invariant probability measure π. Consider $f : \mathsf{X} \to \mathbb{R}_+$ a measurable function. If there exists $x \in \mathsf{X}$ such that the sequence $\{\int_{\mathsf{X}} f(y)P^t(x, \mathrm{d}y)\}_{t\in\mathbb{N}}$ is bounded, then $f \in \mathsf{L}^1(\pi)$.

We show now that the log-stepsize change function is integrable inder the invariant measure of the normalized chain when it is geometrically ergodic.

Lemma 5.1. Suppose that the stepsize change satisfies **Γ**5.

(i) When $c_\sigma \neq 1$, assume that the process $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ ($\{(z_t, p_t, \mathbf{\Sigma}_t)\}$ if $c_c = 1$) defined via (5.11) is geometrically ergodic with invariant probability measure π. Then, the function $(z, p, q, \mathbf{\Sigma}, r) \mapsto \log \Gamma(p)$ ($(z, p, \mathbf{\Sigma}) \mapsto \log \Gamma(p)$ if $c_c = 1$) is integrable with respect to π.

(ii) If $c_\sigma = 1$, assume that the process $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ ($\{(z_t, \mathbf{\Sigma}_t)\}$ if $c_c = 1$) defined via (5.11) is geometrically ergodic with invariant probability measure π. Then, when $(z_0, q_0, \mathbf{\Sigma}_0, r_0)$ ($(z_0, \mathbf{\Sigma}_0)$ if $c_c = 1$) is distributed with respect to π,

$$\mathbb{E}[|\log \Gamma(p_1)|] = \mathbb{E}_{U_1 \sim \nu_U^{d\lambda}}[|\log \Gamma(\sqrt{\mu_{\text{eff}}}\mathbf{w}_m^\top U_1^{s_1})|] < +\infty \ .$$

Proof. By **Γ**5 and Proposition 5.1, it is sufficient to show that the quantities for $t \geqslant 1$

$$\mathbb{E}\|p_t\|^n$$

define a bounded sequence, for any fixed $n \in \mathbb{N}$. Observe that, for every $t \geqslant 1$, we have, by (5.11):

$$\|p_{t+1}\|^n \leqslant (1 - c_\sigma)^n \|p_t\|^n + \sum_{k=0}^{n-1} \binom{n}{k} (1 - c_\sigma)^k (c_\sigma(2 - c_\sigma)\mu_{\text{eff}})^{\frac{n-k}{2}} \|U_{t+1}\|_\infty^{n-k} \|p_t\|^k.$$

Denote $\rho_{n-k} \coloneqq \mathbb{E}\|U_{t+1}\|_\infty^{n-k}$ for $k = 0, \ldots, n-1$. Since the normal distribution has finite moments, we have $\rho_{n-k} < +\infty$. Then,

$$\mathbb{E}\|p_{t+1}\|^n \leqslant (1 - c_\sigma)^n \mathbb{E}\|p_t\|^n + \sum_{k=0}^{n-1} \binom{n}{k} (1 - c_\sigma)^k (c_\sigma(2 - c_\sigma)\mu_{\text{eff}})^{\frac{n-k}{2}} \rho_{n-k} \mathbb{E}\|p_t\|^k.$$

If $c_\sigma = 1$, we obtain

$$\mathbb{E}\|p_{t+1}\|^n \leqslant \mu_{\text{eff}}^{n/2} \rho_n < +\infty$$

and if $(1 - c_\sigma)^n < 1$, by induction on $n \in \mathbb{N}$, the sequence $\{\mathbb{E}\|p_{t+1}\|^n\}_{t\geqslant 1}$ is bounded. □

In the next proposition, we prove that the integrability (with respect to π) of the potential function (5.12) implies the integrability (with respect to π) of the maximum eigenvalue of the covariance matrix.

Lemma 5.2. Assume that the process $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ if $c_\sigma = 1$, $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\geqslant 1}$ if $c_c = 1$, $\{(z_t, \mathbf{\Sigma}_t)\}_{t\geqslant 1}$ if $c_\sigma = c_c = 1$, respectively, defined by (5.11), is a geometrically ergodic Markov chain with unique invariant probability measure π. Suppose that the corresponding potential function V defined in Theorem 4.3 is π-integrable. Then, when $(z_0, q_0, \mathbf{\Sigma}_0, r_0)$, respectively $(z_0, p_0, \mathbf{\Sigma}_0)$, $(z_0, \mathbf{\Sigma}_0)$, is distributed with respect to π, then $\mathbb{E}[\lambda_{\max}(\mathbf{\Sigma}_0)] < +\infty$.

Proof. Indeed, for every $(z, p, q, \mathbf{\Sigma}, r) \in \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$, we have $\lambda_{\max}(\mathbf{\Sigma}) \leqslant V(z, q, \mathbf{\Sigma}, r)$ (or $V(z, p, \mathbf{\Sigma})$, $V(z, \mathbf{\Sigma})$). □

Before stating more integrability results, we first give the following inequality, which follows

from [141, Eq. (31), p. 26]. For $u, v \in \mathbb{R}^d$ and $i = 1, \ldots, \mu$, we have

$$\mathbb{1}\left\{\|(\mathbf{w}_m)^\top u\| \leqslant 1\right\} \nu_U^d\left(u - (w_i^m)^{-1} v\right) \leqslant (2\pi)^{-d/2} \mathbb{1}\left\{\|(\mathbf{w}_m)^\top u\| \leqslant 1\right\} \exp\left(-\frac{(w_i^m)^{-2}}{2}\right) \nu_U^d(u). \tag{5.13}$$

We are now interested in the integrability of $\log \|q\|$ where q is the normalized evolution path for the rank-one update.

Lemma 5.3. Consider a normalization function R satisfying **R**3 for some matrix $\mathbf{H} \in \mathcal{S}_{++}^d$.

(i) When $c_c \neq 1$, assume that the process $\{(z_t, p_t, q_t, \boldsymbol{\Sigma}_t, r_t)\}_{t \geqslant 1}$ ($\{(z_t, q_t, \boldsymbol{\Sigma}_t, r_t)\}$ if $c_\sigma = 1$) defined via (5.11) is geometrically ergodic with invariant probability measure π and that the corresponding potential function V defined in Theorem 4.3 is π-integrable. Then, the function $(z, p, q, \boldsymbol{\Sigma}, r) \mapsto \log \|q\|$ ($(z, q, \boldsymbol{\Sigma}, r) \mapsto \log \|q\|$ if $c_\sigma = 1$) is integrable with respect to π.

(ii) If $c_c = 1$, assume that the process $\{(z_t, p_t, \boldsymbol{\Sigma}_t)\}_{t \geqslant 1}$ ($\{(z_t, \boldsymbol{\Sigma}_t)\}$ if $c_\sigma = 1$) defined via (5.11) is geometrically ergodic with invariant probability measure π and that the corresponding potential function V defined in Theorem 4.3 is π-integrable. Then, when $(z_0, p_0, \boldsymbol{\Sigma}_0)$ ($(z_0, \boldsymbol{\Sigma}_0)$ if $c_\sigma = 1$) is distributed with respect to π,

$$\mathbb{E}[|\log \|q_1\||] = \mathbb{E}_{U_1 \sim \nu_U^{d\lambda}}\left[\left|\log\left\|\sqrt{\mu_{\text{eff}} \boldsymbol{\Sigma}_1} \mathbf{w}_m^\top U_1^{s_1}\right\|\right|\right] < +\infty \ .$$

Proof. We assume for the sake of conciseness that $c_\sigma \neq 1$. Moreover, we set $V(z, p, q, \boldsymbol{\Sigma}, r) = \|\mathbf{H} z\|^2 + \beta \lambda_1(\mathbf{H}^{1/2} \boldsymbol{\Sigma} \mathbf{H}^{1/2}) + \gamma_p \|p\| + \gamma_q \|H^{1/2} q\|^2 + \gamma_r r$. However, the following proof would also apply to the case $c_\sigma = 1$, up to dropping the variable p in the lines below. Let $(z, p, q, \boldsymbol{\Sigma}, r) \in \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$. If $c_c \neq 1$, note that, when $\|q\| \geqslant 1$, then $\log \|q\| \leqslant \|q\| \leqslant V(z, p, q, \boldsymbol{\Sigma}, r)$. Therefore, since V is π-integrable by assumption, it is sufficient to show that the map

$$(z, p, q, \boldsymbol{\Sigma}, r) \mapsto |\log \|q\|| \, \mathbb{1}\{\|q\| \leqslant 1\}$$

is π-integrable. If $c_c = 1$, we observe that

$$\mathbb{E}_{U_1 \sim \nu_U^{d\lambda}}\left[\left(\log\left\|\sqrt{\mu_{\text{eff}} \boldsymbol{\Sigma}_1} \mathbf{w}_m^\top U_1^{s_1}\right\|\right)^+\right] \leqslant \frac{\lambda!}{(\lambda - \mu)!} \mathbb{E}_{U_1 \sim \nu_U^{d\lambda}}\left[\left\|\sqrt{\mu_{\text{eff}} \boldsymbol{\Sigma}_1} \mathbf{w}_m^\top U_1\right\|\right] < +\infty \ .$$

In the rest of proof, we do not make any assumption on the value of c_c. In the case that $c_c = 1$, we denote $q_{t+1} = \sqrt{\mu_{\text{eff}} \boldsymbol{\Sigma}_{t+1}} \mathbf{w}_m^\top U_{t+1}^{s_{t+1}}$. We have, by (5.11):

$$\begin{aligned}
&\mathbb{E}_t\left[|\log \|q_{t+1}\| \mathbb{1}\{\|q_{t+1}\| \leqslant 1\}|\right] \\
&\quad = -\mathbb{E}_t\left[\log\left\|(1 - c_c) r_t^{-1/2} q_t + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}} \sqrt{\boldsymbol{\Sigma}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}\right\|\right. \\
&\qquad \left. \times \mathbb{1}\left\{\left\|(1 - c_c) r_t^{-1/2} q_t + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}} \sqrt{\boldsymbol{\Sigma}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}\right\| \leqslant 1\right\}\right] .
\end{aligned}$$

However, by Lemma 2.4, we have then

$$\begin{aligned}
&\mathbb{E}_t\left[\left|\log\|q_{t+1}\|\mathbb{1}\left\{\|q_{t+1}\|\leqslant 1\right\}\right|\right] \\
&\quad\leqslant -\frac{\lambda!}{(\lambda-\mu)!}\mathbb{E}_t\left[\log\left\|(1-c_c)r_t^{-1/2}q_t+\sqrt{c_c(2-c_c)\mu_{\text{eff}}}\sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu}w_i^m U_{t+1}^i\right\|\right. \\
&\qquad\left.\times\,\mathbb{1}\left\{\left\|(1-c_c)r_t^{-1/2}q_t+\sqrt{c_c(2-c_c)\mu_{\text{eff}}}\sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu}w_i^m U_{t+1}^i\right\|\leqslant 1\right\}\right].
\end{aligned}$$

However, by a change of variable, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\left|\log\|q_{t+1}\|\mathbb{1}\left\{\|q_{t+1}\|\leqslant 1\right\}\right|\right] \\
&\quad\leqslant \frac{\lambda!}{(\lambda-\mu)!}\int\left[\log\left\|(1-c_c)r_t^{-1/2}q_t+\sqrt{c_c(2-c_c)\mu_{\text{eff}}}\sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu}w_i^m u_i\right\|\right]^{-}\mathrm{d}\nu_U^{d\mu}(u_1,\ldots,u_\mu) \\
&\quad\leqslant \frac{\lambda!}{(\lambda-\mu)!}\times\left[\frac{1}{2}\left|\log(c_c(2-c_c)\mu_{\text{eff}})\right|+\frac{1}{2}\max\left\{\left|\log\lambda_1\left(\boldsymbol{\Sigma}_t\right)\right|,\left|\log\lambda_d\left(\boldsymbol{\Sigma}_t\right)\right|\right\}\right] \\
&\quad+\frac{\lambda!}{(\lambda-\mu)!}\int\left[\log\left\|\sum_{i=1}^{\mu}w_i^m u_i\right\|\right]^{-}\tau_{(1-c_c)\left(\sqrt{r_t c_c(2-c_c)\mu_{\text{eff}}}w_i^m\right)^{-1}\boldsymbol{\Sigma}_t^{-1/2}q_t}\nu_U^{d\lambda}(\mathrm{d}u)
\end{aligned}$$

where we denote $\tau_q\nu_U^{d\lambda}(\mathrm{d}u)=\nu_U^{d\lambda}(\mathrm{d}(u_1-q),\ldots,\mathrm{d}(u_\lambda-q))$. Hence, by (5.13), we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\log\|q_{t+1}\|\mathbb{1}\left\{\|q_{t+1}\|\leqslant 1\right\}\right|\right] \\
&\quad\leqslant \frac{\lambda!}{(\lambda-\mu)!}\times\left[\frac{1}{2}\left|\log(c_c(2-c_c)\mu_{\text{eff}})\right|+\frac{1}{2}\mathbb{E}\max\left\{\left|\log\lambda_1\left(\boldsymbol{\Sigma}_t\right)\right|,\left|\log\lambda_d\left(\boldsymbol{\Sigma}_t\right)\right|\right\}\right. \\
&\qquad\left.+\mathbb{E}\left[\log\left\|\sum_{i=1}^{\mu}w_i^m U_{t+1}^i\right\|\right]^{-}\right],
\end{aligned}$$

where we have $\mathbb{E}\left[\log\left\|\sum_{i=1}^{\mu}w_i^m U_{t+1}^i\right\|\right]^{-}<+\infty$. However, by **R**3, $\lambda_1\left(\boldsymbol{\Sigma}_t\right)\geqslant\lambda_d\left(\boldsymbol{\Sigma}_t\right)\geqslant\lambda_d(\mathbf{H}^{1/2}\boldsymbol{\Sigma}_t\mathbf{H}^{1/2})/\lambda_d(\mathbf{H})=R(\boldsymbol{\Sigma}_t)/\lambda_d(\mathbf{H})=1/\lambda_d(\mathbf{H})$, hence $-\log\lambda_d(\mathbf{H})\leqslant\log\lambda_d\left(\boldsymbol{\Sigma}_t\right)\leqslant\log\lambda_1\left(\boldsymbol{\Sigma}_t\right)\leqslant\lambda_1\left(\boldsymbol{\Sigma}_t\right)$. Then, by Lemma 5.2, we obtain:

$$\begin{aligned}
&\limsup_{t\to\infty}\mathbb{E}\left[\left|\log\|q_{t+1}\|\mathbb{1}\left\{\|q_{t+1}\|\leqslant 1\right\}\right|\right] \\
&\qquad\leqslant \frac{\lambda!}{(\lambda-\mu)!}\times\left[\frac{1}{2}\left|\log(c_c(2-c_c)\mu_{\text{eff}})\right|+\frac{1}{2}\mathbb{E}_\pi\log\lambda_1(\boldsymbol{\Sigma})\right. \\
&\qquad\qquad\left.+\left|\log\lambda_d(\mathbf{H})\right|+\mathbb{E}\left[\log\left\|\sum_{i=1}^{\mu}w_i^m U_1^i\right\|\right]^{-}\right]<+\infty,
\end{aligned}$$

which ends the proof after applying Proposition 5.1. □

Next, we study the normalization factor r with respect to π.

Lemma 5.4. Suppose that the objective function f is ellipsoidal and satisfies **F**2 with quasi-Hessian matrix $\mathbf{H}\in\mathcal{S}_{++}^d$, and that the normalization function R satisfies **R**3 with $\mathbf{H}$.

(i) When $c_c \neq 1$, assume that the process $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ ($\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}$ if $c_\sigma = 1$) defined via (5.11) is geometrically ergodic with invariant probability measure π. Then, the function $(z, p, q, \mathbf{\Sigma}, r) \mapsto \log r$ ($(z, q, \mathbf{\Sigma}, r) \mapsto \log r$ if $c_\sigma = 1$) is integrable with respect to π.

(ii) If $c_c = 1$, assume that the process $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\geqslant 1}$ ($\{(z_t, \mathbf{\Sigma}_t)\}$ if $c_\sigma = 1$) defined via (5.11) is geometrically ergodic with invariant probability measure π. Then, when $(z_0, p_0, \mathbf{\Sigma}_0)$ ($(z_0, \mathbf{\Sigma}_0)$ if $c_\sigma = 1$) is distributed with respect to π,

$$\mathbb{E}[|\log \|r_1\|\|] < +\infty \ .$$

Proof. By Corollary 4.4, we have

$$1-c_1-c_\mu \leqslant r_{t+1} \leqslant 1-c_1-c_\mu+(2c_1\mu_{\text{eff}}+c_\mu)d\mu\|U_{t+1}\|^2+2c_1 \underbrace{r_t^{-1}}_{\leqslant(1-c_1-c_\mu)^{-1}} \underbrace{(1-c_c)^2}_{\leqslant 1} \underbrace{\|\mathbf{H}^{1/2}q_t\|^2}_{\leqslant\lambda_1(\mathbf{H})\|q_t\|^2}$$

which, by taking the logarithm, yields to

$$|\log r_{t+1}| \leqslant -2\log(1-c_1-c_\mu) + |\log(2c_1\mu_{\text{eff}}+c_\mu)d\mu| + \left|\log\|U_{t+1}\|^2\right| + |\log(2c_1\lambda_1(\mathbf{H}))| + 2\,|\log\|q_t\||| \ .$$

By Lemma 5.3, this implies that

$$\limsup_{t\to\infty} \mathbb{E}|\log r_{t+1}| \leqslant -2\log(1-c_1-c_\mu) + |\log(2c_1\mu_{\text{eff}}+c_\mu)d\mu| + \mathbb{E}\left|\log\|U_1\|^2\right| + |\log(2c_1\lambda_1(\mathbf{H}))| + 2\mathbb{E}_\pi\,|\log\|q\||| < \infty$$

which ends the proof, by Proposition 5.1. □

Finally, we prove that $\log\|\mathbf{B}z\|$ defines an integrable (with respect to π) function, where z is the normalized mean and $\mathbf{B}$ is an arbitrary fixed matrix.

Lemma 5.5. Let $\mathbf{B} \in \mathsf{GL}_d(\mathbb{R})$. Suppose that the normalization function R satisfies **R3** for some matrix $\mathbf{H} \in \mathcal{S}^d_{++}$. Assume that the normalized process $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$, respectively $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ if $c_\sigma = 1$, $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ if $c_c = 1$, $\{(z_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ if $c_\sigma = c_c = 1$, defined via (5.11) is geometrically ergodic with unique invariant probability measure π, and that the function V in Theorem 4.3 is π-integrable. Then, the function $(z, p, q, \mathbf{\Sigma}, r) \mapsto \log\|\mathbf{B}z\|$, respectively $(z, q, \mathbf{\Sigma}, r) \mapsto \log\|\mathbf{B}z\|$, $(z, p, \mathbf{\Sigma}) \mapsto \log\|\mathbf{B}z\|$, and $(z, \mathbf{\Sigma}) \mapsto \log\|\mathbf{B}z\|$, is integrable with respect to the measure π.

Proof. For the sake of simplicity, we suppose that c_σ, c_c are not 1. However, assuming otherwise would not change the following lines, except for dropping the variables p, q or r. Moreover, we set $V(z, p, q, \mathbf{\Sigma}, r) = \|\mathbf{H}z\|^2 + \beta\lambda_1(\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}) + \gamma_p\|p\| + \gamma_q\|H^{1/2}q\|^2 + \gamma_r r$. Suppose first that $\mathbf{B} = \mathbf{I}_d$. Let $(z, p, q, \mathbf{\Sigma}, r) \in \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$. As in the proof of Lemma 5.3, note that, when $\|z\| \geqslant 1$, then $\log\|z\| \leqslant V(z, p, q, \mathbf{\Sigma}, r)$. Therefore, since V is π-integrable by assumption, it is sufficient to show that the map

$$(z, p, q, \mathbf{\Sigma}, r) \mapsto |\log\|z\|| \, \mathbb{1}\{\|z\| \leqslant 1\}$$

is π-integrable. Moreover, we have, by (5.11):

$$\mathbb{E}_t\left[\log\|z_{t+1}\|\mathbb{1}\left\{\|z_{t+1}\|\leqslant 1\right\}\right] \leqslant \frac{1}{2}\mathbb{E}_t|\log r_{t+1}| + \mathbb{E}_t|\log\Gamma(p_{t+1})| + \mathbb{E}_t\left[\log\left\|z_t + c_m\sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}\right\| \times \mathbb{1}\left\{\left\|z_t + c_m\sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)}\right\| \leqslant 1\right\}\right] .$$

Then, similarly to in the proof of Lemma 5.3, we obtain

$$\mathbb{E}\left[\log\|z_{t+1}\|\mathbb{1}\left\{\|z_{t+1}\|\leqslant 1\right\}\right] \leqslant \frac{1}{2}\mathbb{E}|\log r_{t+1}| + \mathbb{E}|\log\Gamma(p_{t+1})| + \frac{\lambda!}{(\lambda-\mu)!}\times\left[|\log(c_m)| + \frac{1}{2}\mathbb{E}\log\lambda_1\left(\boldsymbol{\Sigma}_t\right) + \mathbb{E}\left[\log\left\|\sum_{i=1}^{\mu} w_i^m U_{t+1}^i\right\|\right]^-\right] .$$

Then, by Lemmas 5.1, 5.2 and 5.4, we have

$$\limsup_{t\to\infty}\mathbb{E}\left[\log\|z_{t+1}\|\mathbb{1}\left\{\|z_{t+1}\|\leqslant 1\right\}\right] \leqslant \frac{1}{2}\mathbb{E}_\pi|\log r| + \mathbb{E}_\pi|\log\Gamma(p)| + \frac{\lambda!}{(\lambda-\mu)!}\times\left[|\log(c_m)| + \frac{1}{2}\mathbb{E}_\pi\log\lambda_1(\boldsymbol{\Sigma}) + \mathbb{E}\left[\log\left\|\sum_{i=1}^{\mu} w_i^m U_1^i\right\|\right]^-\right] < +\infty .$$

This ends the proof, by Proposition 5.1.
When $\mathbf{B}\neq\mathbf{I}_d$, observe that

$$\log\|z\| - \log\left\|\mathbf{B}^{-1}\right\| \leqslant \log\|\mathbf{B}z\| \leqslant \log\|z\| + \log\|\mathbf{B}\| ,$$

which proves the π-integrability of $(z,p,q,\boldsymbol{\Sigma},r)\mapsto\log\|\mathbf{B}z\|$, since $(z,p,q,\boldsymbol{\Sigma},r)\mapsto\log\|z\|$ is π-integrable. □

3.2 *Linear behavior*

We can now prove the first part of Theorem 5.3, i.e., the linear behavior of CMA-ES when the objective function is ellipsoidal.

Proposition 5.2. Let $\mathbf{B}\in\mathrm{GL}_d(\mathbb{R})$. Consider an ellipsoidal objective function f via **F2** for some quasi-Hessian matrix $\mathbf{H}\in\mathcal{S}_{++}^d$. Suppose moreover that the stepsize change function Γ and the weights $\mathbf{w}_m$, $\mathbf{w}_c$ satisfy **Γ1**-**Γ5** and **W1**, respectively. Assume that $\mu\leqslant\lambda/2$ and μ_{eff} is sufficiently large, $c_1+c_\mu<1$, d_σ^{-1} is sufficiently larger than $c_\mu>0$ and $c_m^{3/2}>0$, and sufficiently smaller than c_m, and both $2c_c$ and c_σ. Besides, assume that $c_c=1$ or $c_\sigma=1$.
Then, there exists a real constant $\mathrm{CR}\in\mathbb{R}$, such that for every initialization $(m_0,\sigma_0,\mathbf{C}_0,p_0^\sigma,p_0^c)\in\mathbb{R}^d\times\mathbb{R}_{++}\times\mathcal{S}_{++}^d\times\mathbb{R}^d\times\mathbb{R}^d$, we have

$$\lim_{T\to\infty}\frac{1}{T}\log\frac{\|\mathbf{B}(m_T-x^*)\|}{\|\mathbf{B}(m_0-x^*)\|} = \lim_{T\to\infty}\frac{1}{T}\left[\log\frac{\sigma_T}{\sigma_0} + \frac{1}{2}\log\frac{\lambda_d(\mathbf{H}^{1/2}\mathbf{C}_T\mathbf{H}^{1/2})}{\lambda_d(\mathbf{H}^{1/2}\mathbf{C}_0\mathbf{H}^{1/2})}\right] = -\mathrm{CR} \quad (5.14)$$

and

$$\lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|\mathbf{B}(m_{t+1}-x^*)\|}{\|\mathbf{B}(m_t-x^*)\|}\right] = \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2}\log \frac{\lambda_d(\mathbf{H}^{1/2}\mathbf{C}_{t+1}\mathbf{H}^{1/2})}{\lambda_d(\mathbf{H}^{1/2}\mathbf{C}_t\mathbf{H}^{1/2})}\right] = -\mathrm{CR}. \tag{5.15}$$

Proof. Consider the normalization $R(\cdot)$ satisfying **R**3 with the matrix $\mathbf{H}$. By Theorem 4.3, the Markov chain $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ if $c_\sigma = 1$, $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\geqslant 1}$ if $c_c = 1$, and $\{(z_t, \mathbf{\Sigma}_t)\}_{t\geqslant 1}$ if $c_\sigma = c_c = 1$, respectively, defined by (5.10), is geometrically ergodic, and admits a unique invariant probability measure π such that the function V defined by, respectively,

$$V(z, q, \mathbf{\Sigma}, r) = \|z\|^2 + \lambda_1(\mathbf{\Sigma}) + \|q\|^2 + r \quad \text{for } (z, q, \mathbf{\Sigma}, r) \in \mathbb{R}^{2d} \times R^{-1}(\{1\}) \times \mathbb{R}_{++} \ ,$$

$$V(z, p, \mathbf{\Sigma}) = \|z\|^2 + \lambda_1(\mathbf{\Sigma}) + \|p\| \quad \text{for } (z, p, \mathbf{\Sigma}) \in \mathbb{R}^{2d} \times R^{-1}(\{1\}) \ ,$$

or

$$V(z, \mathbf{\Sigma}) = \|z\|^2 + \lambda_1(\mathbf{\Sigma})(z, \mathbf{\Sigma}) \in \mathbb{R}^d \times R^{-1}(\{1\}) \ ,$$

is integrable with respect to π. Moreover, note that

$$\begin{aligned}
\frac{1}{T}\log \frac{\|\mathbf{B}(m_T - x^*)\|}{\|\mathbf{B}(m_0 - x^*)\|} &= \frac{1}{T}\sum_{t=0}^{T-1} \log\|\mathbf{B}(m_{t+1} - x^*)\| - \log\|\mathbf{B}(m_t - x^*)\| \\
&= \frac{1}{T}\sum_{t=0}^{T-1} \log\|\mathbf{B}z_{t+1}\| - \log\|\mathbf{B}z_t\| + \log\frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2}\log\frac{R(\mathbf{C}_{t+1})}{R(\mathbf{C}_t)} \\
&= \frac{1}{T}\sum_{t=0}^{T-1} \log\|\mathbf{B}z_{t+1}\| - \log\|\mathbf{B}z_t\| + \log\Gamma(p_{t+1}) + \frac{1}{2}\log r_{t+1} \ .
\end{aligned}$$

Likewise, for $t \geqslant 1$,

$$\mathbb{E}\left[\log \frac{\|\mathbf{B}(m_{t+1} - x^*)\|}{\|\mathbf{B}(m_t - x^*)\|}\right] = \mathbb{E}\left[\log\|\mathbf{B}z_{t+1}\| - \log\|\mathbf{B}z_t\| + \log\Gamma(p_{t+1}) + \frac{1}{2}\log r_{t+1}\right]$$

Then, by Lemmas 5.1 to 5.5, and by the Law of Large Numbers [110, Theorem 17.0.1], we have:

$$\lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|\mathbf{B}(m_{t+1} - x^*)\|}{\|\mathbf{B}(m_t - x^*)\|}\right] = \lim_{T\to\infty} \frac{1}{T}\log \frac{\|\mathbf{B}(m_T - x^*)\|}{\|\mathbf{B}(m_0 - x^*)\|} = \mathbb{E}_\pi[\log\Gamma(p)] + \frac{1}{2}\mathbb{E}_\pi[\log r] \ .$$

Note that, when $c_\sigma = 1$, respectively $c_c = 1$, we used the abuse of notation

$$\mathbb{E}_\pi[\log\Gamma(p)] = \mathbb{E}_{(z_0, q_0, \mathbf{\Sigma}_0, r_0)\sim\pi}[\log\Gamma(p_1)] \ ,$$

respectively,

$$\mathbb{E}_\pi[\log r] = \mathbb{E}_{(z_0, p_0, \mathbf{\Sigma}_0)\sim\pi}[\log r_1] \ .$$

□

3.3 *Equal convergence rates for different ellipsoidal functions*

We deduce in the following proposition from the linear behavior of CMA-ES (Proposition 5.2) and the affine-invariance of several CMA-ES variants (Theorem 5.2) that the rate of convergence of CMA-ES remains invariant for every ellispoidal objective function.

Proposition 5.3. Under the conditions of Proposition 5.2, if moreoover $c_\sigma = 1$, then the value of CR is independent of the quasi-Hessian matrix $\mathbf{H}$.

Proof. Assume $c_\sigma = 1$. Let $\{\phi_t\}_{t\in\mathbb{N}} = \{(m_t, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$ be the process produced by CMA-ES when minimizing a function f satisfying **F**2, with random inputs $\mathbf{U} = \{U_{t+1}\}_{t\in\mathbb{N}}$ an i.i.d. process independent of ϕ_0, with $U_1 \sim \nu_U^{d\lambda}$. Then, there exist a matrix $\mathbf{B} = \mathbf{H}^{-1/2} \in \mathcal{S}_{++}^d$ and a vector $b \in \mathbb{R}^d$, such that the function $g = f(\mathbf{B}\cdot + b)$ is sherical, with minimum in 0. Then, define

$$\hat{\phi}_0 = (\hat{m}_0, \hat{p}_0^c, \hat{\sigma}_0, \hat{\mathbf{C}}_0) = \Phi_{\mathbf{B},b}(\phi_0),$$

where $\Phi_{\mathbf{B},b}$ is defined by (5.5). Let $\{\hat{\phi}_t\}_{t\in\mathbb{N}} = \{(\hat{m}_t, \hat{p}_t^c, \hat{\sigma}_t, \hat{\mathbf{C}}_t)\}_{t\in\mathbb{N}}$ be the process produced by CMA-ES when minimizing the spherical function g with random inputs $\hat{\mathbf{U}} = \{\hat{U}_{t+1}\}_{t\in\mathbb{N}}$, with

$$\hat{U}_{t+1} = \Psi_{\mathbf{B},b}(U_{t+1}, \phi_t),$$

where $\Psi_{\mathbf{B},b}$ is defined by (5.6). Then, by Proposition 5.2 applied to the spherical function g, we have

$$\lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|\mathbf{B}\hat{m}_{t+1}\|}{\|\mathbf{B}\hat{m}_t\|}\right] = \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2}\log \frac{\lambda_d(\mathbf{C}_{t+1})}{\lambda_d(\mathbf{C}_t)}\right] = -\mathrm{CR}.$$

Therefore, by Theorem 5.2

$$\begin{aligned}
\lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}\right] &= \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|\mathbf{B}\hat{m}_{t+1}\|}{\|\mathbf{B}\hat{m}_t\|}\right] =: -\mathrm{CR} \in \mathbb{R} \\
&= \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\hat{\sigma}_{t+1}}{\hat{\sigma}_t} + \frac{1}{2}\log \frac{\lambda_d(\hat{\mathbf{C}}_{t+1})}{\lambda_d(\hat{\mathbf{C}}_t)}\right] \\
&= \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2}\log \frac{\lambda_d(\mathbf{B}^{-1}\mathbf{C}_{t+1}\mathbf{B}^{-\top})}{\lambda_d(\mathbf{B}^{-1}\mathbf{C}_t\mathbf{B}^{-\top})}\right] \\
&= \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2}\log \frac{\lambda_d(\mathbf{H}^{1/2}\mathbf{C}_{t+1}\mathbf{H}^{1/2})}{\lambda_d(\mathbf{H}^{1/2}\mathbf{C}_t\mathbf{H}^{1/2})}\right] .
\end{aligned}$$

Thus, applying Proposition 5.2 to f, we find that the convergence rates of CMA-ES when minimizing f and g are equal. □

3.4 *Positivity of the convergence rate*

The goal of this section is to end the proof of Theorem 5.3, more specifically the positivity of the convergence rate of CMA-ES when the stepsize change function is Γ_{CSA}^2.

In that perspective, we provide in the next lemma a tight upper bound of the expected mean shift when the mean is the optimum of the objective function.

Lemma 5.6. Consider an ellipsoidal objective function f via **F**2 for some quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}^d_{++}$ and consider the Markov chain obeying (5.11). There exists $\bar{\varepsilon} > 0$, such that if $z_t = 0$, then

(i)
$$\mathbb{E}\left[\mu_{\text{eff}} \left\| \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \right\|^2\right] \leqslant (1 - 3\bar{\varepsilon}) \times d \ . \tag{5.16}$$

(ii)
$$\mathbb{E}\left[\|p_{t+1}\|^2\right] \leqslant (1 - 3\bar{\varepsilon})d + \frac{(1-c_\sigma)^2}{c_\sigma(2-c_\sigma)} \times \left(\mathbb{E}\|p_t\|^2 - \mathbb{E}\|p_{t+1}\|^2\right) \ . \tag{5.17}$$

Proof. We have

$$\left\| \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \right\|^2 = \sum_{j,k=1}^{d} \sum_{i_1,i_2=1}^{\mu} w_{i_1}^m w_{i_2}^m \left[U_{t+1}^{s_{t+1}(i_1)}\right]_j \left[U_{t+1}^{s_{t+1}(i_2)}\right]_k \ .$$

However, if $i_1 \neq i_2$ or if $j \neq k$, we have by symmetry with respect to 0 that

$$\mathbb{E}\left[\left[U_{t+1}^{s_{t+1}(i_1)}\right]_j \left[U_{t+1}^{s_{t+1}(i_2)}\right]_k\right] = 0 \ .$$

Indeed,

$$\left[U_{t+1}^{s_{t+1}(i_1)}\right]_j \left[U_{t+1}^{s_{t+1}(i_2)}\right]_k \overset{\text{dist.}}{=} - \left[U_{t+1}^{s_{t+1}(i_1)}\right]_j \left[U_{t+1}^{s_{t+1}(i_2)}\right]_k \ .$$

Hence

$$\mathbb{E}\left\| \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \right\|^2 = \mathbb{E}\left[\sum_{i=1}^{\mu} (w_i^m)^2 \left\| U_{t+1}^{s_{t+1}(i)} \right\|^2\right] \ .$$

We write $s_{t+1} = s_{\mathbf{\Sigma}_t; U_{t+1}}$, where $s_{\mathbf{\Sigma}; U_{t+1}}$ is the (almost surely unique) permutation of $\mathfrak{S}_\lambda$ that sorts the $f(\sqrt{\mathbf{\Sigma}} U^i_{t+1})$ for $i = 1, \ldots, \lambda$. Note that $s_{\mathbf{\Sigma}; U_{t+1}} = s_{\mathbf{\Sigma}/\lambda_{\max}(\mathbf{\Sigma}_t); U_{t+1}}$, and by continuity property, see Proposition 4.7, we have that

$$\max\left\{ \mathbb{E}\left[\sum_{i=1}^{\mu} (w_i^m)^2 \left\| U_{t+1}^{s_{\mathbf{\Sigma}; U_{t+1}}(i)} \right\|^2\right] \mid \mathbf{\Sigma} \in \mathcal{S}^d_+, \lambda_{\max}(\mathbf{\Sigma}) = 1 \right\}$$

is well-defined and finite. Let us denote $\mathbf{\Sigma}^* \in \mathcal{S}^d_+$ with $\lambda_{\max}(\mathbf{\Sigma}^*) = 1$ a matrix that reaches the above maximum. Without loss of generality we assume $\mathbf{\Sigma}^*$ to be diagonal.
Then

$$\mathbb{E}\left[\sum_{i=1}^{\mu} (w_i^m)^2 \left\| U_{t+1}^{s_{t+1}(i)} \right\|^2\right] \leqslant \mathbb{E}\left[\sum_{i=1}^{\mu} (w_i^m)^2 \left\| U_{t+1}^{s_{\mathbf{\Sigma}^*; U_{t+1}}(i)} \right\|^2\right]$$

Besides, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{\mu} (w_i^m)^2 \left\| U_{t+1}^{s_{\mathbf{\Sigma}^*; U_{t+1}}(i)} \right\|^2\right] &= \lambda \mathbb{E}\left[\left(w^m_{s^{-1}_{\mathbf{\Sigma}^*; U_{t+1}}(1)}\right)^2 \left\| U_{t+1}^1 \right\|^2\right] \\ &= \lambda \sum_{k=1}^{d} \mathbb{E}\left[\left(w^m_{s^{-1}_{\mathbf{\Sigma}^*; U_{t+1}}(1)}\right)^2 \left[U_{t+1}^1\right]_k^2\right] .\end{aligned}$$

However, for $k = 1, \ldots, d$, note that, as a consequence of **W**1, the map

$$\left[U_{t+1}^1\right]_k^2 \mapsto \left(w^m_{s^{-1}_{\boldsymbol{\Sigma}^*;U_{t+1}}(1)}\right)^2$$

is nonincreasing and not constant everywhere. By the Harris-FKG inequality, see Theorem 4.6, we have then

$$\mathbb{E}\left[\sum_{i=1}^{\mu}(w_i^m)^2\left\|U_{t+1}^{s_{\boldsymbol{\Sigma}^*;U_{t+1}}(i)}\right\|^2\right] < \lambda\mathbb{E}\left[\left(w^m_{s^{-1}_{\boldsymbol{\Sigma}^*;U_{t+1}}(1)}\right)^2\right]\mathbb{E}\left[\left\|U_{t+1}^1\right\|^2\right] = \mu_{\text{eff}}^{-1}d \ ,$$

ending the proof.
For (ii), we have

$$\mathbb{E}\|p_{t+1}\|^2 = (1-c_\sigma)^2\mathbb{E}\|p_t\|^2 + c_\sigma(2-c_\sigma)\mu_{\text{eff}}\mathbb{E}\left\|\sum_{i=1}^{\mu}w_i^m U_{t+1}^{s_{t+1}(i)}\right\|^2 + 2(1-c_\sigma)\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbb{E}\left\langle p_t, \sum_{i=1}^{\mu}w_i^m U_{t+1}^{s_{t+1}(i)}\right\rangle \ .$$

Since $z_t = 0$,

$$\left\langle p_t, \sum_{i=1}^{\mu}w_i^m U_{t+1}^{s_{t+1}(i)}\right\rangle \overset{\text{dist}}{=} -\left\langle p_t, \sum_{i=1}^{\mu}w_i^m U_{t+1}^{s_{t+1}(i)}\right\rangle \ .$$

Hence, by (i):

$$\mathbb{E}\|p_{t+1}\|^2 = (1-c_\sigma(2-c_\sigma))\mathbb{E}\|p_{t+1}\|^2 + c_\sigma(2-c_\sigma)(1-2\bar{\varepsilon})d + (1-c_\sigma)^2\times\left(\mathbb{E}\|p_t\|^2 - \mathbb{E}\|p_{t+1}\|^2\right) \ .$$

ending the proof. □

As a consequence of the previous lemma, we have a similar upper bound when the mean is close enough to the optimum.

Lemma 5.7. Suppose that the objective function f satisfies **F**2 and consider the Markov chain defined by (5.11). For $t \in \mathbb{N}$, denote $y_t = \lambda_1 \left(\boldsymbol{\Sigma}_t\right)^{-1}\boldsymbol{\Sigma}_t^{1/2}z_t$. There exists $\bar{\alpha} > 0$ and $\bar{\varepsilon} > 0$ such that

$$\mathbb{E}\left\|y_t\right\|^2 \leqslant \bar{\alpha} \Rightarrow \mathbb{E}\left[\mu_{\text{eff}}\left\|\sum_{i=1}^{\mu}w_i^m U_{t+1}^{s_{t+1}(i)}\right\|^2\right] \leqslant (1-\bar{\varepsilon})\times d. \tag{5.18}$$

Proof. By Lemma 4.5 and Lemma 5.6, we know that there exists $\bar{\varepsilon} > 0$ and $M_y > 0$ such that if $\|y_t\|^2 \leqslant M_y$, then

$$\mathbb{E}\left[\mu_{\text{eff}}\left\|\sum_{i=1}^{\mu}w_i^m U_{t+1}^{s_{t+1}(i)}\right\|^2\right] \leqslant (1-3\bar{\varepsilon})\times d + \varepsilon d \ ,$$

where $\varepsilon > 0$ can be chosen arbitrary small. Denote

$$M_U = \mathbb{E}\left[\max_{s\in\mathfrak{S}_\lambda}\mu_{\text{eff}}\left\|\sum_{i=1}^{\mu}w_i^m U_{t+1}^{s(i)}\right\|^2\right] < +\infty$$

and set

$$\bar{\alpha} := \frac{\bar{\varepsilon} d}{M_U} \times M_y > 0.$$

Assume that $\mathbb{E}\|y_t\|^2 \leqslant \bar{\alpha}$. Note that we have then

$$\bar{\alpha} \geqslant \mathbb{E}\|y_t\|^2 \geqslant \mathbb{P}\left[\|y_t\|^2 \geqslant M_y\right] \times M_y$$

hence by definition of $\bar{\alpha}$ we have

$$p := \mathbb{P}\left[\|y_t\|^2 \geqslant M_y\right] \leqslant \frac{\bar{\varepsilon} d}{M_U}.$$

Therefore, we have

$$\mathbb{E}\left[\mu_{\text{eff}} \left\| \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \right\|^2\right] \leqslant p \times M_U + (1-p) \times (1 - 2\bar{\varepsilon})d \leqslant \bar{\varepsilon} d + (1 - 2\bar{\varepsilon})d.$$

□

Under the conditions of Lemma 5.7, we can bound the expected squared norm of the path for the stepsize update.

Lemma 5.8. Suppose that the objective function f satisfies **F**2, that $c_\sigma \in (0, 1]$ and consider the Markov chain defined by (5.11). For $t \in \mathbb{N}$, denote $y_t = \lambda_1 \left(\mathbf{\Sigma}_t\right)^{-1} \mathbf{\Sigma}_t^{1/2} z_t$. There exists $\bar{\alpha} > 0$ and $\bar{\varepsilon} > 0$ such that If $\mathbb{E}\|y_t\|^2 < \bar{\alpha}$ and $1 \leqslant \mathbb{E}\|p_{t+1}\|^2 = \mathbb{E}\|p_t\|^2 < +\infty$, then:

$$\mathbb{E}\|p_{t+1}\|^2 \leqslant (1 - \bar{\varepsilon})d.$$

Proof. We adopt the same notations as in Lemma 5.7 and its proof. We have

$$\begin{aligned}\mathbb{E}\|p_{t+1}\|^2 \leqslant (1-c_\sigma)^2 \mathbb{E}\|p_t\|^2 + (1 - (1-c_\sigma)^2)(1-\varepsilon)d \\ + \mathbb{P}\left[\|y_t\|^2 \leqslant M_y \text{ and } s_{t+1} \neq s_{0,\mathbf{\Sigma}_t} \text{ or } \|y_t\|^2 > M_y\right] \times \sqrt{\mathbb{E}\|p_t\|^2 M_U}.\end{aligned}$$

Up to taking $\bar{\alpha} > 0$ smaller we can assume that

$$\mathbb{P}\left[\|y_t\|^2 \leqslant M_y \text{ and } s_{t+1} \neq s_{0,\mathbf{\Sigma}_t} \text{ or } \|y_t\|^2 > M_y\right] \leqslant M_U^{-1/2}(1 - (1-c_\sigma)^2) \times \epsilon,$$

for an arbitrary small $\epsilon > 0$. Thus, since $\mathbb{E}\|p_t\|^2 \geqslant 1$, $\sqrt{\mathbb{E}\|p_t\|^2} \leqslant \mathbb{E}\|p_t\|^2$, so we obtain

$$(1-\epsilon) \times (1 - (1-c_\sigma)^2)\mathbb{E}\|p_{t+1}\|^2 \leqslant (1 - (1-c_\sigma)^2)(1-\varepsilon)d,$$

ending the proof. □

We can now finish the proof of Theorem 5.3. It is achieved by considering two possible cases: first, if the states of the normalized Markov chain, when distributed according to the invariant measure, are such that the previous lemmas stated in this section apply, and then the expected log stepsize change is negative. This implies, under additional assumptions on the learning rates, that the convergence rate is positive. Otherwise, this gives a decrease condition on the expected progress of the mean to

the optimum, which can only be achieved with a positive convergence rate.

Proposition 5.4. Suppose that $\Gamma = \Gamma_{\text{CSA}}^2$, and that the objective function f and the weights $\mathbf{w}_m$, $\mathbf{w}_c$ satisfy **F**2 and **W**1, respectively. Suppose that μ_{eff}/d is sufficiently large, $c_1 + c_\mu < 1$, d_σ^{-1} is sufficiently larger than c_μ and c_m^2, and sufficiently smaller than c_m, and $(1-c_1-c_\mu)^{-1}(1-c_c)^2 \leqslant 1 - c_\mu$. Then, CMA-ES converges linearly, that is, the convergence rate CR in Proposition 5.2 is positive.

Proof. Remark that for $z \in \mathbb{R}^d$ and $\mathbf{\Sigma} \in \mathcal{S}_{++}^d$, we have $\|\lambda_1 (\mathbf{\Sigma})^{-1} \mathbf{\Sigma}^{1/2} z\| \leqslant \|z\|$. Then, by Theorem 4.3 and Lemma 5.1, $(z, p, q, \mathbf{\Sigma}, r) \mapsto \|\lambda_1 (\mathbf{\Sigma})^{-1} \mathbf{\Sigma}^{1/2} z\|^2$ and $(z, p, q, \mathbf{\Sigma}, r) \mapsto \|p\|^2$ both belong to $\mathsf{L}^1(\pi)$. Hence there exists $\alpha \in (0, +\infty)$ such that

$$\lim_{t\to\infty} \mathbb{E}\left[\|y_t\|^2\right] = \mathbb{E}_\pi[\lambda_1 (\mathbf{\Sigma})^{-1} \mathbf{\Sigma}^{1/2} z] = \alpha < +\infty$$

where, for $t \in \mathbb{N}$, we denote $y_t = \lambda_1 (\mathbf{\Sigma}_t)^{-1} \mathbf{\Sigma}_t^{1/2} z_t$.
Let $\bar{\alpha} > 0$ and $\bar{\varepsilon} > 0$ be constants which satisfy Lemma 5.8.
If $\alpha < \bar{\alpha}$*:* then by Lemma 5.8, $\mathbb{E}_\pi \|p\|^2 \leqslant (1 - \bar{\varepsilon})d$. Therefore

$$\mathbb{E}_\pi \left[\log \Gamma_{\text{CSA}}^2(p)\right] \leqslant -\frac{c_\sigma}{2d_\sigma}(1 - \bar{\varepsilon}) < 0.$$

However

$$\text{CR} = -\mathbb{E}_\pi \left[\log \Gamma_{\text{CSA}}^2(p)\right] - \frac{1}{2}\mathbb{E}_\pi \left[\log r_{t+1}\right].$$

Hence, up to choosing $c_1 + c_\mu$ smaller enough than c_σ/d_σ, we have that $\text{CR} > 0$.
If $\alpha \geqslant \bar{\alpha}$*:* Suppose that the random variable $(z_0, p_0, q_0, \mathbf{\Sigma}_0, r_0)$ is distributed with respect to π. Then

$$\forall t \in \mathbb{N}, \quad \mathbb{E}\|y_t\|^2 \geqslant \bar{\alpha}.$$

Then, for any $t \in \mathbb{N}$, there exists $j \in \{1, \ldots, d\}$, such that

$$\mathbb{E}\langle y_t \rangle_j \geqslant d^{-1}\mathbb{E}\|y_t\|^2 \geqslant d^{-1}\bar{\alpha}$$

where $(\langle\cdot\rangle_1, \ldots, \langle\cdot\rangle_d)$ denotes the orthonormal coordinate system in which $\mathbf{\Sigma}_t$ is diagonal with decreasingly ordered diagonal elements, and $\langle y_t \rangle_k$ is nonnegative for $k = 1, \ldots, d$.
Then, by Proposition 4.17, we have

$$\mathbb{E}\left[\left\langle \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \right\rangle_j \mid \mathcal{F}_t\right] \leqslant -\kappa_{\bar{\alpha}}.$$

Yet, we have

$$\|m_{t+1} - x^*\|^2 = \|m_t - x^*\|^2 + 2c_m\sigma_t \sum_{k=1}^{d} \sqrt{\lambda_k(\mathbf{C}_t)}\langle m_t - x^* \rangle_k \left\langle \sum_{i=1}^{\mu} U_{t+1}^{s_{t+1}(i)} \right\rangle_k + c_m^2\sigma_t^2 \left\| \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} U_{t+1}^{s_{t+1}(i)} \right\|^2.$$

Hence,

$$\mathbb{E}\left[\|m_{t+1} - x^*\|^2 \mid \mathcal{F}_t\right] \leqslant \|m_t - x^*\|^2 - 2c_m\kappa_{\bar{\alpha}}\sigma_t\sqrt{\lambda_j(\mathbf{C}_t)}\langle m_t - x^* \rangle_j + c_m^2\sigma_t^2\|\mathbf{C}_t\| \times \mu_{\text{eff}}^{-1} M_U$$

where

$$M_U = \mathbb{E}\left[\max_{s\in\mathfrak{S}_\lambda} \mu_{\text{eff}} \left\| \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s(i)} \right\|^2\right] < +\infty.$$

However, remark that

$$\sigma_t \sqrt{\lambda_j(\mathbf{C}_t)} \langle m_t - x^* \rangle_j = \sigma_t^2 \lambda_1(\mathbf{C}_t) \times \langle y_t \rangle_j.$$

Hence,

$$\mathbb{E}\left[\|m_{t+1} - x^*\|^2 \mid \mathcal{F}_t\right] \leqslant \|m_t - x^*\|^2 + c_m \sigma_t^2 \lambda_1(\mathbf{C}_t) \times \left(c_m \mu_{\text{eff}}^{-1} M_U - 2\kappa_{\bar{\alpha}} \langle y_t \rangle_j\right).$$

Furthermore, up to changing the initialization of CMA-ES, we can assume that the random variables $\langle y_t \rangle_j$ and $\sigma_t^2 \lambda_1(\mathbf{C}_t)$ are independent. Indeed, the objective function f is scaling-invariant, and scaling the Markov chain $\{(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$ by a factor to set a different value for $\sigma_t^2 \lambda_1(\mathbf{C}_t)$ does affect the normalized chain $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ and hence $\langle y_t \rangle_j$ remains unchanged. Besides, by Proposition 5.2, the value of CR is identical for every initialization of CMA-ES. This yields to

$$\mathbb{E}\left[\|m_{t+1} - x^*\|^2\right] \leqslant \mathbb{E}\left[\|m_t - x^*\|^2\right] + c_m \mathbb{E}\left[\sigma_t^2 \lambda_1(\mathbf{C}_t)\right] \times \left(c_m \mu_{\text{eff}}^{-1} M_U - 2\kappa_{\bar{\alpha}}(1-\varepsilon)\bar{\alpha}\right).$$

Then, up to choosing $c_m \leqslant \mu_{\text{eff}} M_U^{-1} \kappa_{\bar{\alpha}} \bar{\alpha}$, we have for $t \in \mathbb{N}$

$$\mathbb{E}\|m_{t+1} - x^*\|^2 \leqslant \mathbb{E}\|m_t - x^*\|^2 - c_m \kappa_{\bar{\alpha}} (1-2\varepsilon)\bar{\alpha} \mathbb{E}\left[\sigma_t^2 \lambda_1(\mathbf{C}_t)\right]$$

so

$$\mathbb{E}\|m_{t+1} - x^*\|^2 < \mathbb{E}\|m_t - x^*\|^2$$

and thus $\mathrm{CR} > 0$. □

4 Limit distribution of the covariance matrix

In this section, we describe the limit of the distribution of the covariance matrix in CMA-ES, when the objective function is ellipsoidal. Indeed, we prove that the limit (in distribution) of the covariance matrix normalized by its determinant is an affine transformation of a probability measure which satisfies an invariance to rotation property (5.19).

To prove this result, we rely on the affine-invariance of CMA-ES (Theorem 5.2), and thus to ensure this property, we assume that no cumulation is used to update the stepsize.

Theorem 5.4. Suppose that the stepsize change function Γ and the weights $\mathbf{w}_m$, $\mathbf{w}_c$ satisfy **Γ**1-**Γ**5 and **W**1, respectively. Suppose that $\mu \leqslant \lambda/2$ and μ_{eff} is sufficiently large, $c_1 + c_\mu < 1$, d_σ^{-1} is sufficiently larger than $c_\mu > 0$ and $c_m^{3/2} > 0$, and sufficiently smaller than c_m, and $2c_c$ is larger than c_μ. Besides assume that $c_\sigma = 1$.
There exists a probability distribution π_C on $\mathcal{B}(\det^{-1}(\{1\}))$ which satisfies for every $\mathbf{P} \in \mathsf{O}(d)$ and every $\mathsf{A} \in \mathcal{B}(\det^{-1}(\{1\}))$ the following invariance property:

$$\pi_C(\mathsf{A}) = \pi_C\left(\{\mathbf{P}\mathbf{\Sigma}\mathbf{P}^\top \mid \mathbf{\Sigma} \in \mathsf{A}\}\right), \tag{5.19}$$

such that, if the objective function f satisfies **F**2 for some quasi-Hessian matrix $\mathbf{H} \in \mathcal{S}_{++}^d$, then,

for every initialization $(m_0, \sigma_0, \mathbf{C}_0, p_0^\sigma, p_0^c) \in \mathbb{R}^d \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \times \mathbb{R}^d \times \mathbb{R}^d$ of CMA-ES, then, when t tends to $+\infty$, $\mathbf{C}_t / \det(\mathbf{C}_t)^{1/d}$ converges in distribution towards a probability measure $\nu_{\mathbf{H}}$ on $\mathcal{B}(\det^{-1}(\{1\}))$ defined by

$$\nu_{\mathbf{H}}(\mathsf{A}) = \pi_C \left(\left\{ \frac{\mathbf{H}^{1/2} \boldsymbol{\Sigma} \mathbf{H}^{1/2}}{\det(\mathbf{H})^{1/d}} \mid \boldsymbol{\Sigma} \in \mathsf{A} \right\} \right) \quad \text{for every } \mathsf{A} \in \mathcal{B}\left(\det^{-1}(\{1\})\right). \tag{5.20}$$

In particular, when $\mathbf{H} = \mathbf{I}_d$, $\mathbf{C}_t / \det(\mathbf{C}_t)^{1/d}$ converges to $\nu_{\mathbf{I}_d} = \pi_C$.

Before proving Theorem 5.4, we give in the next lemma a property of invariant measures of the normalized chain when minimizing a spherical function. For the sake of simplicity we provide this result only for the case $c_\sigma \neq 1$ and $c_c \neq 1$. However, we note that this would not change significantly the result nor the proof otherwise in dropping the variables p, q or r.

Lemma 5.9. Suppose that the objective function f, the normalization function R, the stepsize change function Γ satisfy **F**1, **R**2, **Γ**4, respectively. Assume that the minimum of f is $x^* = 0$. Denote $\mathsf{X} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ the state space of a Markov chain $\Theta = \{\theta_t\}_{t\in\mathbb{N}} = \{(z_t, p_t, q_t, \boldsymbol{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ obeying to (5.11), and suppose that Θ is geometrically ergodic, and denote π its unique invariant probability measure. Let $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ and $\mathbf{P} \in \mathsf{O}_d(\mathbb{R})$ be an orthogonal matrix of size $d \times d$. Define

$$\zeta_{\mathbf{P}}(\mathsf{A}) := \left\{ \theta = (z, p, q, \boldsymbol{\Sigma}, r) \in \mathsf{X} \mid \left(\mathbf{P}z, \mathbf{P}p, \mathbf{P}q, \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}^\top, r \right) \in \mathsf{A} \right\}. \tag{5.21}$$

Then, π is rotational invariant in the sense of $\pi(\mathsf{A}) = \pi(\zeta_{\mathbf{P}}(\mathsf{A}))$.

Proof. Let $\{\phi_t\}_{t\in\mathbb{N}} = \{(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$ be a sequence of iterates produced by CMA-ES when minimizing f, such that (4.16) holds. Since $f(\mathbf{P}\cdot) = f(\cdot)$ by **F**1 with $x^* = 0$, then by Theorem 5.1, if we define

$$(\hat{m}_t, \hat{p}_t^\sigma, \hat{p}_t^c, \hat{\sigma}_t, \hat{\mathbf{C}}_t) = \Phi_{\mathbf{P}^{-1},0}(\phi_t) \ ,$$

then $\{\hat{\phi}_t\}_{t\in\mathbb{N}} = \{(\hat{m}_t, \hat{p}_t^\sigma, \hat{p}_t^c, \hat{\sigma}_t, \hat{\mathbf{C}}_t)\}_{t\in\mathbb{N}}$ is a sequence of iterates produced by CMA-ES when minimizing f. Define then for $t \geqslant 1$:

$$\hat{z}_t = R(\hat{\mathbf{C}}_t)^{-1/2} \hat{\sigma}_t^{-1} \hat{m}_t \quad ; \quad \hat{p}_t = \hat{p}_t^\sigma \quad ; \quad \hat{q}_t = R(\hat{\mathbf{C}}_{t-1})^{-1/2} \hat{p}_t^c \quad ; \quad \hat{\boldsymbol{\Sigma}}_t = R(\hat{\mathbf{C}}_t)^{-1} \hat{\mathbf{C}}_t \quad ;$$
$$\hat{r}_t = R(\hat{\mathbf{C}}_{t-1})^{-1} R(\hat{\mathbf{C}}_t)$$

so that, by Theorem 4.2, $\hat{\Theta} = \{\hat{\theta}_t\}_{t\geqslant 1} = \{(\hat{z}_t, \hat{p}_t, \hat{q}_t, \hat{\boldsymbol{\Sigma}}_t, \hat{r}_t)\}_{t\geqslant 1}$ is a time-homogeneous Markov chain, obeying to (4.17), hence is geometrically ergodic, with π its unique invariant probability measure. However, remark that

$$\mathbb{P}[\theta_t \in \mathsf{A}] = \mathbb{P}[\hat{\theta}_t \in \zeta_{\mathbf{P}}(\mathsf{A})] \ .$$

When $t \to \infty$, by limit property of ergodic chains [110, Theorem 13.0.1], we get $\pi(\mathsf{A}) = \pi(\zeta_{\mathbf{P}}(\mathsf{A}))$. □

We are now able to prove Theorem 5.4.

Proof of Theorem 5.4. We assume here that $c_\sigma = 1$. For the sake of simplicity, the following proof applies to the case $c_c \neq 1$. Assuming otherwise would not change the arguments below more than dropping the variable p^c.
Let $\{\hat{\phi}_t\}_{t\in\mathbb{N}} = \{(\hat{m}_t, \hat{p}^c_t, \hat{\sigma}_t, \hat{\mathbf{C}}_t)\}_{t\in\mathbb{N}}$ be a sequence of iterates produced by CMA-ES when minimizing an ellipsoidal function f satisfying **F**2 for some $\mathbf{H} \in \mathcal{S}^d_{++}$ and $x^* \in \mathbb{R}^d$. Define then for $t \in \mathbb{N}$

$$(m_t, p^c_t, \sigma_t, \mathbf{C}_t) = \Phi_{\mathbf{H}^{-1/2}, -\mathbf{H}^{-1/2}x^*}(\hat{\phi}_t),$$

where $\Phi_{\mathbf{H}^{-1/2}, -\mathbf{H}^{-1/2}x^*}$ is defined by (5.5). Then, by Theorem 5.2, the process $\{\phi_t\}_{t\in\mathbb{N}} = \{(m_t, p^c_t, \sigma_t, \mathbf{C}_t)\}$ is a sequence of iterates produced by CMA-ES when minimizing a spherical function, i.e., satisfying **F**1, with a minimum in 0. Let $\Theta = \{\theta_t\}_{t\geqslant 1} = \{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ be the normalized process defined by (5.11). Then, by Theorem 4.3, Θ is a geometrically ergodic Markov chain, and by Lemma 5.9, its unique invariant probability measure π satisfies (5.21). Then,

$$\lim_{t\to\infty} \|\mathcal{L}(\theta_t) - \pi\|_{\mathsf{TV}} = 0.$$

Define for $\mathsf{A} \in \mathcal{B}(\det^{-1}(\{1\}))$:

$$\pi_C(\mathsf{A}) = \int \mathbb{1}\left\{\frac{\mathbf{\Sigma}}{\det(\mathbf{\Sigma})^{1/d}} \in \mathsf{A}\right\} \mathrm{d}\pi(z, q, \mathbf{\Sigma}, r).$$

Then, by Lemma 5.9, π_C satisfies (5.19) and

$$\lim_{t\to\infty} \left\| \mathcal{L}\left(\frac{\mathbf{\Sigma}}{\det(\mathbf{\Sigma})^{1/d}}\right) - \pi_C \right\|_{\mathsf{TV}} = 0.$$

Since, for $t \geqslant 1$, $\mathbf{C}_t/\det(\mathbf{C}_t)^{1/d} = \mathbf{\Sigma}_t/\det(\mathbf{\Sigma}_t)^{1/d}$, and $\hat{\mathbf{C}}_t = \mathbf{H}^{-1/2}\mathbf{C}_t\mathbf{H}^{-1/2}$, using

$$\frac{\hat{\mathbf{C}}_t}{\det(\hat{\mathbf{C}}_t)^{1/d}} = \frac{\mathbf{H}^{-1/2}\mathbf{C}_t\mathbf{H}^{-1/2}}{\det(\mathbf{H}^{-1/2}\mathbf{C}_t\mathbf{H}^{-1/2})^{1/d}} = \frac{\mathbf{H}^{-1/2}\mathbf{C}_t\mathbf{H}^{-1/2}}{\det(\mathbf{H}^{-1})^{1/d}\det(\mathbf{C}_t)^{1/d}} = \frac{\mathbf{H}^{-1/2}\mathbf{\Sigma}_t\mathbf{H}^{-1/2}}{\det(\mathbf{H}^{-1})^{1/d}\det(\mathbf{\Sigma}_t)^{1/d}}, \tag{5.22}$$

we get:

$$\begin{aligned}\lim_{t\to\infty} \mathbb{P}\left[\frac{\hat{\mathbf{C}}_t}{\det(\hat{\mathbf{C}}_t)^{1/d}} \in \mathsf{A}\right] &= \lim_{t\to\infty} \mathbb{P}\left[\frac{\mathbf{\Sigma}_t}{\det(\mathbf{\Sigma}_t)^{1/d}} \in \left\{\frac{\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}}{\det(\mathbf{H})^{1/d}} \mid \mathbf{\Sigma} \in \mathsf{A}\right\}\right] \\ &= \pi_C\left(\left\{\frac{\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}}{\det(\mathbf{H})^{1/d}} \mid \mathbf{\Sigma} \in \mathsf{A}\right\}\right)\end{aligned}$$

for every $\mathsf{A} \in \mathcal{B}(\det^{-1}(\{1\}))$. □

5 Learning of the inverse Hessian of convex-quadratic functions

As a consequence of Theorems 5.3 and 5.4, we give in this section our final contribution. Theorem 5.5 below shows that, when minimizing convex-quadratic functions, the covariance matrix in CMA-ES approximates the inverse Hessian of the objective function.

The proof of this result relies once more on the affine-invariance of CMA-ES (Theorem 5.2). Moreover, we use Theorem 5.4 that characterizes the limit distribution of the covariance matrix:

when the objective function is spherical, it is invariant to rotation and we deduce that the expected covariance matrix (normalized by its determinants) commutes with every rotation matrix, and thus is proportionnal to $\mathbf{I}_d$. Since an ellipsoidal function is an affine transformation of a spherical function, the affine-invariance of CMA-ES allows us to conclude.

Theorem 5.5. Consider an ellipsoidal objective function f satisfying **F**2 with quasi-Hessian matrix $\mathbf{H}$, as well as a stepsize change function Γ and weights $\mathbf{w}_m$, $\mathbf{w}_c$ satisfying **Γ**1-**Γ**5 and **W**1, respectively. Suppose that $\mu \leqslant \lambda/2$ and μ_{eff} is sufficiently large, $c_1 + c_\mu < 1$, d_σ^{-1} is sufficiently larger than $c_\mu > 0$ and $c_m^{3/2} > 0$, and sufficiently smaller than c_m, and $2c_c$ is larger than c_μ. Besides assume that $c_\sigma = 1$.
Then, there exists $\rho > 0$, such that for every initialization $(m_0, \sigma_0, \mathbf{C}_0, p_0^\sigma, p_0^c) \in \mathbb{R}^d \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \times \mathbb{R}^d \times \mathbb{R}^d$ of CMA-ES, we have

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbf{C}_t}{\det(\mathbf{C}_t)^{1/d}} = \rho \det(\mathbf{H})^{1/d} \mathbf{H}^{-1}$$

and

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\mathbf{C}_t}{\det(\mathbf{C}_t)^{1/d}}\right] = \rho \det(\mathbf{H})^{1/d} \mathbf{H}^{-1}.$$

Proof. Let π be the unique invariant probability measure of the positive Markov chain $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$, or $\{(z_t, \mathbf{\Sigma}_t)\}_{t\geqslant 1}$ if $c_c = 1$.
First assume that $\mathbf{H} = \mathbf{I}_d$ and $x^* = 0$. Since $\det(\cdot)^{1/d}$ is larger that λ_d on $\mathcal{S}_{++}^d$, by Lemma 5.2, we know that $(z, q, \mathbf{\Sigma}, r) \mapsto \det^{-1/d}(\mathbf{\Sigma})\mathbf{\Sigma}$, or if $c_c = 1$, $(z, \mathbf{\Sigma}) \mapsto \det^{-1/d}(\mathbf{\Sigma})\mathbf{\Sigma}$, is π-integrable. Moreover, by Theorem 5.4, we have, for every $\mathbf{P} \in \mathsf{O}_d(\mathbb{R})$,

$$\mathbb{E}_\pi\left[\frac{\mathbf{\Sigma}}{\det(\mathbf{\Sigma})^{1/d}}\right] = \mathbb{E}_\pi\left[\frac{\mathbf{P}\mathbf{\Sigma}\mathbf{P}^\top}{\det(\mathbf{P}\mathbf{\Sigma}\mathbf{P}^\top)^{1/d}}\right].$$

Since the determinant is stable by multiplication, we get then

$$\mathbb{E}_\pi\left[\frac{\mathbf{\Sigma}}{\det(\mathbf{\Sigma})^{1/d}}\right] \times \mathbf{P} = \mathbf{P} \times \mathbb{E}_\pi\left[\frac{\mathbf{\Sigma}}{\det(\mathbf{\Sigma})^{1/d}}\right].$$

Moreover, the only matrices that commute with every orthogonal matrices are homotheties (multiples by a scalar of the identity matrix), hence there exists $\rho \in \mathbb{R}$ such that

$$\mathbb{E}_\pi\left[\frac{\mathbf{\Sigma}}{\det(\mathbf{\Sigma})^{1/d}}\right] = \rho \mathbf{I}_d.$$

Thus, by the Law of Large Numbers [110, Theorem 17.0.1], we get

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\mathbf{\Sigma}_t}{\det(\mathbf{\Sigma}_t)^{1/d}}\right] = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbf{\Sigma}_t}{\det(\mathbf{\Sigma}_t)^{1/d}} = \rho \mathbf{I}_d.$$

Since $\det^{-1/d}(\mathbf{\Sigma}_t)\mathbf{\Sigma}_t \succ 0$ for every $t \geqslant 0$, we get $\rho > 0$. This proves Theorem 5.5 in the case $\mathbf{H} = \mathbf{I}_d$ and $x^* = 0$, since

$$\frac{\mathbf{\Sigma}_t}{\det(\mathbf{\Sigma}_t)^{1/d}} = \frac{\mathbf{C}_t}{\det(\mathbf{C}_t)^{1/d}} \quad \text{for } t \geqslant 0.$$

When $\mathbf{H} \neq \mathbf{I}_d$ or $x^* \neq 0$, consider iterates $\{\phi_t\}_{t\in\mathbb{N}} = \{(m_t, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$ produced by CMA-ES when minimizing f, such that (5.10) holds. Then, define for $t \in \mathbb{N}$,

$$(\hat{m}_t, \hat{p}_t^c, \hat{\sigma}_t, \hat{\mathbf{C}}_t) = \Phi_{\mathbf{H}^{-1/2}, -\mathbf{H}^{-1/2}x^*}(\phi_t),$$

where $\Phi_{\mathbf{H}^{-1/2}, -\mathbf{H}^{-1/2}x^*}$ is given by (5.5). Then, by Theorem 5.2, the process $\{\hat{\phi}_t\}_{t\in\mathbb{N}} = \{(\hat{m}_t, \hat{p}_t^c, \hat{\sigma}_t, \hat{\mathbf{C}}_t)\}_{t\in\mathbb{N}}$ is a sequence of iterates produced by CMA-ES when minimizing $f(\mathbf{H}^{-1/2}\cdot -\mathbf{H}^{-1/2}x^*)$ a spherical function with optimum in 0. Hence,

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\hat{\mathbf{C}}_t}{\det(\hat{\mathbf{C}}_t)^{1/d}}\right] = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{\mathbf{C}}_t}{\det(\hat{\mathbf{C}}_t)^{1/d}} = \rho\mathbf{I}_d.$$

This ends the proof, since $\hat{\mathbf{C}}_t = \mathbf{H}^{1/2}\mathbf{C}_t\mathbf{H}^{1/2}$, and since the determinant is stable by multiplication.

□

Conclusion and discussion

This dissertation consists in a mathematical proof that the algorithm CMA-ES converges at a geometric rate to the optimum of a specific class of functions. We summarize the key contributions of this work, as well as limitations of our results.

Our methodology to proving convergence of the algorithm is to define a stochastic process based on the normalization of CMA-ES. Similar to previous studies on related algorithms [17, 141], we demonstrate that this process forms a time-homogeneous Markov chain when the objective function is scaling-invariant—meaning the ranking of function values remains unchanged when the points are scaled by a multiplicative factor with respect to a reference point. This property is first established in Chapter 1 for a simplified version of CMA-ES, and then extended in Chapter 2 to the algorithm variants analyzed throughout the thesis. This approach seems however to be limited to the class of scaling-invariant objective functions, possibly with noise, which we did not investigate.

The principal idea of our proof is to apply a Law of Large Numbers to the normalized Markov chain, from which we derive a convergence rate. In the first part of the proof, we show that this chain is an irreducible and aperiodic T-chain. In the second part, we establish ergodicity through a geometric drift condition.

However, proving the irreducibility of such a complex process is a complicated problem. Initially, our intension was to use a methodology that provide sufficient conditions for nonlinear state-space models to be irreducible [29]. Yet, due to the normalization of the covariance matrix in CMA-ES, the Markov chain in our analysis is valued in a manifold, rather than in an open subset of Euclidean space, as required by this methodology. Chapter 1 addresses this problem by generalizing the conditions to apply to state spaces that are smooth manifolds. In addition, we incorporate nonsmooth (locally Lipschitz) update functions, allowing us to analyze ES with cumulative stepsize adaptation (CSA). Specifically, we prove that such a Markov chain is an irreducible and aperiodic T-chain when the deterministic control model associated with its update has a steadily attracting state satisfying a controllability condition. In future works, we aim to generalize our assumptions regarding the probability distribution of the random input. Currently, we require it to be absolutely continuous with respect to a sigma-finite measure locally equivalent to the Euclidean Lebesgue measure. Extending this to include discrete components in the distribution would enable the analysis of elitist algorithms, such as the plus strategies in ES.

In Chapter 2, we apply consequently our conditions to normalized Markov chains underlying CMA-ES. To include several hyperparameter settings, we introduce there the notions of redundant Markov chains that we reuse in Chapter 4. Furthermore, in the perspective to analyze nonsmooth normalization functions, we explain how our conditions for irreducibility and aperiodicity may be applied to state spaces that are nonsmooth manifolds. Chapter 2 focus mainly on the proof of existence

of a steadily attracting state for the update of normalized processes underlying CMA-ES, at which a controllability condition holds. This is a nontrivial result and we make several assumptions. On top of being scaling-invariant, the objective function should have Lebesgue negligible level sets. Besides, we exclude several hyperparameter settings that we believe are relevant to analyze.

The next step in our proof is given in Chapter 4 where we show that normalized Markov chains of CMA-ES are geometrically ergodic. This was the most challenging part in our approach, and required to obtain tight upper bounds on several quantities in order to satisfy a geometric drift condition. As an illustration to this difficulty, Chapter 3 shows the upper bounds that we use on the projection of eigenvectors of the updated covariance matrix of CMA-ES to the eigenvectors of the initial covariance matrix. Additionally, in order to address the evolution path for the stepsize update, we adopted a multi-step drift approach. As in Chapter 2, since we desired results for different hyperparameter settings, we extended the theory about the geometric ergodicity of stochastic processes, in particular for the analysis of redundant Markov chains. Yet we exclude for this part certain configurations of hyperparameters that would still be relevant to study. In particular we rely on the rank-mu update for the covariance matrix to obtain our results (even so we we included the rank-one update of the algorithm). Furthermore, at most one of the two evolution paths can be used. Once again, we rely on several assumptions. The objective function is in this chapter ellipsoidal: it has ellipsoidal level sets. Equivalently, it is an increasing transformation of a convex-quadratic function.

Chapter 5 constitutes the final step of our proof and concludes to the linear convergence of CMA-ES. As mentioned above, we apply a Law of Large Numbers and we find a convergence rate. This still required additional work to ensure several integrability properties with respect to the invariant measure of the normalized process. Moreover, in order to prove that this convergence rate is positive (and thus that the algorithm does converge) we assume the stepsize update to be a slightly modified version of CSA. This variant seems however to empirically perform as well as the default ones as shown in Appendix A. Nonetheless, for algorithm variants without the evolution path to update the stepsize, we were able to prove that the convergence rate does not depend on the condition number of the Hessian of a convex quadratic function, and that the inverse Hessian is approximated by the covariance matrix. We rely for these results on the affine-invariance property that is satisfied by several CMA-ES variants.

We summarize and list the limitations of our work that we hope would be dealt with in the near future.

- The methodology that we generalize to find the irreducibility and aperiodicity of Markov chains is still restricted to smooth manifolds for the state space and to distributions of the random inputs that are absolutely continuous with respect to Lebesgue.

- The analysis of convergence of a normalized Markov chain requires to set carefully the hyperparameters of CMA-ES and to use only one evolution path. In particular, it relies on the rank-mu update of the covariance matrix. The active update of CMA-ES [91], i.e., with negative weights for the update of the covariance matrix, is not included in our analysis. This would bring further difficulties as we need to ensure that the updated covariance matrix remains positive definite.

- We obtain the convergence of CMA-ES only for a specific stepsize update only (which slightly differs from the one which is used in practice).

- Other stepsize updates that are not based on an evolution path, like the two point adaptation (TPA) [64], are not mentioned in our works.

- We only prove convergence of non-elitist variants of CMA-ES (usually called comma strategies). However we believe that the analysis of elitist algorithms (plus strategies) [83] should not be more difficult.
- The affine-invariance of the algorithm is not compatible with the cumulation on the stepsize update. Yet, we conjecture that the covariance matrix still learns second-order information in this case and that the convergence rate does not vary with the choice of the ellipsoidal objective function. To extend our results, we may study invariance properties of the limit probability measure of the normalized Markov chain.
- This approach does not allow to estimate the value of the convergence rate. Drift-based approaches have been used for simpler algorithms [3, 4] and they could be investigated in the case of CMA-ES in future works.

Appendix A

Evaluation of the impact of various modifications to CMA-ES that facilitate its theoretical analysis

Comments on Appendix A: This chapter is published as a workshop paper entitled "Evaluation of the impact of various modifications to CMA-ES that facilitate its theoretical analysis" (Armand Gissler) in the proceedings of the conference *GECCO' 23* in 2023 [49]. The experiments presented here have been motivated while an early version of the proofs exposed in the manuscript analyze the algorithm CMA-ES without cumulation, especially when proving the geometric ergodicity of a normalized chain in Chapter 4. We also include in this work an affine-invariant update of CMA-ES. The last modification studied here is a smooth alternative to the CSA update of the step-size. The results of Chapter 1 were not stated yet, and the methodology to prove the irreducibility of a normalized chain in Chapter 2 was then based on previous results [29], which require a continuously differentiable update. Yet, we show these empirical results in this appendix to give some insights on the impact on the performances of CMA-ES by these modifications that might have been useful in our analysis.

We point out that some notations have been slightly changed from the original paper [49] to better fit the rest of the manuscript.

Abstract

In this paper we introduce modified versions of CMA-ES with the objective to help to prove convergence of CMA-ES. In order to ensure that the modifications do not alter the performances of the algorithm too much, we benchmark variants of the algorithm derived from them on problems of the `bbob` test suite. We observe that the main performances losses are observed on ill-conditioned problems, which is probably due to the absence of cumulation in the adaptation of the covariance matrix. However, the versions of CMA-ES presented in this paper have globally similar performances to the original CMA-ES.

1 Introduction . . . 218
2 Algorithm presentation . . . 218
3 Implementation and experimental procedure . . . 220
4 CPU Timings . . . 220
5 Results . . . 220
6 Conclusion / Discussion . . . 221

1 Introduction

In black-box optimization, the derivative-free algorithm CMA-ES (Covariance matrix adapation - Evolution strategies) [76] has shown good performances [73, 63] on many optimization problems. Although its theoretical foundations have progressed in the recent years, establishing a proof of convergence, even for very simple objective functions, remains very challenging.

In order to reduce the complexity of a theoretical analysis via the stability of a normalized Markov chain [110], we introduce in this paper small changes in the update equations. For instance, having continuously differentiable updates allows the use of verifiable conditions, in particular for the irreducibility of the forementioned normalized Markov chain [29]. So far, this approach was succesful to prove linear convergence for some algorithms in the ES class, e.g., $(1+1)$-ES [92] or $(\mu/\mu_w, \lambda)$-ES [141]. However, the adapation of a covariance matrix and cumulation increases the size of the state space of a normalized Markov chain, since it must include these parameters in addition to a normalized mean. Furthermore, a slightly different approach in the update of cumulative paths makes the algorithm that we know is affine-invariant and may extend the class of objective functions for which a proof of convergence exists.

In this paper, we examine modifications of the CMA-ES algorithm to address these different issues and benchmark variants of the CMA-ES algorithm with and without modifications. Our objective is to see whether and how much the performance of the original algorithm deteriorates.

All in all, we compare 10 variants of CMA-ES (including the original one) on the COCO platform [71], on 24 problems in the `bbob` suite.

This paper is organized as follows. In Section 2, we present the algorithm update equations as well as the considered modifications. Section 3 provides details on how the results were produced. We give the CPU timings of each variant in Section 4. In Section 5 we show the obtained results. We conclude in Section 6.

2 Algorithm presentation

The algorithm CMA-ES updates at each iteration $t \in \mathbb{N}$ the mean $m_t \in \mathbb{R}^d$, the stepsize $\sigma_t > 0$ and the covariance matrix $\mathbf{C}_t \in \mathcal{S}^d_{++}$ –that is, $\mathbf{C}_t$ belongs to the set of positive definite symmetric matrices of size $d \times d$– of a multivariate normal distribution $\mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$, as shown by Eqs. (A.2), (A.4) and (A.6) below. Besides, to better adapt the stepsize and the covariance matrix, CMA-ES relies on cumulation, that is it takes into account the favored directions in the previous iterations to update the stepsize and the covariance matrix. This is translated mathematically by the paths $p_t^\sigma \in \mathbb{R}^d$ and $p_t^c \in \mathbb{R}^d$, which are updated according to Eqs. (A.3) and (A.5).

Hence, at each iteration $t \in \mathbb{N}$, given current mean $m_t \in \mathbb{R}^d$, stepsize $\sigma_t > 0$, covariance matrix $\mathbf{C}_t \in \mathcal{S}^d_{++}$, and cumulation paths p_t^c, p_t^σ, we generate a population of $\lambda \in \mathbb{N}$ i.i.d. candidate solutions $x_{t+1}^1, \ldots, x_{t+1}^\lambda \sim \mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$, and we rank them w.r.t. their f-values, i.e. we define indices $i : \lambda$ (for $i = 1, \ldots, \lambda$) such that

$$f\left(x_{t+1}^{1:\lambda}\right) \leqslant f\left(x_{t+1}^{2:\lambda}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{\lambda:\lambda}\right). \tag{A.1}$$

We update then the algorithm parameters according to the following equations

$$m_{t+1} = \sum_{i=1}^{\mu} w_i x_{t+1}^{i:\lambda} \tag{A.2}$$

$$p_{t+1}^\sigma = (1-c_\sigma) p_t^\sigma + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\, \sigma_t^{-1} \mathbf{C}_t^{-1/2} \left[m_{t+1} - m_t\right] \tag{A.3}$$

$$\sigma_{t+1} = \sigma_t \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\left\|p^\sigma_{t+1}\right\|}{\mathbb{E}\|\mathcal{N}(0, I_d)\|} - 1\right)\right) \tag{A.4}$$

$$p^c_{t+1} = (1 - c_c)p^c_t + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}}\sigma_t^{-1}\left[m_{t+1} - m_t\right] \tag{A.5}$$

$$\mathbf{C}_{t+1} = (1 - c_1 - c_\mu \sum w_i)\mathbf{C}_t + c_1[p^c_{t+1}][p^c_{t+1}]^\top + c_\mu\sigma_t^{-2}\sum_{i=1}^{\lambda} w_i\left[x^{i:\lambda}_{t+1} - m_t\right]\left[x^{i:\lambda}_{t+1} - m_t\right]^\top \tag{A.6}$$

where $c_1, c_\mu \geqslant 0$ are such that $c_1 + c_\mu \leqslant 1$, $c_c, c_\sigma > 0$ and the weights $w_1 \geqslant w_2 \geqslant \ldots w_\mu \geqslant 0 \geqslant w_{\mu+1} \geqslant \ldots \geqslant w_\lambda$ are such that $\sum_{i=1}^{\mu} w_i = 1$ and $\sum_{i=1}^{\lambda} w_i \approx 0$. Moreover, the integer $\lambda \geqslant 1$ is called the population size, and $\mu \in \{1, \ldots, \lambda\}$ is the parent number.

First, a possible simplification of the algorithm is to remove cumulation on the covariance matrix and/or the stepsize, i.e. set $c_c = 1$ and/or $c_\sigma = 1$. Second, we can assume that the updates are continuously differentiable w.r.t. the current parameters. Only the update of the stepsize in Eq. (A.4) is concerned, which we replace then by

$$\sigma_{t+1} = \sigma_t \times \exp\left(\frac{c_\sigma}{2d_\sigma}\left(\frac{\left\|p^\sigma_{t+1}\right\|^2}{d} - 1\right)\right). \tag{A.7}$$

Note that this modification requires to change the value of d_σ. In the default algorithm d_σ is chosen proportional to $\sqrt{\mu_{\text{eff}}}$ when μ is large since $\|p^\sigma_t\|$ scales with $\sqrt{\mu_{\text{eff}}}$ on linear selection. Therefore, with this modification, we have to chose d_σ proportional to μ_{eff} when μ is large, while applying Eq. (A.7) instead of Eq. (A.4).

Last, we introduce a modification the replacement of the path p^σ_t by p^c_t, i.e., instead of Eq. (A.4) we have

$$\sigma_{t+1} = \sigma_t \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\left\|\mathbf{C}_t^{-1/2}p^c_{t+1}\right\|}{\mathbb{E}\|\mathcal{N}(0, I_d)\|} - 1\right)\right). \tag{A.8}$$

This modification is indeed helpful in a theoretical context, as it reduces the dimension of the state space of the chain produced by the algorithm, since there would remain only one cumulation path (instead of two in the original algorithm). Besides, we know that this version of the algorithm is *affine-invariant* –we refer to [15] for a formal definition– and allow to extend in some contexts an analysis on e.g. spherical objective function to any (monotonic transformation of a) convex-quadratic objective function. Note that this introduces a bias on the $\log$ stepsize on uniformly random selection[1], whose expectation is 0 with the standard stepsize update.

Therefore, we analyze in total 10 variants of CMA-ES, as described here. We use the following labels to refer to them.

(a) The affine-invariant variant, i.e. when applying Eq. (A.8).

(s) The variant without cumulation on the stepsize, i.e. $c_\sigma = 1$.

(c) The variant without cumulation on the covariance matrix, i.e. $c_c = 1$.

(d) The continuously differentiable variant, i.e. when applying Eq. (A.7).

Moreover, we consider variants which are combinations of the previous ones, e.g., (ad) refers to the variant of CMA-ES applying Eqs. (A.7) and (A.8). Note that if we do not have cumulation, it is already affine-invariant, so we don't have to analyze variants that combine (a) with (s) and/or (c). Finally, the original CMA-ES is labeled (default).

[1]That is, choosing the indices $(i:\lambda)_{i=1,\ldots,\lambda}$ uniformly on the set of permutations of $\{1, \ldots, \lambda\}$

3 Implementation and experimental procedure

The code used for the benchmarking of the variants presented in this paper is available online [48]. This is based on the Python module of `cma`, and used options available in the `development − branch` which can also be found online [66]. To run the different variants we use respectively the following options in Python.

(a) `CSA_invariant_path` = `True`

(s) `es.adaptsigma.damps` += 1 − `es.adaptsigma.cs`
`es.adaptsigma.cs` = 1

(c) `es.sp.cc` = 1

(d) `CSA_squared` = `True`
`CSA_damp_mueff_exponent` = 1

The option `CSA_damp_mueff_exponent` = 1 for the variant (d) is due to a wrong scaling of d_σ with $\mu_{\text{eff}} := \sum w_i^2$ when the stepsize update follows Eq. (A.7). We refer to e.g. Figure A.1 where we see that the variant (d) with d_σ chosen as by default (labeled as (dm)) has lower performances, in particular on the Rastrigin function in higher dimensions. Note that the scaling of d_σ with μ_{eff} when using this update is not changed by default in v.3.3.0. of the `cma` module in Python [66]. The budget used for all experiments here was $2 \times 10^5 \times d$ function evaluations (where d is the dimension of the problem), with possibly up to 9 restarts with doubling the population size. We have used 15 instances, and analyzed 24 objective functions from `bbob` suite in dimensions $2, 3, 5, 10, 20, 40$.

We performed the experiment using COCO [71] v.2.6.2, and post-processed the results (as shown in Section 5) with COCO v.2.6.3.

4 CPU Timings

In order to evaluate the CPU timing of the algorithm, we have run each variant presented above with restarts on the entire `bbob` test suite for the entire budget. The Python code was run on a Linux machine with 32 cores, Intel ®Xeon ®E7 to E3 v4 processor, with Python 2.7.5. The time per function evaluation for dimensions 2, 3, 5, 10, 20, 40 for each variant is presented in Table A.1.

5 Results

Specific results for each algorithm variant presented in this paper are available online [48]. They notably include empirical cumulative distributions of the number of evaluations for a certain number of targets, expected runtimes measured in number of required function evaluations, and how they scale with dimension, on 24 objective functions of the `bbob` suite. We focus here on the main differences in the performances introduced by the modifications presented in this paper w.r.t. the original CMA-ES.

Figure A.5 shows how the expected runtime (measured in number of function evaluations) to reach given targets scales with the dimension of the problem when optimizing `bbob` functions with the default CMA-ES and the affine-invariant versions of CMA-ES (a) and (ad). We observe that (a) and (ad) performs well on many of these problems, in particular the sphere, Rastrigin, Weierstrass, Gallagher 101 peaks and Katsuuras. This is particularly significant for Katsuuras, where the affine-invariant variants solve up to three times more targets than the default, see Figure A.3. Note that there

CPU timings per function evaluation (in 10^{-4}s)						
Dimensions	2	3	5	10	20	40
(default)-CMA-ES	2.8	2.4	2.0	2.3	2.6	3.0
a-CMA-ES	2.8	2.4	2.3	2.3	2.4	3.0
ad-CMA-ES	3.0	2.4	2.0	1.5	1.4	1.7
s-CMA-ES	2.4	2.2	1.9	1.5	1.4	1.7
c-CMA-ES	2.5	2.1	2.0	1.3	1.4	1.7
sc-CMA-ES	4.5	3.5	2.9	3.0	5.0	3.2
d-CMA-ES	3.1	2.6	2.1	1.6	1.5	1.8
sd-CMA-ES	3.0	2.4	2.1	1.5	1.4	1.8
cd-CMA-ES	2.9	2.3	1.8	1.4	1.4	1.7
scd-CMA-ES	2.6	2.4	1.9	1.5	1.3	1.8

Table A.1: Average CPU timings for each variant

is no remarkable performance loss for these variants, as on most objective functions, they perform about as fast as the default.

In Figure A.5 are shown the scaling of the expected runtime with dimension for the modified CMA-ES algorithms without cumulation—(c), (s) and (sc)—compared to the default CMA-ES. The main differences are in favor of algorithms with cumulation on the covariance matrix especially on highly ill-conditioned problems, see Figure A.4. This is particularly significant on the Bent cigar function, for which we show in Figure A.2 the empirical cumulative distribution (ECDF) of the number of evaluations to reach several targets in dimension 40. We see there that variants with cumulation on the covariance matrix perform about twice faster in dimension 20 and thrice faster in dimension 40.

Figures A.6 and A.7 compare the algorithm variant (d), whose stepsize update is continuously differentiable, with the default CMA-ES. We can see that (d) performs slightly worse than the default on some functions, especially on high dimension, see e.g. Schwefel or Rastrigin separable, for which this is significant for some targets in larger dimensions. As explained in Section 3, Figure A.1 shows a possible undesirable effect on the performance when we do not rescale the value of d_σ with μ_{eff} when using the variant (d).

All in all, when applying all modifications, the algorithm (scd) still performs well on most functions of the `bbob` suite, with the same observations that as for the variants (sc) and (d), with results presented in Figures A.6 and A.7 and Table A.2. Note that, on most functions it performs as fast as the default CMA-ES. The highest differences of performances are, again, observed on ill-conditioned functions, see Figure A.4, but it is never more than three times slower.

6 Conclusion / Discussion

We have analyzed the effects of modifications of CMA-ES that allow for easier theoretical analysis. From the results obtained in this paper, we observe that these modifications do not deteriorate too much the performances. Specifically, it appears that cumulation on the covariance matrix helps significantly on ill-conditioned problems, and that the standard CSA update for the stepsize performs slightly better than the continuously differentiable alternative we introduce in this paper. We do not observe any major negative effects of the modification on the paths introduced here to make CMA-ES

affine-invariant.

We also note that the value of d_σ, when the stepsize update follows Eq. (A.7), is not correctly chosen in the version of the `cma` Python package used when running the experiment (v.3.3.0), and it is currently required to change the scaling of d_σ with μ_{eff} when using this variant.

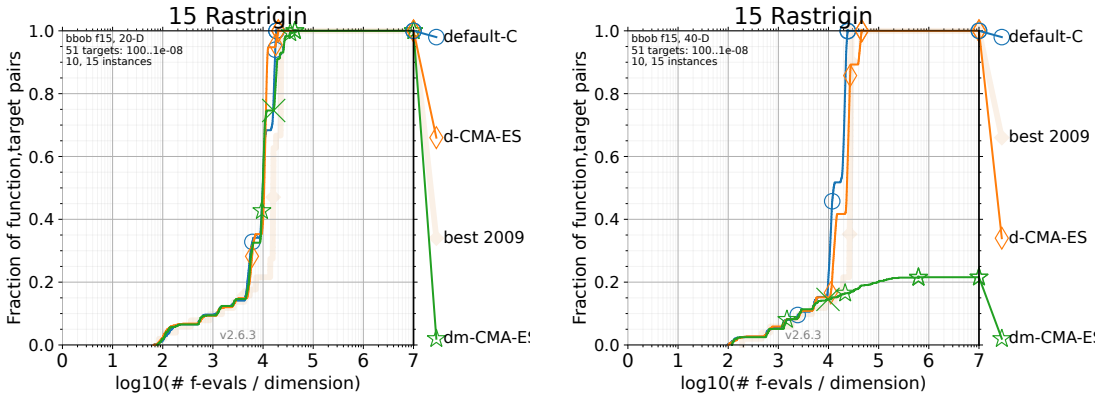


Figure A.1: Empirical cumulative distribution of the number of evaluations for 51 targets on the Rastrigin (left) and the Bent cigar (right) objective function in dimensions 20 (left) and 40 (right). In both cases the budget used is $2 \times 10^5 \times \text{dimension}$, and the variant (dm) does not use its full budget as it can diverge

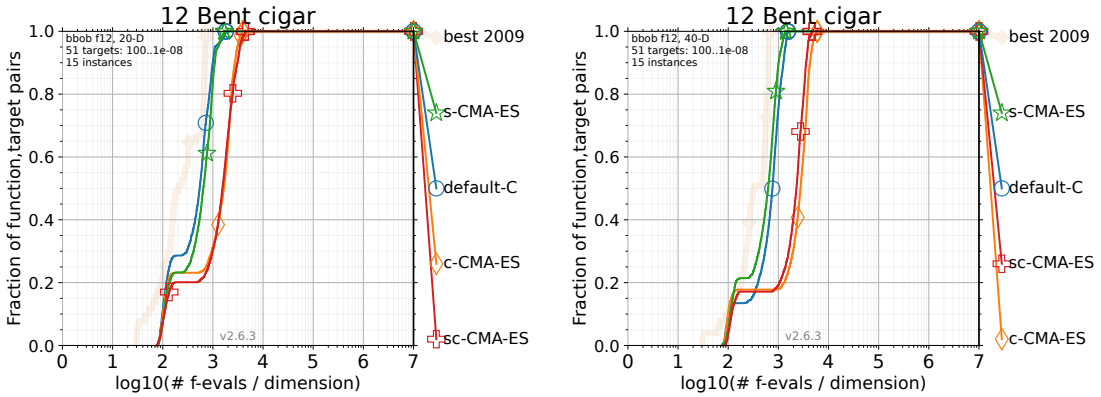


Figure A.2: Empirical cumulative distribution of the number of evaluations for 51 targets on the Bent cigar objective function in dimensions 20 (left) and 40 (right). In both cases the budget used is $2 \times 10^5 \times \text{dimension}$

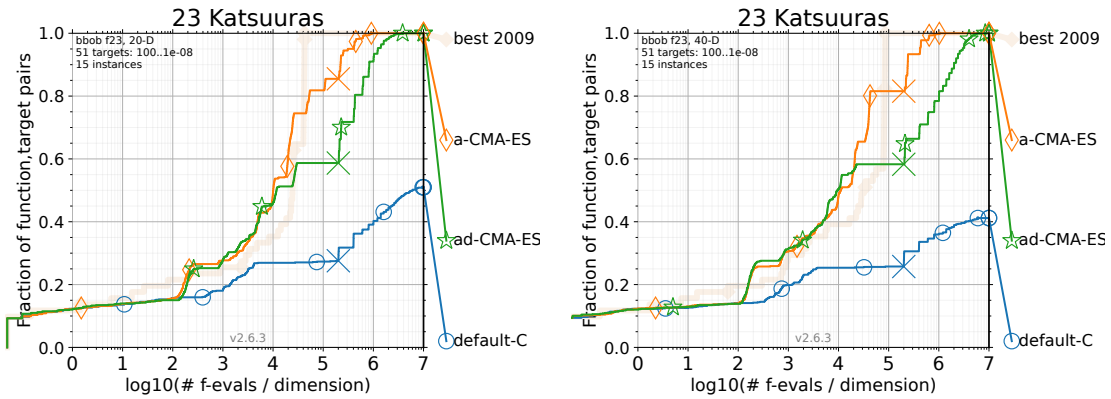


Figure A.3: Empirical cumulative distribution of the number of evaluations for 51 targets on the Katsuuras objective function in dimensions 20 (left) and 40 (right). In both cases the budget used is $2 \times 10^5 \times \text{dimension}$

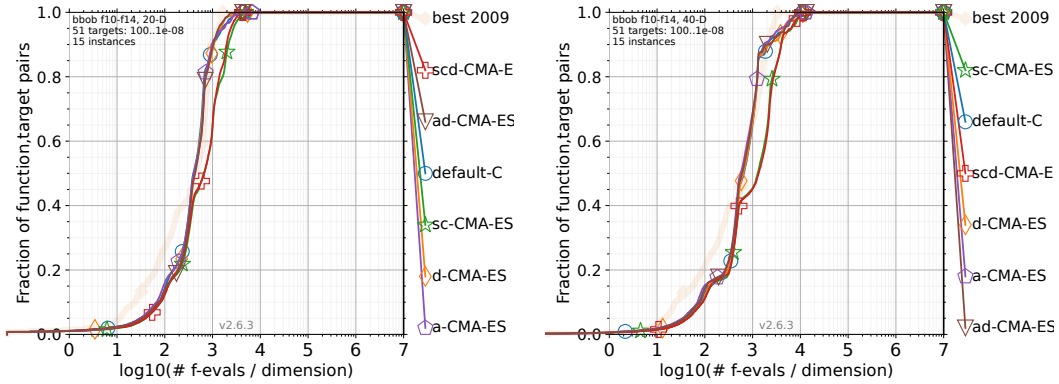


Figure A.4: Empirical cumulative distribution of the number of evaluations for 51 targets on the highly ill-conditioned objective functions group in dimensions 20 (left) and 40 (right). In both cases the budget used is $2 \times 10^5 \times$ dimension

Δf_{opt}	1e1	1e0	1e-1	1e-2	1e-3	1e-5	1e-7	#succ
f1	43	43	43	43	43	43	43	15/15
default	**6.8**(1.0)	**13**(1)	**19**(1)	**25**(2)	**32**(2)$^{\star 2}$	**45**(2)$^{\star 2}$	**57**(3)$^{\star 3}$	15/15
scd-CMA-ES	6.9(2)	15(5)	22(6)	30(5)	38(6)	53(7)	69(7)	15/15
f2	385	386	387	388	390	391	393	15/15
default	**21**(3)$^{\star 4}$	**25**(3)$^{\star 4}$	**28**(3)$^{\star 4}$	**30**(3)$^{\star 4}$	**31**(2)$^{\star 4}$	**32**(2)$^{\star 4}$	**34**(2)$^{\star 4}$	15/15
scd-CMA-ES	32(5)	39(6)	44(4)	47(4)	49(3)	51(4)	52(4)	15/15
f3	5066	7626	7635	7637	7643	7646	7651	15/15
default	**7.9**(8)	∞	∞	∞	∞	∞	∞ *3e6*	0/15
scd-CMA-ES	14(9)	∞	∞	∞	∞	∞	∞ *2e6*	0/15
f4	4722	7628	7666	7686	7700	7758	1.4e5	9/15
default	∞	∞	∞	∞	∞	∞	∞ *2e6*	0/15
scd-CMA-ES	∞	∞	∞	∞	∞	∞	∞ *2e6*	0/15
f5	41	41	41	41	41	41	41	15/15
default	**4.9**(1)	**5.6**(2)	**5.8**(1)	**5.8**(1)	**5.8**(1)	**5.8**(1)	**5.8**(1)	15/15
scd-CMA-ES	5.0(1)	6.3(2)	6.5(1)	6.5(1)	6.5(1)	6.5(1)	6.5(1)	15/15
f6	1296	2343	3413	4255	5220	6728	8409	15/15
default	1.4(0.2)	**1.2**(0.1)	**1.0**(0.1)	**1.0**(0.1)	**1.0**(0.1)	**1.0**(0.1)	**1.1**(0.1)	15/15
scd-CMA-ES	**1.4**(0.2)	1.2(0.1)	1.1(0.1)	1.1(0.1)	1.1(0.1)	1.1(0.1)	1.1(0.1)	15/15
f7	1351	4274	9503	16523	16524	16524	16969	15/15
default	1.0(1)	**2.5**(2)	**1.6**(0.6)*	**0.98**(0.3)*	**0.98**(0.3)*	**0.98**(0.3)*	**0.97**(0.3)*	15/15
scd-CMA-ES	**0.93**(0.3)	3.1(1)	2.2(0.6)	1.4(0.4)	1.4(0.4)	1.4(0.4)	1.4(0.4)	15/15
f8	2039	3871	4040	4148	4219	4371	4484	15/15
default	**3.3**(0.9)$^{\star 3}$	**4.3**(3)	**4.5**(3)	**4.5**(3)	**4.6**(3)	**4.6**(3)	**4.7**(3)	15/15
scd-CMA-ES	5.8(2)	7.4(4)	7.8(4)	8.1(4)	8.2(4)	8.2(4)	8.1(3)	15/15
f9	1716	3102	3277	3379	3455	3594	3727	15/15
default	**3.8**(0.9)$^{\star 3}$	**5.3**(4)	**5.6**(4)	**5.6**(3)	**5.7**(3)	**5.6**(3)	**5.6**(3)	15/15
scd-CMA-ES	6.0(2)	8.0(5)	8.6(5)	8.9(5)	9.0(5)	8.9(5)	8.8(4)	15/15
f10	7413	8661	10735	13641	14920	17073	17476	15/15
default	**1.1**(0.2)$^{\star 4}$	**1.1**(0.2)$^{\star 4}$	**1.00**(0.1)$^{\star 4}$	**0.83**(0.1)$^{\star 4}_{\uparrow 0}$	**0.79**(0.1)$^{\star 4}_{\uparrow 0}$	**0.73**(0.0)$^{\star 4}_{\uparrow 0}$	**0.75**(0.0)$^{\star 4}_{\uparrow 0}$	15/15
scd-CMA-ES	1.7(0.2)	1.8(0.2)	1.6(0.2)	1.4(0.1)	1.3(0.1)	1.2(0.1)	1.2(0.1)	15/15
f11	1002	2228	6278	8586	9762	12285	14831	15/15
default	**4.4**(0.4)	**2.2**(0.1)	**0.85**(0.1)$_{\uparrow 0}$	**0.67**(0.0)$_{\uparrow 0}$	**0.62**(0.0)$_{\uparrow 0}$	**0.55**(0.0)$_{\uparrow 0}$	**0.49**(0.0)$_{\uparrow 0}$	15/15
scd-CMA-ES	4.5(0.4)	2.3(0.2)	0.88(0.1)$_{\uparrow 0}$	0.69(0.0)$_{\uparrow 0}$	0.64(0.0)$_{\uparrow 0}$	0.56(0.0)$_{\uparrow 0}$	0.52(0.0)$_{\uparrow 0}$	15/15
f12	1042	1938	2740	3156	4140	12407	13827	15/15
default	**2.8**(3)	**2.2**(2)	**2.6**(2)	**2.9**(2)*	**2.8**(1)$^{\star 3}$	**1.2**(0.4)$^{\star 3}$	**1.3**(0.3)$^{\star 3}$	15/15
scd-CMA-ES	4.1(5)	5.7(7)	7.1(6)	7.7(6)	7.2(4)	3.1(1)	3.2(0.9)	15/15

Δf_{opt}	1e1	1e0	1e-1	1e-2	1e-3	1e-5	1e-7	#succ
f13	652	2021	2751	3507	18749	24455	30201	15/15
default	**2.3**(0.3)	**2.7**(2)	**4.0**(3)*	**4.0**(2)*	**1.0**(0.6)	**1.2**(0.6)	**1.4**(0.5)	15/15
scd-CMA-ES	4.4(7)	6.2(5)	7.6(4)	7.0(5)	1.5(0.6)	1.6(0.4)	1.6(0.5)	15/15
f14	75	239	304	451	932	1648	15661	15/15
default	**2.5**(1)	**2.3**(0.3)	**3.2**(0.5)	**3.6**(0.4)	**3.1**(0.4)	**3.7**(0.3)$^{\star 4}$	**0.67**(0.1)$_{\uparrow 0}^{\star 4}$	15/15
scd-CMA-ES	3.4(2)	2.7(0.6)	3.6(0.8)	4.0(0.7)	3.4(0.5)	4.8(0.4)	0.89(0.1)$_{\uparrow 0}$	15/15
f15	30378	1.5e5	3.1e5	3.2e5	3.2e5	4.5e5	4.6e5	15/15
default	**0.75**(0.6)*	**1.2**(0.4)	**0.71**(0.4)	**0.71**(0.4)	**0.72**(0.4)	**0.53**(0.3)$_{\uparrow 0}^{\star}$	**0.54**(0.3)$_{\uparrow 0}^{\star}$	15/15
scd-CMA-ES	1.3(0.5)	1.4(0.4)	0.77(0.4)	0.78(0.4)	0.79(0.4)	0.58(0.3)$_{\uparrow 0}$	0.59(0.3)$_{\uparrow 0}$	15/15
f16	1384	27265	77015	1.4e5	1.9e5	2.0e5	2.2e5	15/15
default	**1.9**(0.8)	**0.65**(0.6)	**0.65**(0.4)$_{\uparrow 0}$	0.81(0.8)	**1.2**(1)	**1.1**(1)	**1.1**(1)	15/15
scd-CMA-ES	2.2(0.9)	0.86(0.4)	0.94(0.6)	**0.78**(0.5)	1.4(1)	1.5(1)	1.4(1)	15/15
f17	63	1030	4005	12242	30677	56288	80472	15/15
default	**1.2**(1)	**0.96**(0.4)	**1.2**(2)	**0.98**(0.7)	**0.73**(0.3)	**0.88**(0.3)	**0.85**(0.2)	15/15
scd-CMA-ES	1.5(1)	1.0(0.2)	1.8(1)	1.1(0.6)	0.89(0.4)	0.93(0.4)	0.90(0.3)	15/15
f18	621	3972	19561	28555	67569	1.3e5	1.5e5	15/15
default	**0.86**(0.3)	**0.73**(0.1)$_{\uparrow 0}$	0.87(0.8)	1.3(0.8)	0.86(0.3)	0.82(0.4)	**0.79**(0.3)	15/15
scd-CMA-ES	0.97(0.3)	1.5(2)	**0.75**(0.4)	**1.3**(0.4)	**0.75**(0.3)	**0.74**(0.6)	0.89(0.6)	15/15
f19	1	1	3.4e5	4.7e6	6.2e6	6.7e6	6.7e6	15/15
default	**1**(0)	**1**(0)	**1.1**(0.9)	**0.48**(0.2)$_{\uparrow 0}^{\star 2}$	**0.54**(0.4)*	**0.59**(0.4)*	**0.59**(0.3)*	12/15
scd-CMA-ES	**1**(0)	**1**(0)	1.5(0.8)	1.2(0.8)	1.3(1)	1.8(2)	2.2(2)	3/15
f20	82	46150	3.1e6	5.5e6	5.5e6	5.6e6	5.6e6	14/15
default	**3.8**(0.9)	**4.0**(1)	**1.0**(1)$^{\star 3}$	**1.5**(2)	**1.5**(1)	**1.7**(2)	**1.7**(1)	5/15
scd-CMA-ES	5.0(1)	4.3(1)	∞	∞	∞	∞	∞ *3e6*	0/15
f21	561	6541	14103	14318	14643	15567	17589	15/15
default	**1.6**(4)	115(166)	**74**(74)	**73**(100)	**71**(80)	**67**(110)	**60**(108)	8/15
scd-CMA-ES	2.6(5)	**90**(100)	118(167)	116(137)	114(164)	107(126)	95(112)	6/15
f22	467	5580	23491	24163	24948	26847	1.3e5	12/15
default	171(24)	**141**(218)	∞	∞	∞	∞	∞ *1e6*	0/15
scd-CMA-ES	**8.4**(14)	141(216)	∞	∞	∞	∞	∞ *1e6*	0/15
f23	3.0	1614	67457	3.7e5	4.9e5	8.1e5	8.4e5	15/15
default	1.9(2)	1924(1872)	90(104)	**154**(175)	**115**(203)	∞	∞ *4e6*	0/15
scd-CMA-ES	**1.3**(1)	**391**(1240)	**53**(74)	∞	∞	∞	∞ *4e6*	0/15
f24	1.3e6	7.5e6	5.2e7	5.2e7	5.2e7	5.2e7	5.2e7	3/15
default	**4.5**(3)	7.6(12)	∞	∞	∞	∞	∞ *4e6*	0/15
scd-CMA-ES	7.1(6)	**2.4**(2)	∞	∞	∞	∞	∞ *4e6*	0/15

Table A.2: Expected runtime (ERT in number of f-evaluations) divided by the respective best ERT measured during BBOB-2009 in dimension 20. This ERT ratio and, in braces as dispersion measure, the half difference between 10 and 90run lengths appear for each algorithm and target, the corresponding reference ERT in the first row. The different target Δf-values are shown in the top row. #succ is the number of trials that reached the (final) target $f_{\mathrm{opt}} + 10^{-8}$. The median number of conducted function evaluations is additionally given in italics, if the target in the last column was never reached. Entries, succeeded by a star, are statistically significantly better (according to the rank-sum test) when compared to all other algorithms of the table, with $p = 0.05$ or $p = 10^{-k}$ when the number k following the star is larger than 1, with Bonferroni correction by the number of functions (24). A $\downarrow$ indicates the same tested against the best algorithm from BBOB 2009. Best results are printed in bold. Data produced with COCO v2.6.2

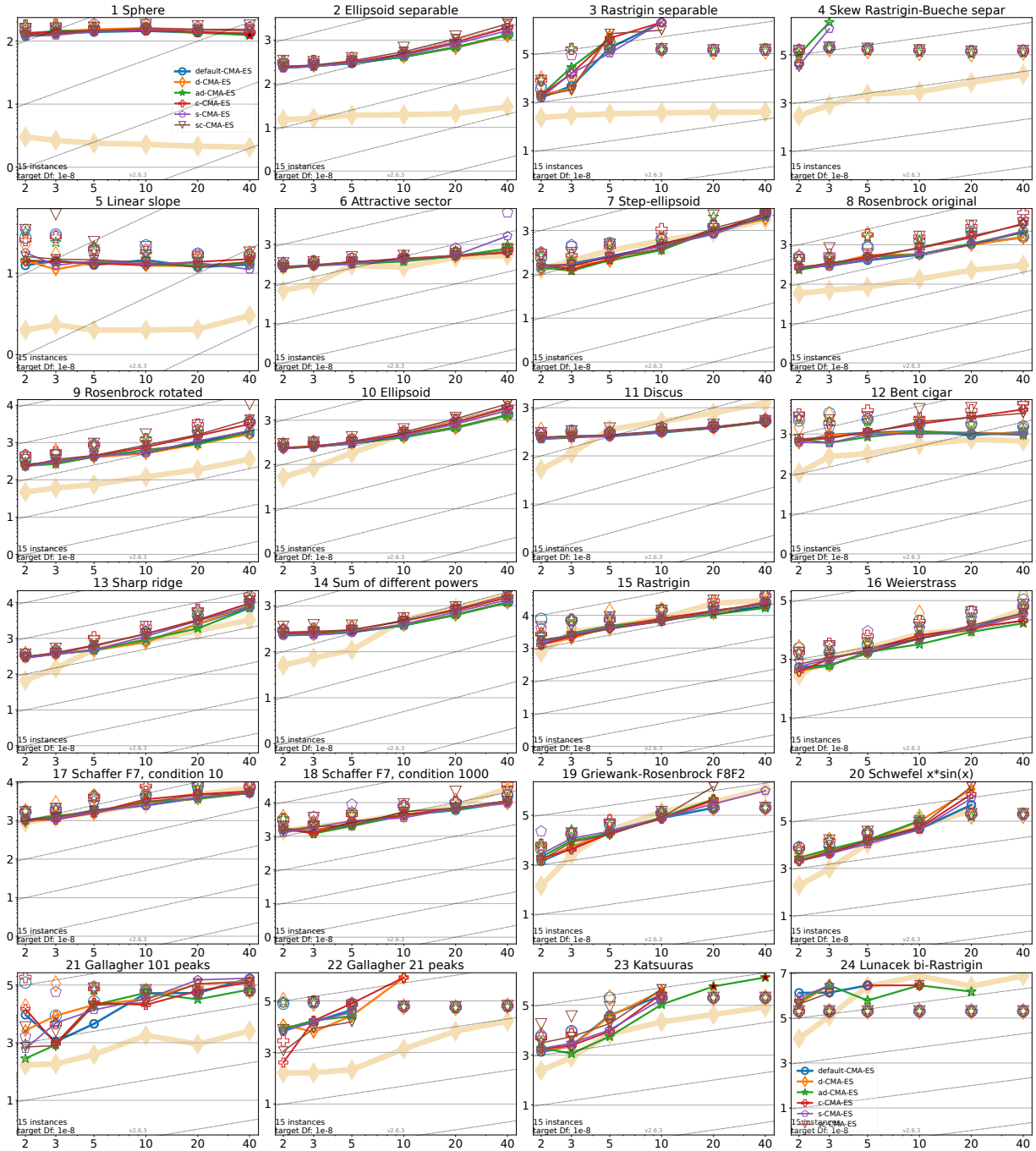


Figure A.5: Scaling of expected runtime (measured in function evaluations) with dimension: on the x-axis the dimension of the problem, on the y-axis $\log_{10}(\#\{\text{function evaluations}\}/\text{dimension})$

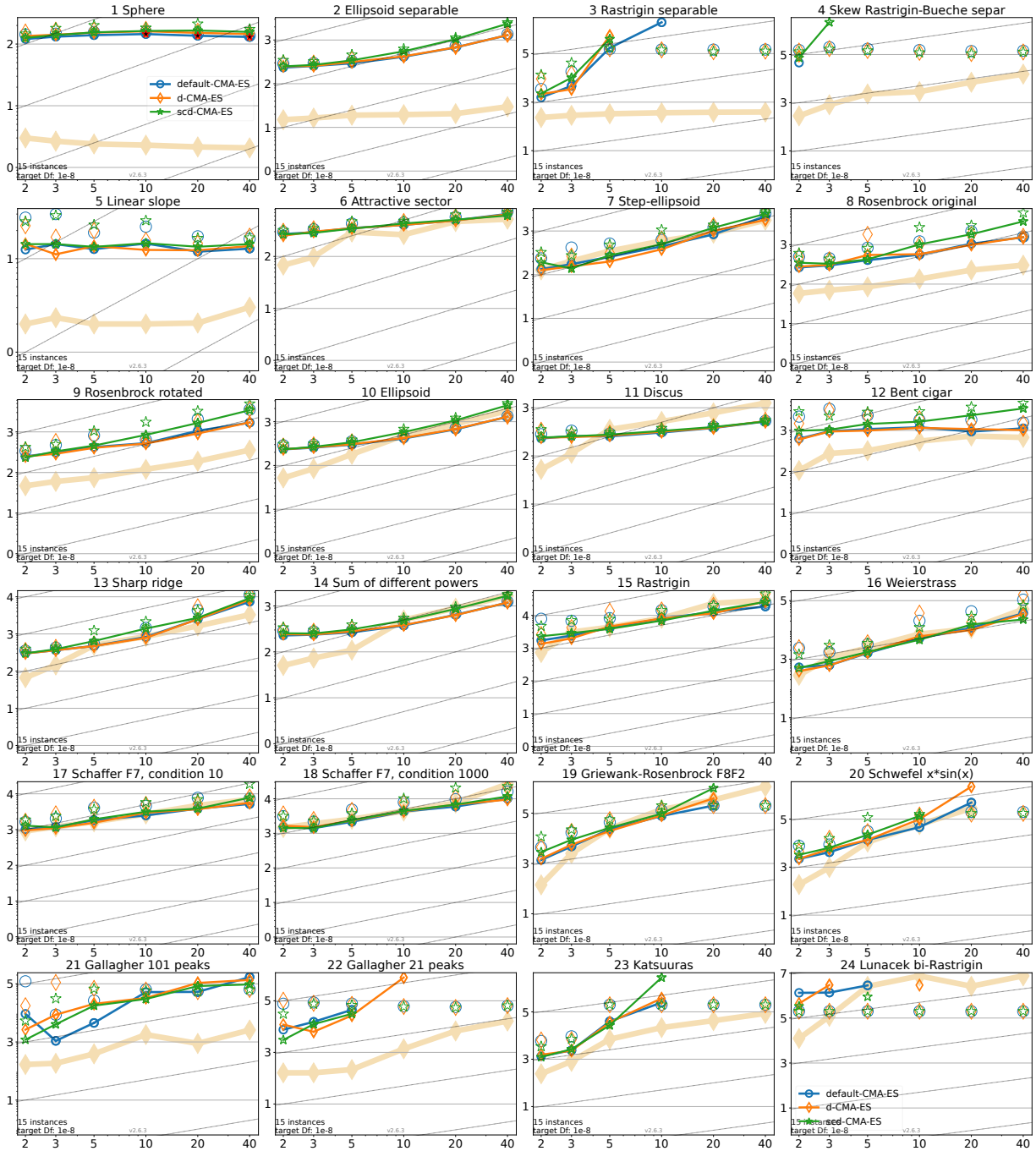


Figure A.6: Scaling of expected runtime (measured in function evaluations) with dimension: on the x-axis the dimension of the problem, on the y-axis $\log_{10}(\#\{\text{function evaluations}\}/\text{dimension})$

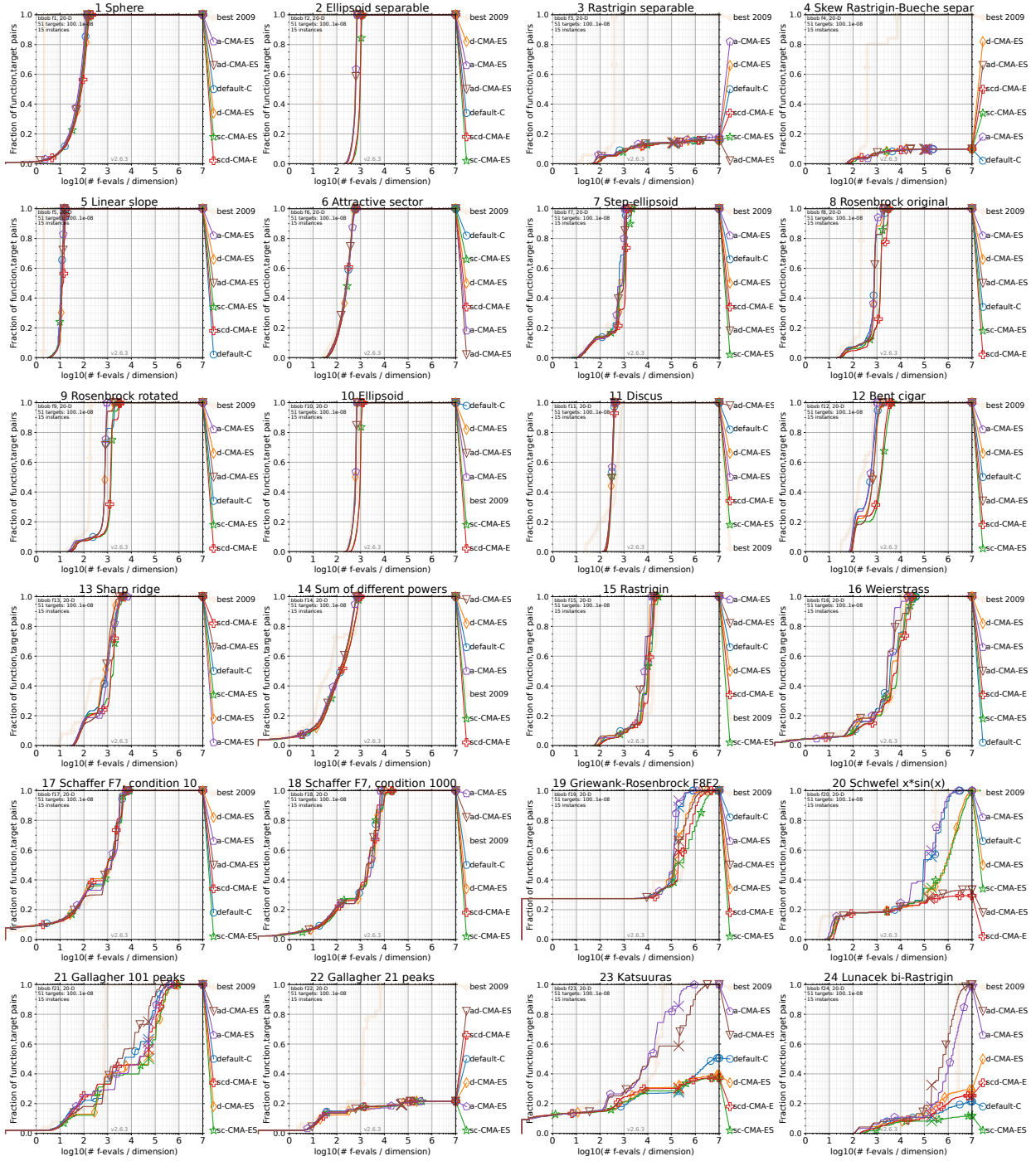


Figure A.7: Empirical cumulative distribution of simulated (bootstrapped) runtimes, measured in number of f-evaluations, divided by dimension for 51 targets $10^{[-8..2]}$ in 20D.

Appendix B

Introduction en français

1 Formulation mathématique de CMA-ES 230
1.1 Échantillonnage et classement des solutions candidates 231
1.2 Mise à jour de la moyenne 231
1.3 Mise à jour du pas 231
1.4 Mise à jour de la matrice de covariance 232
1.5 Résumé de CMA-ES 232
2 D'autres méthodes d'optimisation 233
2.1 Algorithmes sans dérivées 233
2.2 Méthodes de Newton 236
3 Convergence linéaire 237
4 Introduction aux chaînes de Markov 238
4.1 Noyaux et chaînes de Markov 239
4.2 Irréductibilité 239
4.3 Périodicité, apériodicité 240
4.4 Ensembles petits et small, T-chaînes 240
4.5 Récurrence, récurrence Harris 241
4.6 Mesures invaiantes 241
4.7 Ergodicité 241
4.8 Stabilité de chaînes de Markov non-linéaires par l'analyse d'un modèle de contrôle déterministe 242
5 Convergence et garanties théoriques pour les ES 244
5.1 Prouver un comportement linéaire des ES en analysant une chaîne de Markov normalisée 245
6 Méthodologie et aperçu 246
6.1 Chapitre 1: Sur l'irréductibilité et la convergence d'une classe de modèles d'espaces d'états non-lisses sur des variétés [52] 247
6.2 Chapitre 2: Irréductibilité de modèles à espaces d'états non-lisses avec application à CMA-ES[53] 247
6.3 Chapitre 3: Estimation asymptotique d'un problème de vecteurs propres symétrique perturbé [51] 248
6.4 Chapitre 4: Ergodicité géométrique de chaîne de Markov sous-jacentes à CMA-ES . . 249
6.5 Chapitre 5: Convergence linéaire de CMA-ES sur des problèmes ellipsoïdaux et apprentissage d'informations de second ordre 250

6.6 Annexe A: Évaluation de l'impact de modifications variées de CMA-ES qui facilitent son analyse théorique [49] *251*

Le but de l'optimisation est de trouver une solution au problème suivant:

$$\text{trouver } x^* \in \operatorname{Arg\,min} f \tag{P}$$

où $f\colon \Omega \to \mathbb{R}$ est une fonction définie sur un domaine Ω. Dans cette thèse, nous prendrons $\Omega = \mathbb{R}^d$, avec d un entier strictement positif. Dans ce contexte, trouver une solution exacte à (P) est souvent difficile, et cherche donc à approcher le minimum global x^* de f avec des algorithmes qui produisent des suites convergeant vers x^*.

Parmi les algorithmes d'optimisation sans dérivées[1], les stratégies d'évolution (ES) sont une classe d'algorithmes qui ont connu leur premiers développements dans les années 60 et 70 [131, 124]. Ces méthodes s'inspirent le principe de la sélection naturelle en biologie évolutive. La stratégie d'évolution avec adaptation de la matrice de covariance (CMA-ES) est l'algorithme à l'état de l'art de cette classe, très utilisé par exemple en biologie [129, 22] et en médecine [121], dans l'énergie [93, 125], pour l'apprentissage automatique [60, 80, 123] and en optimisation d'hyperparamètres [103, 2, 118], en imagerie numérique [31, 122, 140], ce qui inclut des applications pour les véhicules autonomes[104, 1], les pages web [95] ou les niveaux de jeu vidéo [144, 137]. Elle approche la solution x^* par une distribution normale multivariée, et adapte une matrice de covariance pour favoriser l'échantillonnage dans les directions qui augmentent les valeurs par f des solutions candidates précédentes.

Bien que des propositions initiales de CMA-ES prenaient la moyenne de la matrice de covariance avec une *mise à jour de rang un* [74, 75, 76], des améliorations successives introduisent une *mise à jour de rang mu* avec des poids égaux [73], puis avec des poids de recombinaison [72] et plus tard une *mise à jour active* reposant sur des poids négatives pour les pires solutions candidates [91]. Des preuves empiriques [76, 73] suggèrent que la moyenne produite par CMA-ES converge vers la solution de beaucoup de problèmes d'optimisation avec haute probabilité, y compris des problèmes non-convexes, non-séparable, mal conditionnés et multi-modaux. De plus, CMA-ES semble apprendre de l'information de second ordre, une propriété satisfaite par les méthodes quasi-Newton, ce qui est remarquable puisque CMA-ES repose uniquement sur des comparaisons d'évaluations par f.

Cependant, une démonstration mathématiques de ces observations manque. Alors que plusieurs ES ont vu une preuve de convergence vers l'optimum global pour des classes spécifiques de fonctions objectives [23, 39, 17, 4, 6, 54, 141], la convergence linéaire vers l'optimum global et l'approximation de l'inverse de la Hessienne par CMA-ES restent des problèmes ouverts. Cette thèse cherche à répondre à ces deux questions.

1 Formulation mathématique de CMA-ES

Le principe de CMA-ES est d'approcher le minimum x^* de la fonction objectif $f\colon \mathbb{R}^d \to \mathbb{R}$ par une distribution nirmale multivariée $\mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$ où $t \in \mathbb{N}$ est l'itération présente, le vecteur $m_t \in \mathbb{R}^d$ est la moyenne, le scalaire $\sigma_t > 0$ est le pas, et $\mathbf{C}_t \in \mathcal{S}_{++}^d$ la matrice de covariance à l'itération t. De plus, nous utilisons deux chemins $p_t^\sigma, p_t^c \in \mathbb{R}^d$ pour les mises à jours respectives du pas et de la matrice de covariance. À l'itération $t \in \mathbb{N}$, étant donné $(m_t, \sigma_t, \mathbf{C}_t, p_t^\sigma, p_t^c) \in \mathbb{R}^d \times \mathbb{R}_{++} \times \mathcal{S}_{++}^d \times \mathbb{R}^d \times \mathbb{R}^d$, la mise à jour des états de CMA-ES est donnée ci-dessous. Cette thèse ne s'intéresse pas à la mise à jour active [91] qui utilise des poids négatifs pour adapter la matrice de covariance.

[1] qui n'utilisent pas les dérivées ou même supposent leur existence

1.1 *Échantillonnage et classement des solutions candidates*

Premièrement, une population de λ enfants $x_{t+1}^1, \ldots, x_{t+1}^\lambda$ est générée depuis la distribution $\mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$. Dans cette section et dans tout le manuscrit, l'entier $\lambda \geqslant 2$ represent la taille de population de CMA-ES. Plus précisément, considérons les variables aléatoires

$$U_{t+1}^1, \ldots, U_{t+1}^\lambda \sim \mathcal{N}(0, \mathbf{I}_d) \text{ i.i.d.}, \tag{B.1}$$

qui définissent les solutions candidates suivantes

$$x_{t+1}^i = m_t + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^i \quad \text{for } i = 1, \ldots, \lambda. \tag{B.2}$$

Pour une matrice définie positive $\mathbf{C} \in \mathcal{S}_{++}^d$, sa racine carrée $\sqrt{\mathbf{C}}$ est l'unique matrice dans $\mathcal{S}_{++}^d$ telle que $\sqrt{\mathbf{C}}^2 = \mathbf{C}$. L'étape suivante est de classer les solutions candidates par rapport à leurs valeurs par f. Ici uniquement, CMA-ES utilise la fonction f à cette itération. Formellement, nous définissons une permutation $s_{t+1} \in \mathfrak{S}_\lambda$ satisfaisant

$$f\left(x_{t+1}^{s_{t+1}(1)}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{s_{t+1}(\lambda)}\right). \tag{B.3}$$

Le choix de la permutationquand les valeurs par f de deux solutions sont égales. Pour départager, nous imposons alors que si $i < j$ et $f(x_{t+1}^i) = f(x_{t+1}^j)$, alors $s_{t+1}^{-1}(i) < s_{t+1}^{-1}(j)$.

1.2 *Mise à jour de la moyenne*

Afin de mettre à jour la moyenne, nous prenons la moyenne des μ meilleurs enfants. L'entier $\mu \in \{1, \ldots, \lambda\}$ est appelé le nombre de parents, est est souvent choisi tel que $\mu \approx \lambda/2$. Aussi, nous fixons des poids $\mathbf{w}_m = (w_1, \ldots, w_\mu)$ satisfaisant $w_1 \geqslant \cdots \geqslant w_\mu > 0$ et $\sum_{i=1}^\mu w_i = 1$. Alors, la moyenne obéit à

$$m_{t+1} = \sum_{i=1}^\mu w_i x_{t+1}^{s_{t+1}(i)}. \tag{B.4}$$

Cependant, tout au long du manuscrit, nous considérons une formule plus générale [77]. Nous choisissons un taux d'apprentissage $c_m > 0$, et nous mettons à jour la moyenne ainsi:

$$m_{t+1} = (1 - c_m) m_t + c_m \sum_{i=1}^\mu w_i x_{t+1}^{s_{t+1}(i)} = m_t + c_m \sigma_t \sqrt{\mathbf{C}_t} \sum_{i=1}^\mu w_i U_{t+1}^{s_{t+1}(i)}. \tag{B.5}$$

Nous retrouvons d'ailleurs (B.4) lorsque $c_m = 1$, ce qui est la valeur par défaut en pratique pour c_m.

1.3 *Mise à jour du pas*

Nous mettons à jour le pas en utilisant le chemin p_t^σ, qui cherche à pondérer les meilleurs directions précédentes et à se comparer à son espérance sous sélection neutre i.e., si on ne classait pas les enfants en défénissant une permutation par (B.3). La mise à jour du chemin est:

$$p_{t+1}^\sigma = (1 - c_\sigma) p_t^\sigma + \sqrt{c_\sigma (2 - c_\sigma) \mu_{\text{eff}}} \sum_{i=1}^\mu w_i U_{t+1}^{s_{t+1}(i)}. \tag{B.6}$$

Le nombre positif $c_\sigma \in (0, 1]$ est le taux de décroissance pour le chemin. Quand $c_\sigma = 1$, seule la dernière itération est utilisée pour mettre à jour le pas et on dit alors qu'il n'y a pas d'accumulation

sur le pas. Aussi, la quantité $\mu_{\text{eff}} > 0$ est définie par $\mu_{\text{eff}} = \|\mathbf{w}_m\|_2^2$. Le pas alors mis à jour par une fonction abstraite de changement de pas $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$, i.e.,

$$\sigma_{t+1} = \sigma_t \times \Gamma(p_{t+1}^\sigma). \tag{B.7}$$

Un choix classique pour le changement de pas est l'adaptation du pas cumulative (CSA) [63]:

$$\Gamma_{\text{CSA}}^1\colon p \in \mathbb{R}^d \mapsto \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|p\|}{\mathbb{E}\|\mathcal{N}(0,\mathbf{I}_d)\|} - 1\right)\right). \tag{B.8}$$

La quantité $d_\sigma > 0$ (souvent appelée paramètre d'amortissement) est en pratique choisie proportionnelle à $\sqrt{\mu_{\text{eff}}}$. Nous nous intéressons dans cette thèse à l'alternative lisse suivante [68]:

$$\Gamma_{\text{CSA}}^2\colon p \in \mathbb{R}^d \mapsto \exp\left(\frac{c_\sigma}{2d_\sigma}\left(\frac{\|p\|^2}{d} - 1\right)\right). \tag{B.9}$$

Cependant, notre analyse considère des changements de pas plus généraux, incluant (B.8) et (B.9).

1.4 *Mise à jour de la matrice de covariance*

Comme pour le pas, la matrice de covariance $\mathbf{C}_t$ se met à jour grâce à un chemin p_t^c, tel que

$$p_{t+1}^c = (1-c_c)p_t^c + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\sqrt{\mathbf{C}_t}\sum_{i=1}^{\mu} w_i U_{t+1}^{s_{t+1}(i)}. \tag{B.10}$$

La matrice de covariance mise à jour doit favoriser l'échantillonnage dans la direction p_{t+1}^c. De plus, elle exploite les μ meilleures directions de l'itération présente. En tout, nous avons

$$\mathbf{C}_{t+1} = (1-c_1-c_\mu)\mathbf{C}_t + c_1\left[p_{t+1}^c\right]\left[p_{t+1}^c\right]^\top + c_\mu\sqrt{\mathbf{C}_t}\times\sum_{i=1}^{\mu} w_i^c\left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top\times\sqrt{\mathbf{C}_t}. \tag{B.11}$$

Les taux d'apprentissage $c_1, c_\mu \in [0,1]$ doivent satisfaire $c_1 + c_\mu \leqslant 1$. La mise à jour de la matrice de covariance utilise des poids $\mathbf{w}_c = (w_1^c, \ldots, w_\mu^c)$ qui ont les mêmes hypothèses que $\mathbf{w}_m$. La matrice $[p_{t+1}^c][p_{t+1}^c]^\top$ est habituellement appelée la mise à jour de rang un, et $\sqrt{\mathbf{C}_t}\times\sum_{i=1}^{\mu} w_i^c[U_{t+1}^{s_{t+1}(i)}][U_{t+1}^{s_{t+1}(i)}]^\top\times\sqrt{\mathbf{C}_t}$ la mise à jour de rang mu, en tant que matrice de rang $\min\{\mu, d\}$ presque sûrement.

1.5 *Résumé de CMA-ES*

L'algorithme CMA-ES est résumé ci-dessous et s'illustre par la figure B.1. Nous initialisons $m_0 \in \mathbb{R}^d$, $\sigma_0 > 0$, $\mathbf{C}_0 \in \mathcal{S}_{++}^d$, $p_0^\sigma \in \mathbb{R}^d$, $p_0^c \in \mathbb{R}^d$. Pour $t \in \mathbb{N}$, nous répétons ce qui suit jusqu'à atteindre un critère d'arrêt.

1. **Échantillonner** $U_{t+1}^1, \ldots, U_{t+1}^\lambda \sim \mathcal{N}(0, \mathbf{I}_d)$ i.i.d., indépendamment de $(m_t, \sigma_t, \mathbf{C}_t, p_t^\sigma, p_t^c)$.

2. **Classer** les solutions candidates par une permutation $s_{t+1} \in \mathfrak{S}_\lambda$ qui satisfait:

$$f\left(m_t + \sigma_t\sqrt{\mathbf{C}_t}U_{t+1}^{s_{t+1}(1)}\right) \leqslant \cdots \leqslant f\left(m_t + \sqrt{\mathbf{C}_t}\sigma_t U_{t+1}^{s_{t+1}(\lambda)}\right) \quad .$$

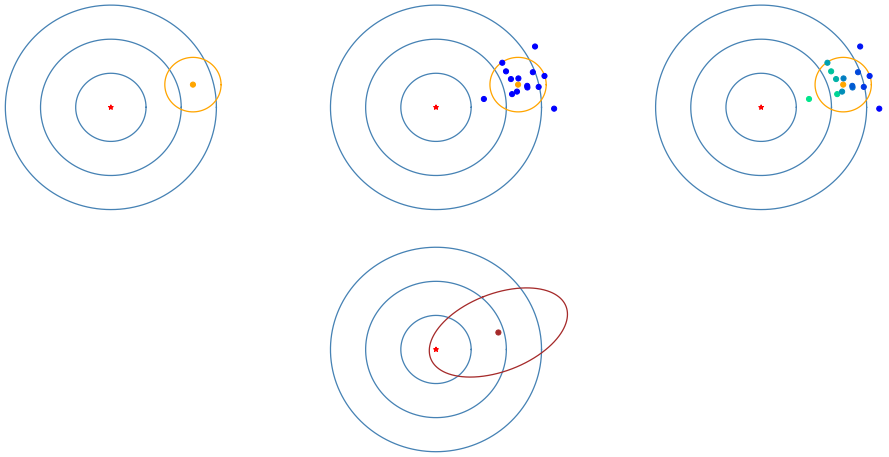


Figure B.1: Une itération de CMA-ES: la fonction à minimiser est représentée par ses lignes de niveaux en bleu et son minimum global en rouge. De gauche à droite: 0. la moyenne initiale m_t est en orange, et le cercle orange est la zone où les solutions candidates (qui suivent $\mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$) sont probables de se situer; 1. les solutions candidates $m_t + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^i$ sont en bleu; 2. après leur classement (selon f), les solutions candidates sont coloriées différemment; 3. la moyenne mise à jour m_{t+1} (marron) est obtenue par une moyenne pondérée des solutions candidates classées.

3. **Mettre à jour**:

$$
\begin{aligned}
m_{t+1} &= m_t + c_m \sigma_t \sqrt{\mathbf{C}_t} \textstyle\sum_{i=1}^{\mu} w_i U_{t+1}^{s_{t+1}(i)} \\
p_{t+1}^{\sigma} &= (1 - c_\sigma) p_t^{\sigma} + \sqrt{c_\sigma (2 - c_\sigma) \mu_{\text{eff}}} \textstyle\sum_{i=1}^{\mu} w_i U_{t+1}^{s_{t+1}(i)} \\
\sigma_{t+1} &= \sigma_t \times \Gamma(p_{t+1}^{\sigma}) \\
p_{t+1}^{c} &= (1 - c_c) p_t^{c} + \sqrt{c_c (2 - c_c) \mu_{\text{eff}}} \sqrt{\mathbf{C}_t} \textstyle\sum_{i=1}^{\mu} w_i U_{t+1}^{s_{t+1}(i)} \\
\mathbf{C}_{t+1} &= (1 - c_1 - c_\mu) \mathbf{C}_t + c_1 [p_{t+1}^{c}][p_{t+1}^{c}]^\top + c_\mu \sqrt{\mathbf{C}_t} \textstyle\sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right] \left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\mathbf{C}_t} \ .
\end{aligned}
$$

2 D'autres méthodes d'optimisation

Nous présentons ici d'autres algorithmes qui sont pertinents dans notre contexte.

2.1 *Algorithmes sans dérivées*

Les stratégies d'évolution n'utilisent pas les des dérivées de la fonction objectif. Ceci dit, il existe d'autres méthodes sans dérivées qui sont utiliséesen pratique. On se concentrera d'abord sur d'autres ES,avec ou sans matrice de covariance, puis sur le recuit simuléqui est dérivé de l'algorithme Metropolis-Hasting, et enfin sur l'algorithme Nelder-Mead.

2.1.1 *Stratégies d'évolution*

Nous présentons brièvement plusieurs ES qui ne seront pas analysés dans cette thèse. Il doit être noté que les ES sont divisés en deux catégories: les stratégies plus et virgule [68]. La stratégie plus correspond à des algorithmes *élitistes*, pour lesquels la valeur par f de la moyenne m_t ne fait

que décroître [124, Chapter C]. Au contraire, la stratégie virgule utilise une moyenne des solutions candidates pour mettre à jour m_{t+1}, qui n'est pas comparée à m_t [132, Section 5.2]. Ainsi la dénomination d'un ES dépend de quelle stratégie il exploite. Par exemple, CMA-ES présenté en section 1 est appelé $(\mu/\mu_w, \lambda)$-CMA-ES, où μ dest le nombre parent, μ_w indique que des poids de recombinaison sont utilisés, et λ est la taille de la population totale.

Nous présentons d'abord le (1+1)-ES avec adaptation du pas minimisant $f\colon \mathbb{R}^d \to \mathbb{R}$. Nous choisissons $\gamma \in (0,1)$ comme facteur de changement de pas fixé, nous initialisons $m_0 \in \mathbb{R}^d$ et $\sigma_0 > 0$, et nous procédons comme suit pour $t \in \mathbb{N}$.

1. **Échantillonner** $x_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 \mathbf{I}_d)$.
2. **Si** $f(x_{t+1}) \leqslant f(m_t)$: $m_{t+1} = x_{t+1}$ and $\sigma_{t+1} = \gamma^{4/5}\sigma_t$.
3. **Sinon**, $m_{t+1} = m_t$ et $\sigma_{t+1} = \gamma^{-1/5}\sigma_t$.

Il existe aussi un (1+1)-CMA-ES [83]. Nous posons $\gamma \in (0,1)$, nous initialisons $m_0 \in \mathbb{R}^d$, $p_0 \in \mathbb{R}^d$, $\sigma_0 > 0$ et $\mathbf{C}_0 \in \mathcal{S}_{++}^d$, et pour $t \in \mathbb{N}$:

1. **Échantillonner** $U_{t+1} \sim \mathcal{N}(0, \mathbf{I}_d)$ i.i.d., indépendamment de $(m_t, p_t, \sigma_t, \mathbf{C}_t)$ et calculer les solutions candidates $x_{t+1} = m_t + \sigma_t \mathbf{C}_t^{1/2} U_{t+1}$.
2. **Si** $f(x_{t+1}) \leqslant f(m_t)$:

$$\begin{aligned} m_{t+1} &= x_{t+1} \\ \sigma_{t+1} &= \gamma^{4/5}\sigma_t \\ p_{t+1} &= (1-c_c)p_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\sqrt{\mathbf{C}_t}U_{t+1} \\ \mathbf{C}_{t+1} &= (1-c_1)\mathbf{C}_t + c_1 p_{t+1}p_{t+1}^\top \quad . \end{aligned}$$

3. **Sinon**:

$$\begin{aligned} m_{t+1} &= m_{t+1} \\ \sigma_{t+1} &= \gamma^{-1/5}\sigma_t \\ p_{t+1} &= p_t \\ \mathbf{C}_{t+1} &= \mathbf{C}_t \quad . \end{aligned}$$

Dernièrement, nous introduisons un $(\mu/\mu_w, \lambda)$-ES sans matrice de covariance [76]. Nous initialisons $m_0 \in \mathbb{R}^d$ et $\sigma_0 > 0$, et pour $t \in \mathbb{N}$ nous répétons ce qui suit.

1. **Échantillonner** $U_{t+1}^1, \ldots, U_{t+1}^\lambda \sim \mathcal{N}(0, \mathbf{I}_d)$ i.i.d., indépendamment de $(m_t, \sigma_t, C_t, p_t^\sigma, p_t^c)$.
2. **Classer** les solutions candidates par la permutation $s_{t+1} \in \mathfrak{S}_\lambda$ qui satisfait:

$$f\left(m_t + \sigma_t U_{t+1}^{s_{t+1}(1)}\right) \leqslant \cdots \leqslant f\left(m_t + \sigma_t U_{t+1}^{s_{t+1}(\lambda)}\right) \quad .$$

3. **Mettre à jour**:

$$\begin{aligned} m_{t+1} &= m_t + \sigma_t \textstyle\sum_{i=1}^\mu w_i U_{t+1}^{s_{t+1}(i)} \\ \sigma_{t+1} &= \sigma_t \times \Gamma\left(\sqrt{\mu_{\text{eff}}}\textstyle\sum_{i=1}^\mu w_i U_{t+1}^{s_{t+1}(i)}\right) \quad . \end{aligned}$$

Cet algorithme peut être obtenu à partir d'un $(\mu/\mu_w, \lambda)$-CMA-ES avec $c_c = c_\sigma = c_m = 1$ et $c_1 = c_\mu = 0$.

2.1.2 *Metropolis-Hastings et le recuit simulé*

Les méthodes de Monte-Carlo par chaînes de Markov (MCMC) [25] et en particulier l'algorithme Metropolis-Hastings [106, 78] cherchent à générer des nombres aléatoires suivant une loi cible π sur $\mathbb{R}^d$ dont la densité $p(\cdot)$ n'est connue qu'à une constante multiplicative près. Le principe est de définir une chaîne de Markov $\{z_t\}_{t\in\mathbb{N}}$ qui admet π comme unique distribution de probabilité invariante. Cela se rapproche de notre travail pour deux raisons: comme présenté ci-dessous, Metropolis-Hastings peut être adapté en un recuit simulé pour résoudre des problèmes d'optimisation, et l'analyse théorique des MCMC repose essentiellement sur les résultats de convergence de chaînes de Markov – tout comme la preuve de convergence de CMA-ES présentée dans cette thèse.

L'algorithme Metropolis-Hastings est résumé ci-dessous. Étant donné $z_0 \in \mathbb{R}^d$, nous répétons pour $t = 0, 1, \ldots$:

1. **Échantillonner** $y_{t+1} \sim \nu_{z_t}(\cdot)$ et $U_{t+1} \sim \mathrm{Uniform}([0,1])$ indépendamment de z_t;

2. **Mettre à jour** $z_{t+1} = z_t + (y_{t+1} - z_t)\,\mathbb{1}\left\{U_{t+1} \leqslant \frac{p(y_{t+1})}{p(z_t)}\right\}$.

À l'étape 1., nous utilisons une distribution de probabilité $\nu_{z_t}(\cdot)$ avec une densité positive sur $\mathbb{R}^d$, telle que $\nu_z(y) = \nu_y(z)$, e.g., $\nu_z = \mathcal{N}(z, \mathbf{I}_d)$. Le recuit simulé [96] est une méthode dérivée de Metropolis-Hastings pour l'optimisation. Bien qu'initialement prévue pour des problèmes discrets (comme le problème du voyageur de commerce [28]), la version suivante du recuit simulé cherche à minimiser une fonction objectif $f\colon \mathbb{R}^d \to \mathbb{R}$ [24, 37]. Étant donné $m_0 \in \mathbb{R}^d$, nous répétons pour $t = 0, 1, \ldots$:

1. **Échantillonner** $x_{t+1} \sim \mathcal{N}(m_t, \mathbf{I}_d)$ and $U_{t+1} \sim \mathrm{Uniform}([0,1])$ indépendamment de m_t;

2. **Choisir** $m_{t+1} = m_t + (x_{t+1} - m_t)\,\mathbb{1}\left\{U_{t+1} \leqslant \exp\left(-\frac{f(x_{t+1}) - f(m_t)}{T_t}\right)\right\}$.

Les constantes positives T_t pour $t \in \mathbb{N}$ sont des paramètres de température et sont choisis tels que T_t tendent vers 0 quand t tend vers $+\infty$ pour assurer la convergence. Cependant, contrairement aux ES, le recuit simulé ne converge pas linéairement – à une vitesse géométrique. Aussi, il n'exploite pas uniquement des comparaisons de valeurs par f, comme pour les ES, puisqu'il utilise les valeurs de la fonction pour calculer la probabilité d'acceptation à l'étape 2.

2.1.3 *Nelder-Mead*

Nous mentionnons ici l'algorithme Nelder-Mead [115]. Bien qu'il soit déterministe, il se rapproche de CMA-ES de plusieurs manières. Il s'agit d'un algorithme sans dérivées et repose sur la comparaison de valeurs de la fonction. En plus, comme CMA-ES pour certains choix d'hyperparamètres, il est affine-invariant [99, Lemma 3.2]. De plu, il a été prouvé qu'il convergeait linéairement pour une classe de fonctions strictement convexes en dimensions 1 et 2 [99]. Le principe de Nelder-Mead algorithm est de mettre à jour un simplexe $\theta = (x_1, \ldots, x_n)$ pour lequel nous remplaçons à chaque itération le point de θ avec la plus haute f-valeur par un point transformé (soit réfléchi, étendu ou contracté par rapport au centroïde du simplexe), ou en cas d'échec en réduisant tous les points sauf celui avec la plus faible f-valeur.

2.2 *Méthodes de Newton*

Dans cette section, nous dressons un parallèle entre CMA-ES et les méthodes de Newton pour l'optimisation. À l'origine, la méthode de Newton exacte approche itérativement la fonction objectif f par une fonction quadratique en utilisant l'information de second ordre. Cependant, cela requiert de résoudre un système linéaire ce qui implique un coût élevé lorque la dimension du problème est grande. Pour résoudre cel, des approximations de la méthode de Newton, habituellement appelées méthodes quasi-Newton, sont apparues. Elles approchent ainsi l'inverse de la Hessienne de f. Comme pour CMA-ES (si on choisit bien ses hyperparamètres), les méthodes quasi-Newton sont affine-invariantes. De plus, un résultat de cette thèse est l'apprentissage par CMA-ES de l'inverse de la Hessienne.

Nous donnons d'abord une formulation général pour les méthodes de Newton. Pour une fonction objectif deux fois différentiable $f\colon \mathbb{R}^d \to \mathbb{R}$, nous initialisons $x_0 \in \mathbb{R}^d$ et $\mathbf{B}_0 \in \mathcal{S}^d_{++}$. Nous répétons pour $t \in \mathbb{N}$ ce qui suit.

1. **Mettre à jour** $x_{t+1} = x_t - \sigma_t \mathbf{B}_t \nabla f(x_k)$.
2. **Mettre à jour** $\mathbf{B}_{t+1} = F_B(x_{t+1}, x_t, \mathbf{B}_t; f)$.

Nous ne précisons pas comment le pas $\sigma_t > 0$ est choisi.[2] La fonction de mise à jour F_B est alors spécifique à une méthode de Newton donnée, avec deux exemples ci-dessous. Une condition suffisant pour l'affine-invariance est que

$$F_B\left(\mathbf{A}^{-1}(x_{t+1} - a), \mathbf{A}^{-1}(x_t - a), \mathbf{A}^{-1}\mathbf{B}_t\mathbf{A}^{-\top}; f(\mathbf{A}\cdot + a)\right) = \mathbf{A}^{-1} F_B(x_{t+1}, x_t, \mathbf{B}_t; f)\mathbf{A}^{-\top} \tag{B.12}$$

pour tout $\mathbf{A} \in \mathsf{GL}_d(\mathbb{R})$ et $a \in \mathbb{R}^d$. Cependant, cela n'implique qu'on apprend l'information de second ordre. Par example, si $F_B(x_{t+1}, x_t, \mathbf{B}_t; f) = \mathbf{B}_t$, alors (B.12) est vraie mais la matrice $\mathbf{B}_t$ reste constante et donc ne peut pas approcher l'inverse de la Hessienne de f.

2.2.1 *Méthode de Newton exacte*

La méthode de Newton exacte est obtenue à partir du précédent algorithme lorque $F_B(x_{t+1}, x_t, \mathbf{B}_t; f) = \nabla^2 f(x_{t+1})^{-1}$. Quand la Hessienne $\nabla^2 f$ est lipschitzienne, alors la méthode de Newton converge vers un point stationnaire à une vitesse quadratique [116, Theorem 3.5].

2.2.2 *BFGS*

Un algorithme quasi-Newton populaire est nommé d'après Broyden, Fletcher, Goldfarb et Shanno (BFGS) [26, 44, 56, 134]. L'algorithme BFGS utilise alors la fonction de mise à jour suivante :

$$F_B(x_{t+1}, x_t, \mathbf{B}_t; f) = \left(\mathbf{I}_d - \frac{\Delta x_t \Delta d_t^\top}{\Delta d_t^\top \Delta x_t}\right) \mathbf{B}_t \left(\mathbf{I}_d - \frac{\Delta d_t \Delta x_t^\top}{\Delta d_t^\top \Delta x_t}\right) + \frac{\Delta x_t \Delta x_t^\top}{\Delta d_t^\top \Delta x_t} \tag{B.13}$$

où $\Delta x_t = x_{t+1} - x_t$ et $\Delta d_t = \nabla f(x_{t+1}) - \nabla f(x_t)$. En plus d'être affin-invariant, BFGS apprend l'inverse de la Hessienne d'une fonction convexe quadratique ne moins de d itérations [116, Theorem 6.4] et converge super-linéairement vers un minimum d'une fonction convexe deux fois différentiable avec une Hessienne lipschitzienne [116, Theorem 6.6].

[2] D'habitude pour assurer la convergence, une recherche linéaire satisfaisant la condition de Wolfe est requise [116, Eq. (3.7)].

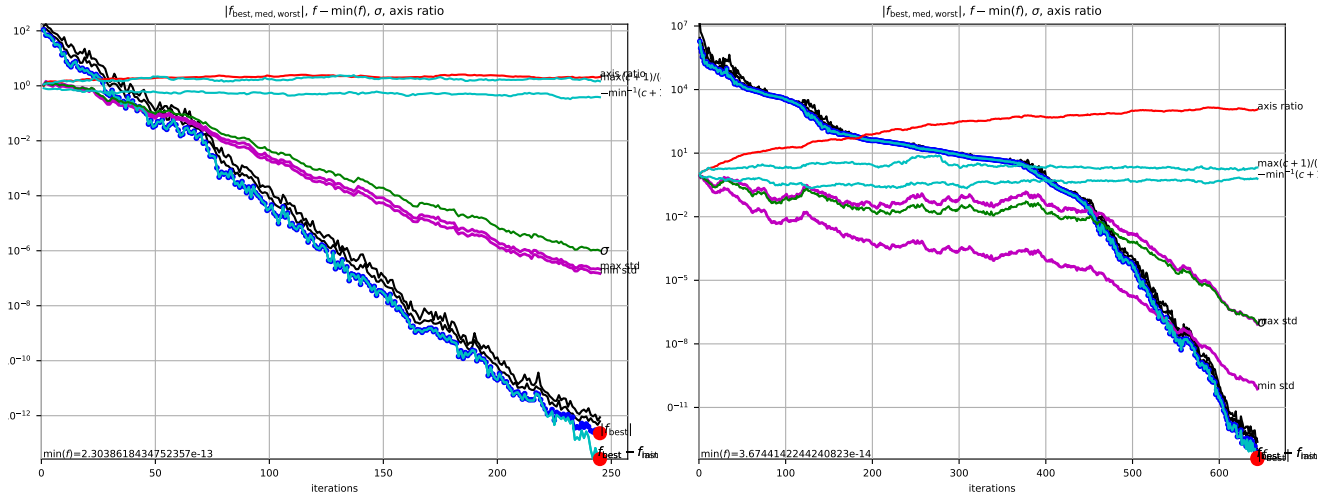


Figure B.2: Résultats empirique de CMA-ES minisant la fonction sphère (gauche) ou une fonction ellipsoïde mal conditionnée ($\mathrm{Cond}(\mathbf{H}) = 10^6$) (droite) en dimension 10. Les lignes bleues et noires montrent les f-valeurs de solutions candidates évaluées pendant l'optimisation. Puisque l'axe des ordonnées est dans une échelle logarithmique, une décroissance linéaire est interprétée comme une convergence linéaire.

3 Convergence linéaire

Il a été observé [76, 73] que CMA-ES converge linéairement (ou géométriquement) avec probabilité élevée vers le minimum global x^* de beaucoup de problèmes. Dans les figures B.2 et 3, nous avons des résultats numériques (produits avec le package Python `pycma` [66]) avec cette observation. Nous pouvons aussi voir (partie droite de la figure B.3) que pour des fonctions multi-modales, CMA-ES peut converger vers des minima locaux uniquement. Mathématiquement, la convergence linéaire de CMA-ES s'écrit par la limite presque sûre suivante :

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{\|m_t - x^*\|}{\|m_0 - x^*\|} = \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|m_t - x^*\|}{\|m_{t-1} - x^*\|}\right] = -\mathrm{CR} \tag{B.14}$$

où le taux de convergence CR est une constante positive. Nous prouvons cette limite dans thèse sous des hypothèses diverses (notamment la fonction objectif doit avoir des lignes de niveau ellipsoïdales). Nous prouvons aussi que CR est positif lorsque le changement de pas est donné par (B.9) et en choisissant correctement les hyperparamètres de CMA-ES.

De plus, il semble que CMA-ES approche l'information de second ordre de f afin de résoudre de hauts conditionnements. Effectivement, les taux de convergence de CMA-ES minimisant une fonction sphère ou une fonction ellipsoïde mal conditionnée semblent être égaux, voir la figure B.2. En plus de cette égalité, nous prouvons la limite suivante pour une constante $\rho > 0$:

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\mathbf{C}_t}{\det(\mathbf{C}_t)^{1/d}}\right] = \rho \mathbf{H}^{-1} \tag{B.15}$$

quand la fonction f est ellipsoïdale, i.e., une transformation croissante d'une fonction convexe quadratique $x \mapsto x^\top \mathbf{H} x$, avec $\mathbf{H}$ une matrice définie positive. Nous faisons l'analogie avec les fonctions quasi-convexes qui sont des transformations croissantes de fonctions convexes: la matrice $\mathbf{H}$ est la Hessienne d'une transformation croissante de f, et sera donc appelée la quasi-Hessienne de f.

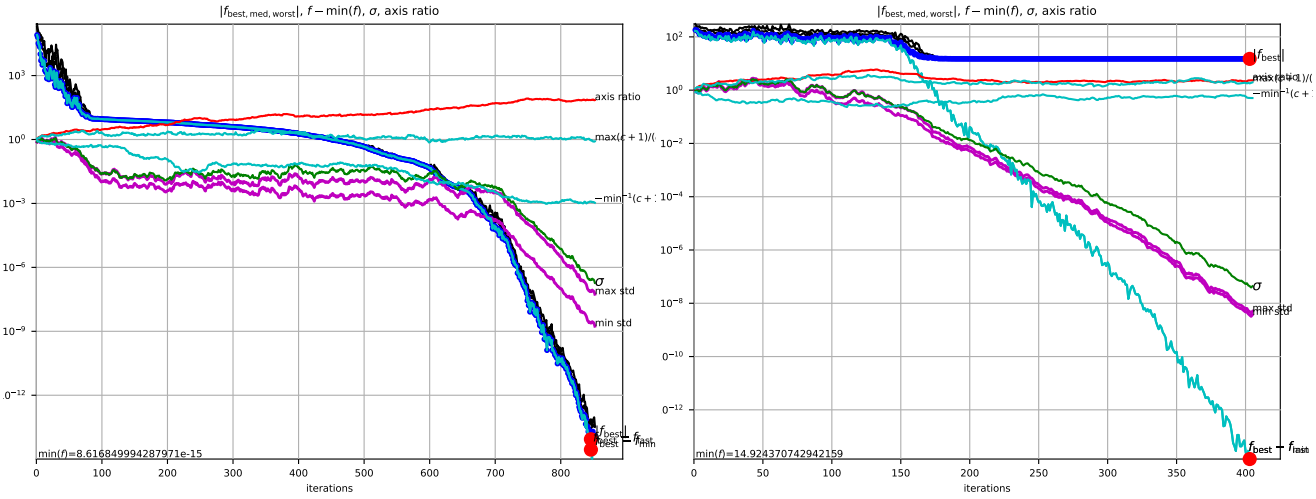


Figure B.3: Résultats empirique de CMA-ES minisant la fonction Rosenbrock (gauche) ou la fonction Rastrigin (droite) en dimension 10. Les lignes bleues et noires montrent les f-valeurs de solutions candidates évaluées pendant l'optimisation. Puisque l'axe des ordonnées est dans une échelle logarithmique, une décroissance linéaire est interprétée comme une convergence linéaire.

$$\begin{array}{ccc} (m_t, \sigma_t, \mathbf{C}_t) & \xrightarrow{\min f} & (m_{t+1}, \sigma_{t+1}, \mathbf{C}_{t+1}) \\ \Phi_{f\to\hat{f}} \downarrow & & \uparrow \Phi^{-1}_{f\to\hat{f}} \\ (\hat{m}_t, \hat{\sigma}_t, \hat{\mathbf{C}}_t) & \xrightarrow{\min \hat{f}} & (\hat{m}_{t+1}, \hat{\sigma}_{t+1}, \hat{\mathbf{C}}_{t+1}) \end{array}$$

Figure B.4: Diagramme commutatif de CMA-ES.

Cette dernière observation est essentiellement une conséquence de l'affine-invariance, illustrée en figure B.4, et établie pour certaines versions de CMA-ES [70, 15], et en particulier vraie dans la version présentée ici lorsque le taux de décroissance c_σ est égal à un.

Les résultats de convergence linéaire de CMA-ES et de son apprentissage d'information de second ordre sont exposés au chapitre 5.

4 Introduction aux chaînes de Markov

Cette section rappelle les fondamentaux nécessaires pour cette thèse sur les chaînes de Markov et leurs noyaux de transition. En plus des ES, l'étude des chaînes de Markov et de leur convergence a de nombreuses applications. Par exemples les méthodes MCMC [25] cherchent à générer depuis un loi $\pi(\cdot)$ connue à une constane multiplicative près. La convergende d'une chaîne de MArkov à une vitesse géométrique vers sa loi stationnaire π peut être établie grâce à des critères de dérive tels que le théorème 1 ci-dessous. Ceci a déjà été utilisé pour l'algorithme Metropolis-Hastings [128] ou des MCMC adaptatifs [127].

Nous nous référons à un livre plus spécialisé [110] pour les lecteurs intéressés.

Soit $(\Omega, \mathcal{F})$ un espace mesurable. L'ensemble Ω est appelé *univers*, et $\mathcal{F}$ est une tribu sur Ω,

appelée *ensemble d'événements*. Un *événement* est un élément de $\mathcal{F}$. Nous équipons $(\Omega, \mathcal{F})$ d'une probabilité $\mathbb{P}$. La probabilité d'un événement $\mathsf{W} \in \mathcal{F}$ est donc $\mathbb{P}[\mathsf{W}]$. Une variable aléatoire à valeurs dans l'ensemble mesurable $(\mathsf{X}, \mathcal{X})$ est une fonction mesurable $X : \Omega \to \mathsf{X}$. Si ν est une mesure sur $\mathcal{X}$, on dit que X suit ν quand

$$\mathbb{P}[X \in \mathsf{A}] \coloneqq \mathbb{P}\left[\{\omega \in \Omega \mid X(\omega) \in \mathsf{A}\}\right] = \nu(\mathsf{A}) \quad \text{for } \mathsf{A} \in \mathcal{X}. \tag{B.16}$$

4.1 Noyaux et chaînes de Markov

Nous supposons que l'espace d'états X est un espace topologique, localement compact, séparable et métrisable (e.g., un espace euclidien). Nous notons $\mathcal{B}(\mathsf{X})$ la tribu borélienne de X. Un noyau sur $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ est une fonction $P : \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}_+$ telle que

(a) pour tout $x \in \mathsf{X}$, $\mathsf{A} \in \mathcal{B}(\mathsf{X}) \mapsto P(x, \mathsf{A})$ est une mesure,

(b) pour tout $\mathsf{A} \in \mathcal{B}(\mathsf{X})$, $x \in \mathsf{X} \mapsto P(x, \mathsf{A})$ est mesurable.

Quand pour tout $x \in \mathsf{X}$, $P(x, \mathsf{X}) = 1$,on dit que Pest un noyau de transition [110, Chapter 3]. On dit que P est un noyau sous-stochastique quand $P(x, \mathsf{X}) \leqslant 1$ pour $x \in \mathsf{X}$.

Dans la suite, P est un noyau de transition sur $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. Afin de définir une chaîne de Markov homogène $\{\theta_k\}_{k\in\mathbb{N}}$ sur X avec noyau de transition P, nous équipons l'univers $(\Omega, \mathcal{F})$ d'une probabilité $\mathbb{P}[\cdot \mid \theta_0 \sim \nu]$, où ν est une probabilité sur $\mathcal{B}(\mathsf{X})$. Alors, nous définissons une suite $\{\theta_k\}_{k\in\mathbb{N}}$ telle que la variable aléatoire θ_0 suit ν, et pour tout $k \in \mathbb{N}^*$, la variable aléatoire θ_k est telle que

$$\mathbb{P}[\theta_k \in \mathsf{A} \mid \theta_0 \sim \nu] = \nu P^k(\mathsf{A}) \coloneqq \int_{\mathsf{X}^{k+1}} P(x^k, \mathsf{A}) P(x^{k-1}, \mathrm{d}x^k) \dots P(x^0, \mathrm{d}x^1) \nu(\mathrm{d}x^0) \tag{B.17}$$

pour tout $\mathsf{A} \in \mathcal{B}(\mathsf{X})$. Pour tout $x \in \mathsf{X}$, nous définissons le noyau de transition à k étapes $P^k(x, \cdot) = \delta_x P^k(\cdot)$, où δ_x est la mesure de Dirac en x.

Nous définissons ci-dessous des notions essentielles pour la convergence des chaînes de Markov.

4.2 Irréductibilité

L'irréductibilité est centrale dans la convergence des chaînes de Markov [110, Chapter 4]. Sur des expaces d'états finis, une chaîne de Markov est irréductible quand tout état est accessible en un nombre fini d'étapes avec probabilité positive, quel que soit l'état de départ. Pour des espaces d'états plus généraux, nous définissons l'irréductibilité par rapport à une mesure φ. Pour une mesure positive φ qur $\mathcal{B}(\mathsf{X})$, le support de φ est l'ensemble $\operatorname{supp} \varphi$ comprenant les sous-ensembles $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ tels que $\varphi(A) > 0$.

Définition 1. Le noyau de transition P sur $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ est dit *irréductible* s'il existe une mesure non-triviale φ sur $\mathcal{B}(\mathsf{X})$ telle que pour tout $\mathsf{A} \in \operatorname{supp} \varphi$ et tout $x \in \mathsf{X}$, nous avons

$$\sum_{k\geqslant 1} P^k(x, \mathsf{A}) > 0.$$

Nous introduisons aussi la mesure d'irréductibilité maximale.

Définition 2. Soit Pun noyau de transition irréductible sur $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. Une *mesure d'irréductibilité maximale* de P est une mesure d'irréductibilité ψ de P telle que pour toute

mesure d'irréductibilité φ de P, nous avons

$$\operatorname{supp} \varphi \subset \operatorname{supp} \psi.$$

4.3 *Périodicité, apériodicité*

Pour un noyau irréductible P sur $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ avec mesure d'irréductibilité maximale ψ, un *d-cycle* est la collection d'ensemble disjoints $\{\mathsf{A}_1, \dots, \mathsf{A}_d\}$ de $\mathcal{B}(\mathsf{X})$ telle que

(1) pour $k = 1, \dots, d-1$ et $x_k \in \mathsf{A}_k$, $P(x_k, \mathsf{A}_{k+1}) = 1$, et pour $x_d \in \mathsf{A}_d$, $P(x_d, \mathsf{A}_1) = 1$;

(2) le support de ψ est inclus dans $\mathsf{A}_1 \cup \cdots \cup \mathsf{A}_d$.

Définition 3 ([110, Chapter 5])**.** La *période* d'un noyau irréductible P est le plus grand entier d tel qu'un d-cycle existe. Quand la période de P est égale à 1, on dit que P est *apériodique*.

4.4 *Ensembles petits et small, T-chaînes*

Les ensembles petits et small [110, Chapter 5] sont eux aussi importants pour la convergence des chaînes de Markov. Ci-dessous, au théorème 1, nous donnons une *condition de dérive* pour la convergence des chaînes de Markov qui repose sur les ensembles small.

Définition 4. Soient P un noyau de transition $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ et $\mathsf{C} \in \mathcal{B}(\mathsf{X})$.

(i) S'il existe un probabilité b sur $\mathbb{N}^*$ et une mesure non-triviale ν_b sur $\mathcal{B}(\mathsf{X})$ telles que

$$\sum_{k \geqslant 1} b(k) P^k(x, \mathsf{A}) \geqslant \nu_b(\mathsf{A}) \quad \text{pour tout } x \in \mathsf{C} \text{ et } \mathsf{A} \in \mathcal{B}(\mathsf{X}),$$

alors C est dit *petit*.

(ii) S'il existe $a \in \mathbb{N}^*$ et une mesure non-triviale ν_a on $\mathcal{B}(\mathsf{X})$ tels que

$$P^a(x, \mathsf{A}) \geqslant \nu_a(\mathsf{A}) \quad \text{pour tout } x \in \mathsf{C} \text{ et } \mathsf{A} \in \mathcal{B}(\mathsf{X}),$$

alors C est dit *small*.

Nous définissons à présent les T-chaînes [110, Chapter 6], qui ont des propriétés topologiques intéressantes. Cela facilite l'identification des ensembles petits : pour une T-chaîne, les compacts sont petits [110, Theorem 6.2.5].

Définition 5. Soient P un noyau de transition irréductible et T un noyau sous-stochastique sur $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$.

(i) Si $x \in \mathsf{X} \mapsto T(x, \mathsf{A})$ est semi-continu inférieurement pour tout $\mathsf{A} \in \mathcal{B}(\mathsf{X})$, et s'il existe une mesure b sur $\mathbb{N}$ telle que

$$\sum_{k \in \mathbb{N}} b(k) P^k(x, \mathsf{A}) \geqslant T(x, \mathsf{A}) \quad \text{for } x \in \mathsf{X} \text{ et } \mathsf{A} \in \mathcal{B}(\mathsf{X}),$$

alors T est appelé *composante continue* de P.

(ii) Si P possède une composante continue T telle que $T(x, \mathsf{X}) > 0$ pour tout $x \in \mathsf{X}$, alors P est le noyau d'une *T-chaîne*.

4.5 *Récurrence, récurrence Harris*

Une des étapes centrales de notre preuve de convergence est de prouver qu'une chaîne de Markov normalisée sous-jacente à CMA-ES est récurrente. La définition de la récurrence est associée au temps d'occupation d'une chaîne de Markov. Pour $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ at $x \in \mathsf{X}$, nous notons $N_{\mathsf{A}}(x)$ le temps d'occupation espéré par $\{\theta_k\}_{k\in\mathbb{N}}$ conditionnellement à $\theta_0 = x$, défini par :

$$N_{\mathsf{A}}(x) = \sum_{k\geqslant 1} P^k(x, \mathsf{A}) \ . \tag{B.18}$$

Définition 6 ([110, Chapter 8]). Soit P un noyau de transition sur $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$.

(i) Un ensemble $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ est *récurrent* quand pour tout $x \in \mathsf{A}$, $N_{\mathsf{A}}(x) = +\infty$.

(ii) Supposons que P est irréductible et notons ψ sa mesure d'irréductibilité maximale. Alors, P est *récurrent* si tout $\mathsf{A} \in \operatorname{supp} \psi$ est récurrent.

De même, nous pouvons définit des ensembles transients quand ils sont visités qu'un nombre fini de fois (en espérance), et les noyaux transients qui sont des noyaux irréductibles qui ne sont pas récurrents. Cependant, seuls les noyaux récurrents seront d'intérêt dans cette thèse. Une notion plus forte est la récurrence Harris. Nous définissons la variable aléatoire η_{A} comme le temps d'occupation en A par

$$\eta_{\mathsf{A}} = \sum_{k\geqslant 1} \mathbb{1}_{\theta_k\in\mathsf{A}} \ . \tag{B.19}$$

Définition 7 ([110, Chapter 9]). Soit P un noyau de transition sur $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$.

(i) Un ensemble $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ est *Harris récurrent* quand pour tout $x \in \mathsf{A}$, on a $\eta_{\mathsf{A}} = +\infty$ presque sûrement.

(ii) Supposons que P est irréductible et notons ψ sa mesure d'irréductibilité maximale. Alors, P est *Harris récurrent* si tout $\mathsf{A} \in \operatorname{supp} \psi$ est Harris récurrent.

4.6 *Mesures invaiantes*

Nous définissons maintenant les mesures invariantes de noyaux de transition.

Définition 8 ([110, Chapter 10]). Soit P un noyau de transition sur $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. Une mesure π sur $\mathcal{B}(\mathsf{X})$ est *invariante* quand $\pi P = \pi$. Lorsqu'une mesure de probabilité invariante sur $\mathcal{B}(\mathsf{X})$ existe, alors P est dit *positif*.

4.7 *Ergodicité*

L'ergodicité [110, Chapter 13] caractérise la convergence d'une chaîne de Markov vers une mesure stationnaire. Nous nous intéressons à l'ergodicité géométrique [110, Chapter 15] dans cette thèse. Soit $V : \mathsf{X} \to [1, +\infty]$ une fonction mesurable appelée *potentiel*. Étant donnée ν une mesure (signée) sur $\mathcal{X}$, nous posons

$$\|\nu\|_V = \sup_{|g|\leqslant V} \int g \, \mathrm{d}\nu. \tag{B.20}$$

Définition 9. Soit P un noyau de transition sur $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ et $V : \mathsf{X} \to [1, +\infty]$ une fonction mesurable. Nous disons que P est *V-géométriquement ergodique* quand il est positif Harris récurrent avec mesure de probabilité π, avec $V \in \mathsf{L}^1(\pi)$, et quand il existe $r > 0$ tel que pour tout $x \in \mathsf{X}$

$$\sum_{k \geqslant 1} r^k \left\| P^k(x, \cdot) - \pi(\cdot) \right\|_V < +\infty \quad .$$

Quand P est 1-géométriquement ergodique, nous disons juste qu'il est *géométriquement ergodique*.

Notre preuve de convergence s'appuie sur le critère d'ergodicité suivant, connu comme condition de dérive géométrique, ou condition de Foster-Lyapunov pour l'ergodicité. Cependant, remarquons qu'on utilise une condition *état-dépendante*, puisque la condition de dérive classique prendrait $n(x) = 1$ dans le théorème ci-dessous.

Théorème 1 ([110, Theorem 19.1.3])**.** Supposons que P est un noyau de transition irréductible et apériodique sur $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. Soient $n : \mathsf{X} \to \mathbb{N}^*$ et $V : \mathsf{X} \to [1, +\infty]$ deux fonctions mesurables. S'il existe un ensemble small $\mathsf{C} \in \mathcal{B}(\mathsf{X})$ tel que V est bornée sur C, et des constantes positives $\rho < 1$ et $b < +\infty$ telles que

$$\int P^{n(x)}(x, \mathrm{d}y) V(y) \leqslant \rho^{n(x)} \times (V(x) + b \mathbb{1}_{x \in \mathsf{C}}) \quad \text{pour } x \in \mathsf{X}, \tag{B.21}$$

alors, P est géométriquement ergodique. De plus, nous avons

$$\sum_{k \geqslant 1} r^k \left\| P^k(x, \cdot) - \pi(\cdot) \right\|_1 \leqslant R V(x) \quad \text{for } x \in \mathsf{X}, \tag{B.22}$$

pour des constantes $R < \infty$ et $r > 1$, où π est l'unique mesure de probabilité invariante de P.

4.8 Stabilité de chaînes de Markov non-linéaires par l'analyse d'un modèle de contrôle déterministe

Dans cette section, nous rappelons des résultats [109, 29] qui donnent des conditions suffisantes pour qu'une chaîne de Markov soit une T-chaîne irréductible et apériodique.

Soient X et W des espaces localement compacts, séparables, métrisables équipées avec leur tribus boréliennes respectives $\mathcal{B}(\mathsf{X})$ et $\mathcal{B}(\mathsf{W})$, et soit $(\mathsf{U}, \mathcal{U})$ un espace mesurables. Considérons $\{\phi_k\}_{k \in \mathbb{N}}$ une chaîne de Markov avec espace d'états X tel que

$$\phi_{k+1} = F\left(\phi_k, \alpha(\phi_k, U_{k+1})\right) \tag{B.23}$$

où $F : \mathsf{X} \times \mathsf{W} \to \mathsf{X}$ et $\alpha : \mathsf{X} \times \mathsf{U} \to \mathsf{W}$ sont des fonctions mesurables, et $\{U_{k+1}\}_{k \in \mathbb{N}}$ un processus i.i.d. à valeurs dans U indépendant de ϕ_0.

Nous définissons la *fonction de transition étendue* S_x^k pour $x \in \mathsf{X}$ et $k \in \mathbb{N}$ par

$$\begin{aligned} &S_x^0 = x \\ &S_x^{k+1}(w_{1:k+1}) = F\left(S_x^k(w_{1:k}), w_{k+1}\right) \quad \text{pour } w_{1:k+1} = (w_1, \ldots, w_{k+1}) \in \mathsf{W}^{k+1}. \end{aligned} \tag{B.24}$$

De même, nous définissons la *densité de probabilité étendue* p_x^k pour $x \in \mathsf{X}$ et $k \geqslant 1$ par

$$\begin{aligned} &p_x^1(w_1) = p_x(w_1) \quad \text{pour } w_1 \in \mathsf{W} \\ &p_x^{k+1}(w_{1:k+1}) = p_x^k(w_{1:k}) p_{S_x^k(w_{1:k})}(w_{k+1}) \quad \text{for } w_{1:k+1} = (w_1, \ldots, w_{k+1}) \in \mathsf{W}^{k+1}, \end{aligned} \tag{B.25}$$

où $p_x(\cdot)$ représente une densité de probabilité par rapport à une mesure σ-finie ζ_{W} sur W de la variable aléatoire $\alpha(x, U_1)$. Dans cette section, nous supposons que ζ_{W} est la mesure de Lebesgue sur un ouvert W de $\mathbb{R}^p$. De plus, nous supposons que $(x, w) \mapsto p_x(w)$ est semi-continue inférieurement sur $\mathsf{X} \times \mathsf{W}$. Nos définissons les ensembles de contrôle :

$$\mathcal{O}_x^k = \{w_{1:k} \in \mathsf{W}^k \mid p_x^k(w_{1:k}) > 0\} \ . \tag{B.26}$$

Alors, l'ensemble des états attaignables par $\{\phi_k\}_{k\in\mathbb{N}}$ commençant en $x \in \mathsf{X}$ est

$$A_+(x) = \bigcup_{k\in\mathbb{N}} A_+^k(x) := \bigcup_{k\in\mathbb{N}} \left\{S_x^k(w_{1:k}) \mid w_{1:k} \in \mathcal{O}_x^k\right\} \ . \tag{B.27}$$

Quand un état $x^* \in \mathsf{X}$ satisfait

$$x^* \in \bigcap_{x\in\mathsf{X}} \overline{A_+(x)} \ , \tag{B.28}$$

on dit que x^* est un *état globalement attracteur*. Quand F est différentiable, nous définissons pour $x \in \mathsf{X}$ la *condition de controllabilité* suivante :

$$\text{il existe } k \geqslant 1 \text{ et } w_{1:k} \in \mathcal{O}_x^k \text{ tels que } \mathcal{D}S_x^k(w_{1:k}) \text{ est de rang maximal.} \tag{B.29}$$

Nous rappelons un critère pour que $\{\phi_k\}_{k\in\mathbb{N}}$ soit une T-chaîne irréductible, avec des hypothèses plus fortes sur F et α.

Théorème 2 ([110, Theorems 7.1.1 and 7.2.6, Propositions 7.1.2 and 7.2.5])**.** Supposons que $\mathsf{X} = \mathbb{R}^n$ et $\mathsf{U} = \mathsf{W} = \mathbb{R}^p$. Si de plus F est infiniment différentiable continuement, et si $\alpha(x, u) = u$ pour $(x, u) \in \mathsf{X} \times \mathsf{U}$, alors nous avons que :

(i) si tout $x \in \mathsf{X}$ satisfait (B.29), alors $\{\phi_k\}_{k\in\mathbb{N}}$ est une T-chaîne ;

(ii) s'il existe un état globalement attracteur x^* satisfaisant (B.29), alors $\{\phi_k\}_{k\in\mathbb{N}}$ une T-chaîne irréductible.

Afinde caractériser l'apériodicité, on dit qu'un état $x^* \in \mathsf{X}$ est régulièrement attracteur quand pour tout $x \in \mathsf{X}$ et tout voisinage U de x^*, il existe $T \in \mathbb{N}$ tel que $A_+^k(x)$ intersecte U pour tout $k \geqslant T$.

Le critère suivant généralise le théorème 2.

Théorème 3 ([29, Corollary 4.1, Theorems 4.2 and 4.4])**.** Supposons que X, U et W sont des ouverts respectifs de $\mathbb{R}^n$, $\mathbb{R}^m$, $\mathbb{R}^p$. Si F est $\mathcal{C}^1$, alors nous avons :

(i) so tout $x \in \mathsf{X}$ satisfait (B.29), alors $\{\phi_k\}_{k\in\mathbb{N}}$ est une T-chaîne;

(ii) s'il existe un état globalement attracteur x^* satisfaisant (B.29), alors $\{\phi_k\}_{k\in\mathbb{N}}$ est une T-chaîne irréductible ;

(iii) s'il existe un état régulièrment attracteur x^* satisfaisant (B.29), alors $\{\phi_k\}_{k\in\mathbb{N}}$ est une T-chaîne irréductible et apériodique.

Dans le chapitre 1, nous généralisons le théorème 3 au cas où X et W sont des variétés, et la fonction F est localement lipschitzienne. Dans le chapitre 2, nous appliquons ces résultats à des chaînes normalisées sous-jacentes à CMA-ES et nous prouvons qu'elles sont des T-chaînes irréductibles et apériodiques.

5 Convergence et garanties théoriques pour les ES

Dans cette section, nous donnons un aperçu des travaux précédents qui fournissent des aspects théoriques de différents ES, en particulier en ce qui concerne la convergence. Les premiers travaux [23] ont prouvé que les ES avec un pas proportionnel à la distance à l'optimum, et plus tard les ES avec un pas adaptatif [14], convergent linéairement vers l'optimum d'une fonction sphère $f = g(|\cdot|_2)$, avec $g\colon \mathbb{R} \to \mathbb{R}$ une fonction monotone. Cette approche est très similaire à celle utilisée dans cette thèse : en définissant un processus normalisé $z_t = (m_t - x^*)/\sigma_t$, nous obtenons une chaîne de Markov géométriquement ergodique. Cela conduit à un comportement linéaire du progrès logarithmique de la moyenne vers l'optimum.

Plus tard, la convergence linéaire du (1+1)-ES avec adaptation du pas a été obtenue, d'abord pour des problèmes sphériques [87] et ellipsoïdaux [88, 89, 90], ainsi que pour des fonctions objectifs bruitées [92]. Plus récemment, une analyse de type dérive a permis d'estimer le taux de convergence et le temps moyen d'atteinte pour des problèmes sphériques [3] et pour des problèmes fortement convexes et lisses [4]. Cette méthodologie diffère de la nôtre car elle nécessite d'obtenir une condition de dérive pour chaque initialisation de l'espace des états (alors que nous pouvons commencer en dehors d'un compact arbitrairement grand). Elle s'inspire des approches utilisées dans l'analyse des algorithmes évolutionnaires pour les problèmes combinatoires [42, 40, 145, 101].

De plus, différents résultats théoriques ont été établis pour les algorithmes stochastiques et en particulier pour la classe des ES provenant du point de vue de l'optimisation géométrique de l'information (IGO) [119, 8]. Il a été montré qu'une version simplifiée de CMA-ES, sans adaptation du pas ni mise à jour de rang un, peut être considérée comme une mise à jour de gradient naturel sur l'espace des probabilités. De plus, une analyse de convergence des équations différentielles ordinaires (EDO) découlant du cadre IGO pour des ES à pas adaptatifs a été réalisée pour les transformations croissantes de fonctions objectifs deux fois continuement différentiables [5]. L'analyse d'une EDO sous-jacente associée aux ESà pas adaptatifs a récemment permis de déduire la convergence linéaire pour les fonctions objectifs sphériques (c'est-à-dire des transformations croissantes de la norme euclidienne) [6].

La première preuve de convergence d'algorithmes incluant des ES avec des mécanismes similaires à ceux de CMA-ES pour l'adaptation de la matrice de covariance a exploité une mise à jour différente du pas [39], garantissant la limite $\sigma_t \to 0$ pour des fonctions objectifs bornées inférieurement, et une condition de décroissance suffisante sur f-valeur de la moyenne avant la mise à jour. Sous des hypothèses supplémentaires, il a été prouvé que la moyenne converge alors vers un point stationnaire. Cependant, ces modifications ne reflètent pas les caractéristiques importantes de CMA-ES, car elles n'autorisent pas l'augmentation de la f-valeur de la moyenne, ce qui limite l'exploration de l'espace d'états, et forcent le pas à diminuer rapidement, affectant les performances de l'algorithme sur des problèmes multimodaux, pour lesquels l'exploration est essentielle.

Plus récemment, l'analyse de convergence des ES avec une adaptation de la matrice de covariance différente, à savoir la stratégie d'évolution avec estimation de la Hessienne (HE-ES), a montré la convergence linéaire de la moyenne vers l'optimum et l'apprentissage de l'inverse de la Hessienne [54] pour des problèmes convexes quadratiques. De plus, il a été démontré que le taux de convergence de HE-ES ne dépendait pas du conditionnement du problème. L'algorithme HE-ES, bien qu'il ait été "conçu pour être conceptuellement proche de CMA-ES", impose que les enfants soient échantillonnés dans des directions symétriques orthogonales, ce qui facilite l'analyse théorique de la mise à jour de la matrice de covariance. De plus, il approche la Hessienne par un schéma de différences finies, en se reposant sur les valeurs de la fonction objectif, contrairement à CMA-ES qui utilise uniquement des

comparaisons de valeurs de fonction. Cette approche repose également sur une analyse de chaînes de Markov, mais suppose que l'algorithme est élitiste, c'est-à-dire qu'il n'autorise pas l'augmentation de la f-valeur de la moyenne.

Des travaux récents ont utilisé une approche par chaînes de Markov pour analyser la convergence globale des ES avec adaptation du pas sur des fonctions invariantes par changement d'échelle [17, 141]. La définition d'une fonction invariante par changement d'échelle permet dans ces travaux de définir un processus normalisé et de prouver qu'il s'agit d'une chaîne de Markov géométriquement ergodique. Cela a été la principale source d'inspiration pour le travail présenté dans cette thèse, car les idées clés des démonstrations suivent le même schéma. Toutefois, il y a plusieurs améliorations de la méthode utilisée à l'époque sur lesquelles repose notre démonstration. Tout d'abord, une précédente preuve de convergence pour les ES adaptatifs en taille de pas [141] reposait sur des conditions d'irréductibilité, ainsi que sur des propriétés d'apériodicité et topologiques, qui nécessitaient d'avoir des mises à jour continuement différentiables sur des ouverts d'espaces euclidiens [29]. Le chapitre 1 généralise ces conditions pour inclure des mises à jour localement lipschitziennes sur des variétés, ce qui permet d'analyser la mise à jour non lisse CSA dans (8), et facilite l'analyse d'une chaîne impliquant une matrice de covariance normalisée, évaluée dans une variété de matrices définies positives normalisées. De plus, nos résultats reposent sur une condition de dérive dépendante de l'état (plutôt qu'une condition de dérive standard) pour obtenir l'ergodicité géométrique d'une chaîne de Markov normalisée, simplifiant l'analyse lorsque nous utilisons la cumulation sur le pas.

5.1 Prouver un comportement linéaire des ES en analysant une chaîne de Markov normalisée

Nous nous concentrons ici sur des approches basées sur l'analyse d'une chaîne de Markov normalisée. C'est l'approche adoptée dans cette thèse pour prouver la convergence linéaire de CMA-ES.

Bien que cela ait été appliqué à plusieurs ES, nous présentons les résultats sur un $(\mu/\mu_w, \lambda)$-ES à pas adaptatifs (sans adaptation de la matrice de covariance) minimisant une fonction objectif $f\colon \mathbb{R}^d \to \mathbb{R}$ introduite dans Section 2.1.1.

La convergence de cet algorithme [141] a été prouvée pour différents changements de taille de pas $\Gamma\colon \mathbb{R}^d \to \mathbb{R}$ satisfaisant diverses hypothèses. Nous présentons ces résultats dans le cas particulier où $\Gamma = \Gamma^2_{\text{CSA}}$, voir (9).

La preuve de convergence repose principalement sur l'hypothèse que la fonction objectif f est *invariante par changement d'échelle* [142], c'est-à-dire lorsqu'il existe un point $x^* \in \mathbb{R}^d$ tel que

$$f(x^* + x) \leqslant f(x^* + y) \Rightarrow f(x^* + \rho x) \leqslant f(x^* + \rho y) \quad \text{for } x, y \in \mathbb{R}^d \text{ and } \rho > 0. \tag{B.30}$$

Dans ce cas, nous définissons le processus suivant

$$z_t = \frac{m_t - x^*}{\sigma_t} \quad \text{for } t \in \mathbb{N}, \tag{B.31}$$

qui est une chaîne de Markov homogène [141, Proposition 2]. Par le théorème 3, c'est une T-chaîne irréductible et apériodique, et par une condition de dérive géométrique comme le théorème 1, c'est une chaîne ergodique.

Théorème 4 ([141, Theorem 6])**.** Quand f est lisse et invariante par changement d'échelle par rapport à x^*, et si $\Gamma = \Gamma^2_{\text{CSA}}$, alors $\{z_t\}_{t\in\mathbb{N}}$ est une chaîne de MArkov géométriquement ergodique.

Cela nous permet de déduire le comportement linéaire de l'algorithme:

$$\begin{aligned}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} &= \frac{1}{T}\sum_{t=0}^{T-1}\log\|m_{t+1} - x^*\| - \log\|m_t - x^*\| \\ &= \frac{1}{T}\sum_{t=0}^{T-1}\log\|z_{t+1}\| - \log\|z_t\| + \log\Gamma_{\text{CSA}}^2\left(\sqrt{\mu_{\text{eff}}}\sum_{i=1}^{\mu} w_i U_{t+1}^{s_{t+1}(i)}\right).\end{aligned}$$

Alors, des propriétes d'intégrabilité [141, Propositions 10 and 11] permettent de trouver une limite presque sûre dans la formule ci-dessus.

Théorème 5 ([141, Theorem 7])**.** Si f is lisse et invariante par changement d'échelle par rapport à x^*, et si $\Gamma = \Gamma_{\text{CSA}}^2$, alors il existe $\text{CR} \in \mathbb{R}$ tel que pour toute initialisation:

$$\lim_{T\to\infty}\frac{1}{T}\log\frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{T\to\infty}\frac{1}{T}\log\frac{\sigma_T}{\sigma_0} = -\text{CR}$$

et

$$\lim_{t\to\infty}\mathbb{E}\left[\log\frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}\right] = \lim_{t\to\infty}\mathbb{E}\left[\log\frac{\sigma_{t+1}}{\sigma_t}\right] = -\text{CR}\ .$$

Le cas $\text{CR} > 0$ correspond à la convergence linéaire.

Une preuve que CR soit positif manque. Nous démontrons dans le chapitre 5 que pour des fonctions objectifs ellipsoïdes, sous des conditions sur les hyperparamètres, CMA-ES avec changement de pas Γ_{CSA}^2 converge linéairement. Les mêmes arguments devraient pouvoir s'appliquer pour des ES avec pas adaptatifs.

6 Méthodologie et aperçu

Nous donnons dans cette section un aperçu de la thèse. Chaque sous-section ci-dessous présente les contributions des chapitres de la thèse.

L'objectif principal de la thèse est de fournir une preuve rigoureuse de la convergence linéaire de la moyenne m_t produite par CMA-ES vers l'optimum x^* pour une certaine classe de fonctions objectifs $f\colon \mathbb{R}^d \to \mathbb{R}$. De plus, nous prouvons que la matrice de covariance $\mathbf{C}_t$ apprend des informations de second ordre.

Tout au long du manuscrit, nous considérons plusieurs hypothèses requises pour différentes parties de la preuve. Ainsi, dans le chapitre 2, nous supposons que la fonction objectif f est invariante par changement d'échelle et une transformation monotonement croissante d'une fonction continue, avec des ensembles de niveaux négligeables pour Lebesgue, tandis que dans le chapitre **??**, nous nous restreignons aux fonctions objectifs ellipsoïdales, c'est-à-dire lorsqu'il existe une fonction strictement croissante $g\colon \mathbb{R} \to \mathbb{R}$ telle que $f(x) = g(x^\top \mathbf{H} x)$ pour une certaine matrice $\mathbf{H} \in \mathcal{S}_{++}^d$. Le nom *ellipsoïdal* est dû à la forme ellipsoïdale des ensembles de niveaux de f.

De plus, afin de préparer une analyse plus générale, dans Chapter 4, nous ne supposons pas que les variables aléatoires U_{t+1}^i dans (B.1) suivent une distribution normale standard, mais plutôt une distribution de probabilité générique sur $\mathcal{B}(\mathbb{R}^d)$ avec des hypothèses minimales. Cependant, dans les chapitres 4 et 5, nos hypothèses incluront uniquement des distributions normales.

6.1 *Chapitre 1: Sur l'irréductibilité et la convergence d'une classe de modèles d'espaces d'états non-lisses sur des variétés [52]*

Tout au long de la thèse, et en particulier dans les chapitres 2, 4 et 5, nous analysons une chaîne de Markov normalisée sous-jacente à CMA-ES afin de déduire finalement la convergence linéaire. Avant cela, nous introduisons dans le chapitre 1 la méthodologie utilisée dans le chapitre 2. En effet, bien que des travaux précédents [109, 29] aient fourni des conditions pour analyser l'irréductibilité des chaînes de Markov sous-jacentes à divers algorithmes ES, ils présentent des limitations qui empêchent leur application à CMA-ES. Nous généralisons ces conditions dans Chapter 1, et nous donnons ici un bref aperçu du principal résultat de ce chapitre.

Considérons une chaîne de Markov $\Theta = \theta_{tt\in\mathbb{N}}$ telle que

$$\theta_{t+1} = F(\theta_t, \alpha(\theta_t, U_{t+1})) \tag{B.32}$$

où $F\colon \mathsf{X} \times \mathsf{W} \to \mathsf{X}$ est une fonction localement lipschitzienne entre des variétés $\mathsf{X} \times \mathsf{W}$ et W, α est une fonction mesurable entre des espaces mesurables $\mathsf{X} \times \mathsf{U}$ et W satsifaisant des hypothèses diverses, et $\{U_{t+1}\}_{t\in\mathbb{N}}$ est un processus i.i.d. indépendant de θ_0.

Nous donnons dans le chapitre 1 des définitions variées associées à (B.32) et certaines sont rappelées dans la section 4.8. La contribution principale du chapitre 1 est résumée dans le théorème suivant.

> **Théorème 6** (Théorème 1.2 dans le chapitre 1)**.** Supposons qu'il exists un état globalement attracteur θ^* telle que la condition (B.29) tienne. Alors, Θ est une T-chaîne irréductible. Si de plus θ^* est régulièrment attracteur, alors Θ est apériodique.

Il reste plusieurs limitations à ce résultat.

Tout d'abord, nous rencontrons dans Chapter 2 le problème selon lequel la chaîne de Markov normalisée sous-jacente à CMA-ES est définie sur une variété non lisse lorsque la normalisation est égale, par exemple, à la plus petite valeur propre. En définissant un homéomorphisme entre deux chaînes de Markov, l'une définie sur une variété lisse et satisfaisant les conditions requises pour appliquer le théorème 6, nous sommes en mesure de prouver nos résultats. Il serait cependant intéressant d'avoir des conditions plus générales qui incluent des variétés non lisses.

Deuxièmement, l'hypothèse selon laquelle la variable aléatoire $\alpha(\theta, U_1)$ est absolument continue par rapport à une mesure σ-finie localement équivalente à la mesure de Lebesgue ne peut pas s'appliquer aux ES avec une stratégie de plus (voir la section 2.1.1 où nous introduisons la différence entre les stratégies plus et virgule). Une extension intéressante serait d'inclure des cas où $\alpha(\theta, U_1)$ est un mélange de distributions avec densité et de distributions de Dirac.

6.2 *Chapitre 2: Irréductibilité de modèles à espaces d'états non-lisses avec application à CMA-ES[53]*

Le chapitre 2 constitue le point de départ de notre preuve de convergence de CMA-ES. Dans ce chapitre, nous définissons un processus normalisé sous-jacent à CMA-ES par

$$\begin{aligned} z_t &= \frac{m_t - x^*}{\sigma_t \sqrt{R(\mathbf{C}_t)}} \\ \mathbf{\Sigma}_t &= \frac{1}{R(\mathbf{C}_t)} \mathbf{C}_t \\ p_t &= p_t^\sigma \\ q_t &= \frac{p_t^c}{\sqrt{R(\mathbf{C}_{t-1})}} \\ r_t &= \frac{R(\mathbf{C}_t)}{R(\mathbf{C}_{t-1})} \end{aligned} \tag{B.33}$$

où $R\colon \mathcal{S}_{++}^d \to \mathbb{R}_{++}$ est une fonction de normalisation de la matrice de covariance. Prouver que ce processus est une chaîne de Markov géométriquement ergodique est l'étape centrale de notre preuve. Nous prouvons d'abord qu'il s'agit d'une chaîne de Markov lorsque la fonction est invariante par changement d'échelle.

Théorème 7 (Théorème 2.2 dans le chapitre 2)**.** Si f est invariante par changement d'échelle par rapport à x^*, et si R est positivement homogène, alors $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ est une chaîne de Markov homogène.

De plus, nous utilisons le théorème 6 et la méthode du chapitre 1 pour trouver que c'est une T-chaîne irréductible et apériodique. Cependant, le théorème 6 ne permet pas d'inclure tous les schémas d'hyperparamètres en une seule preuve. Par exemple, si c_σ est égal à 1, nous ne prouvons pas l'irréductibilité de la chaîne (B.33), mais de $\{(z_t, \mathbf{\Sigma}_t, q_t, r_t)\}_{t\in\mathbb{N}}$. Pour gérer ces chaînes de Markov différentes, on introduit les notions de chaînes *redondantes* et *projectées*. Plus précisément, une chaîne de MArkov $\{(\theta_t, \chi_t)\}_{t\in\mathbb{N}}$ à valeurs dans un espace d'états $\mathsf{X}\times\mathsf{Y}$ est dite redondante, quand le processus $\{\theta_t\}_{t\in\mathbb{N}}$ à valeurs dans X est elle-même une chaîne de Markov. Dans ce cas, la chaîne $\{\theta_t\}_{t\in\mathbb{N}}$ est dite projetée. Nous donnons dans le chapitre 2 une généralisation du théorème 6 pour des chaînes redondantes. Nous considérons la chaine de Markov

$$(\theta_{t+1}, \chi_{t+1}) = \tilde{F}((\theta_t, \chi_t), \tilde{\alpha}((\theta_t, \chi_t), U_{t+1})) \tag{B.34}$$

où $\tilde{F}$ et $\tilde{\alpha}$ obéissent à des hypothèses que F et α dans la section précédente. Nous introduisons la condition de controllabilité en $(\theta, \chi) \in \mathsf{X}$:

$$\exists w_{1:k} \in \overline{\tilde{\mathcal{O}}^k_{(\theta,\chi)}}, \forall (h_\theta, h_\chi) \in \mathsf{T}_{\tilde{S}^k_{(\theta,\chi)}(w_{1:k})}(\mathsf{X}\times\mathsf{Y}), h_\theta \in \operatorname{range} \mathcal{D}(\Pi_{\mathsf{X}} \circ \tilde{S}^k_{(\theta,\chi)})(w_{1:k}) \tag{B.35}$$

où Π_{X} est la projection de $\mathsf{X}\times\mathsf{Y}$ sur X.

Théorème 8 (Théorème 2.3 dans le chapitre 2)**.** Supposons qu'il existe un état régulièrement attracteur (θ^*, χ^*) tel que (B.35) tienne. Alors, $\{\theta_t\}_{t\in\mathbb{N}}$ est une T-chaîne irréductible et apériodique.

Ceci permet de prouver l'irréductibilité de la chaîne normalisée.

Théorème 9 (Théorème 2.1 dans le chapitre 2)**.** Sous des hypothèses supplémentaires sur f, R, Γ, ν_U^d et les hyperparamètres de CMA-ES, le processus

(i) $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ si $c_c \neq 1$ et $c_\sigma \neq 1$;

(ii) $\{z_t, q_t, \mathbf{\Sigma}_t, r_t\}_{t\in\mathbb{N}}$ si $c_c \neq 1$ et $c_\sigma = 1$;

(iii) $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ si $c_c = 1$ et $c_\sigma \neq 1$;

(iv) $\{(z_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ si $c_c = 1$ et $c_\sigma = 1$;

est une T-chaîne irréductible et apériodique.

6.3 Chapitre 3: Estimation asymptotique d'un problème de vecteurs propres symétrique perturbé [51]

L'étape centrale de la preuve de convergence présentée pour CMA-ES est décrite dans Chapter 4. Nous y montrons que le processus normalisé introduit précédemment satisfait une condition de

dérive et est donc géométriquement ergodique. Cependant, cette étape est difficile, car elle nécessite d'obtenir des bornes supérieures précises sur les paramètres estimés de CMA-ES. En particulier, nous devons déterminer une borne précise pour la plus grande valeur propre espérée de la matrice de covariance dans les cas où celle-ci doit diminuer, lorsque le conditionnement de la matrice de covariance est élevé. Pourtant, la normalisation de la matrice de covariance impose également de contrôler l'augmentation de la plus petite valeur propre, et par conséquent, de comprendre comment les axes de la matrice de covariance changent au fil des itérations est déterminant.

C'est le sujet de Chapter 3. Nous y analysons une perturbation d'une matrice définie positive similaire à la mise à jour de rang mu de CMA-ES, et nous fournissons des bornes sur la projection des vecteurs propres du système mis à jour sur ceux de l'initial, lorsque le conditionnement de la matrice est élevé. Formulons le problème comme suit :

$$\mathbf{B} = \mathbf{A} + \sqrt{\mathbf{A}} \sum_{i=1}^{\mu} v_i v_i^\top \sqrt{\mathbf{A}} \tag{B.36}$$

où $\mathbf{A} \in \mathcal{S}_{++}^d$ représente la matrice initiale, et les vecteurs $v_i \in \mathbb{R}^d$, $i = 1, \ldots, \mu$, composent la matrice de rang mu. Dans la suite, on note λ_i la fonction i-eme plus grande (comptée avec multiplicité) valeur propre d'une matrice définie positive, et e_i une fonction vecteur propre associée.

Théorème 10 (Théorème 3.1 dans le chapitre 3)**.** Nous avons

$$|\langle e_i(\mathbf{A}), e_j(\mathbf{B})\rangle| \leqslant C \times \sqrt{\frac{\min\left(\lambda_i(\mathbf{A}), \lambda_j(\mathbf{A})\right)}{\max\left(\lambda_i(\mathbf{A}), \lambda_j(\mathbf{A})\right)}}$$

où $C > 0$ ne dépend que de μ, d et $\|v_i\|$, $i = 1, \ldots, \mu$.

Bien que ce résultat suffise dans notre contexte, nous estimons que la valeur de C pourrait être donnée plus précisément.

6.4 *Chapitre 4: Ergodicité géométrique de chaîne de Markov sous-jacentes à CMA-ES*

Après l'irréductibilité et l'aperiodicité, la prochaine étape dans l'analyse de convergence de la chaîne de Markov normalisée (B.33) est la preuve de l'ergodicité géométrique. À cet effet, nous nous appuyons dans le chapitre 4 sur une condition de dérive dépendante de l'état (voir le théorème 1 ci-dessus). Cependant, prouver qu'elle s'applique à un processus aussi complexe est une tâche très compliquée et nécessite des hypothèses supplémentaires. À ce stade de la preuve, nous ne considérons que des fonctions objectifs ellipsoïdales, c'est-à-dire des fonctions objectifs qui sont des transformations croissantes de fonctions convexes quadratiques. De plus, nous restreignons notre analyse à des fonctions de normalisation spécifiques $R(\cdot)$, plus précisément aux transformations de la plus petite valeur propre.

Théorème 11 (Théorème 4.4 dans le chapitre 4)**.** Considérons la chaîne de Markov normalisée $\{\theta_t\} = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ via (B.33) avec une fonction objectif ellipsoïdale f qui est une transformation croissante de $x \mapsto (x - x^*)^\top \mathbf{H}(x - x^*)$ pour une matrice $\mathbf{H} \in \mathcal{S}_{++}^d$, et une normalisation $R \mapsto \mathbf{C} \in \mathcal{S}_{++}^d \mapsto \lambda_{\min}(\mathbf{H}^{1/2}\mathbf{C}\mathbf{H}^{1/2})$. Soit V le potentiel défini par

$$V(z, p, q, \mathbf{\Sigma}, r) = \|\mathbf{H}^{1/2} z\|^2 + \beta \lambda_{\max}(\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}) + \gamma_p \|p\| + \gamma_q \|\mathbf{H}^{1/2} q\|^2 + \gamma_r r \tag{B.37}$$

où les constantes $\beta, \gamma_p, \gamma_q, \gamma_r > 0$ sont bien choisies. Alors, si les hyperparamètres et le

changement de pas $\Gamma(\cdot)$ de CMA-ES sont bien choisis, nous avons

$$\mathbb{E}\left[V(\theta_{n(\theta)}) \mid \theta_0 = \theta\right] \leqslant \rho \times V(\theta) \tag{B.38}$$

pour tout θ en-dehors d'un compact K, et où $\rho \in (0,1)$ et $n(\theta) = 1$ or 2.

Comme dans Chapter 2, nous souhaitons inclure différents hyperparamètres et nous étendons donc le critère de dérive dépendante de l'état pour l'ergodicité aux chaînes de Markov projetées. En effet, nous prouvons que si une chaîne de Markov redondante satisfait une condition de dérive dépendante de l'état en dehors d'un ensemble small et qu'elle possède une chaîne projetée irréductible et aperiodique, alors cette dernière est ergodique. Nous utilisons cela pour prouver directement que plusieurs variantes d'algorithmes (en particulier lorsque c_c ou c_σ est égal à 11) sont ergodiques en n'analysant que la chaîne (B.33).

Cependant, nous n'avons pas obtenu l'ergodicité géométrique de la chaîne normalisée lorsque deux chemins d'évolution sont utilisées pour mettre à jour le pas et la matrice de covariance.

Théorème 12 (Théorème 4.3 dans le chapitre 4)**.** Supposons que f est ellipsoïdale et une transformation croissante de $x \mapsto (x - x^*)^\top \mathbf{H}(x - x^*)$ pour $\mathbf{H} \in \mathcal{S}^d_{++}$. Considérons une normalisation $R \mapsto \mathbf{C} \in \mathcal{S}^d_{++} \mapsto \lambda_{\min}(\mathbf{H}^{1/2}\mathbf{C}\mathbf{H}^{1/2})$ qui définit une chaîne de Markov $\{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ via (B.33). Lorsque les hyperparamètres et le changement de pas $\Gamma(\cdot)$ de CMA-ES sont bien choisis, nous avons :

(i) quand $c_c \neq 1$ et $c_\sigma = 1$, la chaîne projetée $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\in\mathbb{N}}$ est géométriquement ergodique ;

(ii) quand $c_c = 1$ et $c_\sigma \neq 1$, la chaîne projetée $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ est géométriquement ergodique ;

(iii) quand $c_c = c_\sigma = 1$, la chaîne projetée $\{(z_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ est géométriquement ergodique.

6.5 *Chapitre 5: Convergence linéaire de CMA-ES sur des problèmes ellipsoïdaux et apprentissage d'informations de second ordre*

Le chapitre 5 conclut notre preuve de convergence de CMA-ES. Puisque dans le chapitre 4, nous avons obtenu l'ergodicité géométrique de la chaîne de Markov (B.33), nous obtenons le comportement linéaire de la moyenne vers l'optimum de la fonction objective lorsque celle-ci est ellipsoïdale. De plus, lorsque nous n'avons pas de cumulation sur le pas (c'est-à-dire $c_\sigma = 1$), nous pouvons utiliser l'affine-invariance de CMA-ES et constater que le taux de convergence est le même lors de la minimisation de toute fonction objective ellipsoïdale et que la matrice de covariance approche l'inverse de la Hessienne d'une fonction convexe quadratique. En outre, nous prouvons que pour une mise à jour spécifique du pas, le taux de convergence est positif, ce qui prouve que CMA-ES converge linéairement.

Théorème 13 (Théorème **??** dans le chapitre 5)**.** Considérons une fonction objective ellipsoïdale f qui est une transformation croissante de $x \mapsto (x - x^*)^\top \mathbf{H}(x - x^*)$ pour $\mathbf{H} \in \mathcal{S}^d_{++}$ et $x^* \in \mathbb{R}^d$. Si les hyperparamètres de CMA-ES sont bien choisis et si le changement de pas est donné par (B.9), alors la moyenne m_t converge vers x^* à une vitesse géométrique, c'est-à-dire

il exitse $C > 0$ et $\rho \in (0, 1)$ tels que pour tout $t \in \mathbb{N}$:

$$\|m_t - x^*\| \leqslant C\rho^t \ . \tag{B.39}$$

De plus, la matrice de covariance approche $\mathbf{H}^{-1}$ dans le sens qu'il existe $\alpha > 0$ tel que :

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\mathbf{C}_t}{\det(\mathbf{C}_t)^{1/d}}\right] = \alpha\mathbf{H}^{-1} \ . \tag{B.40}$$

6.6 *Annexe A: Évaluation de l'impact de modifications variées de CMA-ES qui facilitent son analyse théorique [49]*

Nous fournissons dans l'Annexe A plusieurs résultats expérimentaux qui évaluent l'impact des modifications de CMA-ES qui ont été ou pourraient avoir été utilisées dans la preuve de convergence présentée dans cette thèse.

On y trouve l'impact de la cumulation sur le pas et sur la matrice de covariance sur les performances de l'algorithme. Cela est particulièrement intéressant à la lumière de nos résultats, car nous avons pu prouver l'apprentissage de l'inverse du Hessien, ainsi que le fait que le taux de convergence est le même pour toutes les fonctions ellipsoïdales, uniquement lorsque nous n'avons pas de cumulation sur le pas.

Nous avons également analysé les performances de la mise à jour alternative du pas pour laquelle nous avons pu prouver que le taux de convergence est positif. Nous n'avons pas trouvé de différences suffisamment significatives qui suggéreraient que cette mise à jour ne soit pas pertinente en pratique. Une autre mise à jour du pas incluse dans ce travail est basée sur le chemin p_t^c, et modifie l'algorithme CMA-ES pour le rendre affine-invariant. Cependant, nous n'avons pas analysé cette variante dans nos travaux théoriques.

Bibliography

[1] Esther Tolulope Aboyeji, Oladayo S. Ajani, and Rammohan Mallipeddi. Covariance matrix adaptation evolution strategy based on ensemble of mutations for parking navigation and maneuver of autonomous vehicles. *Expert Systems with Applications*, 249:123565, September 2024. (pp. 2, 58, and 230)

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631, New York, NY, USA, July 2019. Association for Computing Machinery. (pp. 2, 25, 58, and 230)

[3] Youhei Akimoto, Anne Auger, and Tobias Glasmachers. Drift theory in continuous search spaces: Expected hitting time of the (1 + 1)-ES with 1/5 success rule. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '18, pages 801–808, New York, NY, USA, July 2018. Association for Computing Machinery. (pp. 16, 215, and 244)

[4] Youhei Akimoto, Anne Auger, Tobias Glasmachers, and Daiki Morinaga. Global Linear Convergence of Evolution Strategies on More than Smooth Strongly Convex Functions. *SIAM Journal on Optimization*, 32(2):1402–1429, June 2022. (pp. 2, 16, 63, 215, 230, and 244)

[5] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. Convergence of the Continuous Time Trajectories of Isotropic Evolution Strategies on Monotonic C2-composite Functions. In Carlos A. Coello Coello, Vincenzo Cutello, Kalyanmoy Deb, Stephanie Forrest, Giuseppe Nicosia, and Mario Pavone, editors, *Parallel Problem Solving from Nature - PPSN XII*, pages 42–51, Berlin, Heidelberg, 2012. Springer. (pp. 16 and 244)

[6] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. An ODE method to prove the geometric convergence of adaptive stochastic algorithms. *Stochastic Processes and their Applications*, 145:269–307, March 2022. (pp. 2, 16, 230, and 244)

[7] Youhei Akimoto, Yuichi Nagata, Isao Ono, and Shigenobu Kobayashi. Bidirectional Relation between CMA Evolution Strategies and Natural Evolution Strategies. In *Parallel Problem Solving from Nature, PPSN XI*, Lecture Notes in Computer Science, pages 154–163, Berlin, Heidelberg, 2010. Springer. (p. 28)

[8] Youhei Akimoto, Yuichi Nagata, Isao Ono, and Shigenobu Kobayashi. Theoretical Foundation for CMA-ES from Information Geometry Perspective. *Algorithmica*, 64(4):698–716, December 2012. (pp. 16, 28, and 244)

[9] Herbert Amann and Joachim Escher. *Analysis III*. Birkhäuser, Basel, 2009. (pp. 29 and 48)

[10] H. Z. An and S. G. Chen. A note on the ergodicity of non-linear autoregressive model. *Statistics & Probability Letters*, 34(4):365–372, June 1997. (p. 24)

[11] Charly Andral, Randal Douc, Hugo Marival, and Christian P. Robert. The importance Markov chain. *Stochastic Processes and their Applications*, 171:104316, May 2024. (p. 117)

[12] Dirk V. Arnold. Weighted multirecombination evolution strategies. *Theoretical Computer Science*, 361(1):18–37, August 2006. (p. 142)

[13] D.V. Arnold and H.-G. Beyer. Performance analysis of evolutionary optimization with cumulative step length adaptation. *IEEE Transactions on Automatic Control*, 49(4):617–622, April 2004. (p. 62)

[14] Anne Auger. Convergence results for the (1,λ)-SA-ES using the theory of ϕ-irreducible Markov chains. *Theoretical Computer Science*, 334(1):35–69, April 2005. (pp. 15, 59, 116, and 244)

[15] Anne Auger. Analysis of Comparison-based Stochastic Continuous Black-Box Optimization Algorithms. Thèse d'habilitation à diriger des recherches, Université Paris-Sud, May 2016. (pp. 9, 27, 63, 65, 138, 178, 193, 219, and 238)

[16] Anne Auger and Nikolaus Hansen. Linear Convergence on Positively Homogeneous Functions of a Comparison Based Step-Size Adaptive Randomized Search: The (1+1) ES with Generalized One-fifth Success Rule, October 2013. (pp. 59 and 116)

[17] Anne Auger and Nikolaus Hansen. Linear Convergence of Comparison-based Step-size Adaptive Randomized Search via Stability of Markov Chains. *SIAM Journal on Optimization*, 26(3):1589–1624, January 2016. (pp. 2, 16, 25, 26, 33, 59, 63, 65, 109, 116, 117, 123, 213, 230, and 245)

[18] Jacques Bénasséni and Alain Mom. Inequalities for the eigenvectors associated to extremal eigenvalues in rank one perturbations of symmetric matrices. *Linear Algebra and its Applications*, 570:123–137, June 2019. (p. 108)

[19] Hans-Georg Beyer and Bernhard Sendhoff. Simplify Your Covariance Matrix Adaptation Evolution Strategy. *IEEE Transactions on Evolutionary Computation*, 21(5):746–759, October 2017. (p. 178)

[20] Rajendra Bhatia. *Perturbation Bounds for Matrix Eigenvalues*. Number 53 in Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2007. (p. 108)

[21] Rabi Bhattacharya and Chanho Lee. On geometric ergodicity of nonlinear autoregressive models. *Statistics & Probability Letters*, 22(4):311–315, March 1995. (p. 24)

[22] Jonathan Bieler, Rosamaria Cannavo, Kyle Gustafson, Cedric Gobet, David Gatfield, and Felix Naef. Robust synchronization of coupled circadian and cell cycle oscillators in single mammalian cells. *Molecular Systems Biology*, 10(7):739, July 2014. (pp. 2, 25, 58, and 230)

[23] Alexis Bienvenüe and Olivier François. Global convergence for evolution strategies in spherical problems: Some simple proofs and difficulties. *Theoretical Computer Science*, 306(1):269–289, September 2003. (pp. 2, 15, 59, 116, 230, and 244)

[24] Ihor O. Bohachevsky, Mark E. Johnson, and Myron L. Stein. Generalized Simulated Annealing for Function Optimization. *Technometrics*, 28(3):209–217, August 1986. (pp. 7 and 235)

[25] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, May 2011. (pp. 6, 10, 58, 235, and 238)

[26] C. G. Broyden. The convergence of single-rank quasi-Newton methods. *Mathematics of Computation*, 24(110):365–382, 1970. (pp. 8 and 236)

[27] James R. Bunch, Christopher P. Nielsen, and Danny C. Sorensen. Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik*, 31(1):31–48, March 1978. (p. 108)

[28] V. Černý. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, January 1985. (pp. 7 and 235)

[29] Alexandre Chotard and Anne Auger. Verifiable conditions for the irreducibility and aperiodicity of Markov chains by analyzing underlying deterministic models. *Bernoulli*, 25(1):112–147, February 2019. (pp. 14, 15, 16, 18, 23, 24, 25, 27, 29, 30, 31, 33, 34, 35, 37, 38, 39, 40, 44, 59, 213, 217, 218, 242, 243, 245, and 247)

[30] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, January 1990. (pp. 31, 39, 41, 49, 50, 51, 52, 65, 75, and 105)

[31] Sebastian Colutto, Florian Fruhauf, Matthias Fuchs, and Otmar Scherzer. The CMA-ES on Riemannian Manifolds to Reconstruct Shapes in 3-D Voxel Images. *IEEE Transactions on Evolutionary Computation*, 14(2):227–245, April 2010. (pp. 2, 58, and 230)

[32] S. B. Connor and G. Fort. State-dependent Foster–Lyapunov criteria for subgeometric convergence of Markov chains. *Stochastic Processes and their Applications*, 119(12):4176–4193, December 2009. (p. 117)

[33] J. T. Cox and Richard Durrett. Limit theorems for the spread of epidemics and forest fires. *Stochastic Processes and their Applications*, 30(2):171–191, December 1988. (p. 116)

[34] Hans Crauel. *Random Probability Measures on Polish Spaces*. July 2002. (p. 132)

[35] H. A. David and H. N. Nagaraja. Order Statistics. In *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Ltd, 2004. (p. 182)

[36] Chandler Davis and W. M. Kahan. The Rotation of Eigenvectors by a Perturbation III. 7(1), 1970. (p. 109)

[37] Anton Dekkers and Emile Aarts. Global optimization and simulated annealing. *Mathematical Programming*, 50(1):367–393, March 1991. (pp. 7 and 235)

[38] Jiu Ding and Aihui Zhou. Eigenvalues of rank-one updated matrices with some applications. *Applied Mathematics Letters*, 20(12):1223–1226, December 2007. (p. 108)

[39] Y. Diouane, S. Gratton, and L. N. Vicente. Globally convergent evolution strategies. *Mathematical Programming*, 152(1):467–490, August 2015. (pp. 2, 16, 230, and 244)

[40] Benjamin Doerr, Daniel Johannsen, and Carola Winzen. Drift analysis and linear functions revisited. In *IEEE Congress on Evolutionary Computation*, pages 1–8, July 2010. (pp. 16 and 244)

[41] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Cham, 2018. (pp. 117 and 126)

[42] Stefan Droste, Thomas Jansen, and Ingo Wegener. On the analysis of the (1+1) evolutionary algorithm. *Theoretical Computer Science*, 276(1):51–81, April 2002. (pp. 16 and 244)

[43] Lawrence Craig Evans and Ronald F Gariepy. *Measure Theory and Fine Properties of Functions, Revised Edition*. Chapman and Hall/CRC, New York, April 2015. (pp. 41, 49, and 74)

[44] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, January 1970. (pp. 8 and 236)

[45] Nicolas Fournier and Sylvie Méléard. A microscopic probabilistic description of a locally regulated population and macroscopic approximations. *The Annals of Applied Probability*, 14(4):1880–1919, November 2004. (p. 116)

[46] Garuda Fujii, Youhei Akimoto, and Masayuki Takahashi. Exploring optimal topology of thermal cloaks by CMA-ES. *Applied Physics Letters*, 112(6):061108, February 2018. (p. 58)

[47] Marco A. Gallegos-Herrada, David Ledvinka, and Jeffrey S. Rosenthal. Equivalences of Geometric Ergodicity of Markov Chains. *Journal of Theoretical Probability*, May 2023. (p. 58)

[48] Armand Gissler. Github repository. https://github.com/agissler/Benchmarking_proof_variants_CMA-ES/. (p. 220)

[49] Armand Gissler. Evaluation of the impact of various modifications to CMA-ES that facilitate its theoretical analysis. In *GECCO 2023 - Genetic and Evolutionary Computation Conference*, July 2023. (pp. x, xiii, 1, 22, 62, 73, 121, 178, 179, 217, 230, and 251)

[50] Armand Gissler, Anne Auger, and Nikolaus Hansen. Learning Rate Adaptation by Line Search in Evolution Strategies with Recombination. In *GECCO 2022*, July 2022. (p. 179)

[51] Armand Gissler, Anne Auger, and Nikolaus Hansen. Asymptotic estimations of a perturbed symmetric eigenproblem. *Applied Mathematics Letters*, 150:108951, April 2024. (pp. ix, xiii, 1, 20, 107, 139, 140, 229, and 248)

[52] Armand Gissler, Alain Durmus, and Anne Auger. On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds, February 2024. (pp. ix, xiii, 1, 18, 23, 59, 71, 72, 82, 91, 104, 121, 229, and 247)

[53] Armand Gissler, Shan-Conrad Wolf, Anne Auger, and Nikolaus Hansen. Irreducibility of nonsmooth state-space models with an application to CMA-ES, September 2024. (pp. ix, xiii, 1, 19, 57, 117, 122, 123, 124, 126, 127, 131, 135, 229, and 247)

[54] Tobias Glasmachers and Oswin Krause. Convergence Analysis of the Hessian Estimation Evolution Strategy. *Evolutionary Computation*, 30(1):27–50, March 2022. (pp. 2, 16, 230, and 244)

[55] Peter W. Glynn, Sanatan Rai, and John E. Glynn. Recurrence classification for a family of non-linear storage models. *Probability and Mathematical Statistics*, 37(2):337–353, 2017. (p. 24)

[56] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970. (pp. 8 and 236)

[57] Gene H. Golub. Some Modified Matrix Eigenvalue Problems. *SIAM Review*, 15(2):318–334, April 1973. (p. 108)

[58] Geoffrey Grimmett. *Percolation*, volume 321 of *Grundlehren Der Mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 1999. (p. 133)

[59] Victor Guillemin and Alan Pollack. *Differential Topology*. American Mathematical Soc., 2010. (pp. 27, 40, and 48)

[60] David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. *Advances in neural information processing systems*, 2018. (pp. 2, 25, and 230)

[61] Heikki Haario, Eero Saksman, and Johanna Tamminen. An Adaptive Metropolis Algorithm. *Bernoulli*, 7(2):223–242, 2001. (p. 108)

[62] Piotr Hajłasz. Change of variables formula under minimal assumptions. In *Colloquium Mathematicae*, volume 64, pages 93–101, 1993. (p. 41)

[63] Nikolaus Hansen. The CMA Evolution Strategy: A Comparing Review. 2006. (pp. 3, 116, 121, 218, and 232)

[64] Nikolaus Hansen. CMA-ES with Two-Point Step-Size Adaptation, May 2008. (p. 214)

[65] Nikolaus Hansen. The CMA Evolution Strategy: A Tutorial, April 2016. (p. 120)

[66] Nikolaus Hansen, Youhei Akimoto, and Petr Baudis. CMA-ES/pycma on Github. Zenodo, DOI:10.5281/zenodo.2559634, February 2019. (pp. 8, 220, and 237)

[67] Nikolaus Hansen, Youhei Akimoto, and Petr Baudis. CMA-ES/pycma on Github. Zenodo, January 2023. (p. 116)

[68] Nikolaus Hansen, Dirk V Arnold, and Anne Auger. Evolution Strategies. 2015. (pp. 4, 5, 58, 116, 121, 232, and 233)

[69] Nikolaus Hansen and Anne Auger. CMA-ES: Evolution strategies and covariance matrix adaptation. In *Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '11, pages 991–1010, New York, NY, USA, July 2011. Association for Computing Machinery. (p. 58)

[70] Nikolaus Hansen and Anne Auger. Principled Design of Continuous Stochastic Search: From Theory to Practice. In Yossi Borenstein and Alberto Moraglio, editors, *Theory and Principled Methods for the Design of Metaheuristics*, Natural Computing Series, pages 145–180. Springer, Berlin, Heidelberg, 2014. (pp. 9, 58, 63, 65, 91, 138, 178, 193, and 238)

[71] Nikolaus Hansen, Anne Auger, Raymond Ros, Olaf Mersmann, Tea Tušar, and Dimo Brockhoff. COCO: A Platform for Comparing Continuous Optimizers in a Black-Box Setting. *Optimization Methods and Software*, 36(1):114–144, January 2021. (pp. 218 and 220)

[72] Nikolaus Hansen and Stefan Kern. Evaluating the CMA Evolution Strategy on Multimodal Test Functions. In *Parallel Problem Solving from Nature - PPSN VIII*, Lecture Notes in Computer Science, pages 282–291, Berlin, Heidelberg, 2004. Springer. (pp. 2, 58, 142, and 230)

[73] Nikolaus Hansen, Sibylle D. Müller, and Petros Koumoutsakos. Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18, March 2003. (pp. 2, 8, 25, 26, 58, 108, 109, 116, 218, 230, and 237)

[74] Nikolaus Hansen and Andreas Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of IEEE International Conference on Evolutionary Computation*, pages 312–317, May 1996. (pp. 2, 58, and 230)

[75] Nikolaus Hansen and Andreas Ostermeier. Convergence properties of evolution strategies with the derandomized covariance matrix adaptation : The (μ/μI,λ)-CMA-ES. *Proceedings of the Fifth European Congress on Intelligent Techniques and Soft Computing (EUFIT'97), Aachen, Germany*, 650, 1997. (pp. 2 and 230)

[76] Nikolaus Hansen and Andreas Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, June 2001. (pp. 2, 6, 8, 25, 26, 58, 109, 116, 122, 218, 230, 234, and 237)

[77] Nikolaus Hansen and Raymond Ros. Benchmarking a weighted negative covariance matrix update on the BBOB-2010 noiseless testbed. In *Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '10, pages 1673–1680, New York, NY, USA, July 2010. Association for Computing Machinery. (pp. 3, 63, and 231)

[78] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970. (pp. 6, 58, and 235)

[79] Horn R. and Johnson C. *Matrix Analysis*. Cambridge University Press, 2013. (pp. 65, 108, 109, 110, and 134)

[80] Biwei Huang, Chaochao Lu, Liu Leqi, Jose Miguel Hernandez-Lobato, Clark Glymour, Bernhard Schölkopf, and Kun Zhang. Action-Sufficient State Representation Learning for Control with Structural Constraints. In *Proceedings of the 39th International Conference on Machine Learning*, pages 9260–9279. PMLR, June 2022. (pp. 2 and 230)

[81] Jianyi Huang, Ioannis Kontoyiannis, and Sean P. Meyn. The ODE Method and Spectral Theory of Markov Operators. In *Stochastic Theory and Control*, Lecture Notes in Control and Information Sciences, pages 205–221, Berlin, Heidelberg, 2002. Springer. (p. 24)

[82] Kanji Ichihara and Hiroshi Kunita. A classification of the second order degenerate elliptic operators and its probabilistic characterization. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 30(3):235–254, 1974. (p. 24)

[83] Christian Igel, Thorsten Suttorp, and Nikolaus Hansen. A computational efficient covariance matrix update and a (1+1)-CMA for evolution strategies. In *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*, GECCO '06, pages 453–460, New York, NY, USA, July 2006. Association for Computing Machinery. (pp. 5, 215, and 234)

[84] I. C. F. Ipsen and B. Nadler. Refined Perturbation Bounds for Eigenvalues of Hermitian and Non-Hermitian Matrices. *SIAM Journal on Matrix Analysis and Applications*, 31(1):40–53, January 2009. (p. 108)

[85] Ilse C. F. Ipsen. Relative perturbation results for matrix eigenvalues and singular values. *Acta Numerica*, 7:151–201, January 1998. (pp. 108, 109, and 141)

[86] Ilse C. F. Ipsen. An overview of relative sinΘ theorems for invariant subspaces of complex matrices. *Journal of Computational and Applied Mathematics*, 123(1):131–153, November 2000. (p. 108)

[87] Jens Jägersküpper. Analysis of a Simple Evolutionary Algorithm for Minimization in Euclidean Spaces. In Jos C. M. Baeten, Jan Karel Lenstra, Joachim Parrow, and Gerhard J. Woeginger, editors, *Automata, Languages and Programming*, pages 1068–1079, Berlin, Heidelberg, 2003. Springer. (pp. 15 and 244)

[88] Jens Jägersküpper. Rigorous Runtime Analysis of the (1+1) ES: 1/5-Rule and Ellipsoidal Fitness Landscapes. In Alden H. Wright, Michael D. Vose, Kenneth A. De Jong, and Lothar M. Schmitt, editors, *Foundations of Genetic Algorithms*, pages 260–281, Berlin, Heidelberg, 2005. Springer. (pp. 15 and 244)

[89] Jens Jägersküpper. How the (1+1) ES using isotropic mutations minimizes positive definite quadratic forms. *Theoretical Computer Science*, 361(1):38–56, August 2006. (pp. 15 and 244)

[90] Jens Jägersküpper. Algorithmic analysis of a basic evolutionary algorithm for continuous optimization. *Theoretical Computer Science*, 379(3):329–347, June 2007. (pp. 15 and 244)

[91] G.A. Jastrebski and D.V. Arnold. Improving Evolution Strategies through Active Covariance Matrix Adaptation. In *2006 IEEE International Conference on Evolutionary Computation*, pages 2814–2821, July 2006. (pp. 2, 63, 214, and 230)

[92] Mohamed Jebalia, Anne Auger, and Nikolaus Hansen. Log-Linear Convergence and Divergence of the Scale-Invariant (1+1)-ES in Noisy Environments. *Algorithmica*, 59(3):425–460, March 2011. (pp. 15, 218, and 244)

[93] Jérôme Henri Kämpf and Darren Robinson. A hybrid CMA-ES and HDE optimisation algorithm with application to solar energy potential. *Applied Soft Computing*, 9(2):738–745, March 2009. (pp. 2 and 230)

[94] Michael Karow and Daniel Kressner. On a Perturbation Bound for Invariant Subspaces of Matrices. 2014. (p. 108)

[95] K. Kikuchi, M. Otani, K. Yamaguchi, and E. Simo-Serra. Modeling Visual Containment for Web Page Layout Optimization. *Computer Graphics Forum*, 40(7):33–44, 2021. (pp. 2 and 230)

[96] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, May 1983. (pp. 7 and 235)

[97] G. Kjellstrom and L. Taxen. Stochastic optimization in system design. *IEEE Transactions on Circuits and Systems*, 28(7):702–715, July 1981. (p. 108)

[98] Wolfgang Kliemann. Recurrence and Invariant Measures for Degenerate Diffusions. *The Annals of Probability*, 15(2):690–707, 1987. (p. 24)

[99] Jeffrey C. Lagarias, James A. Reeds, Margaret H. Wright, and Paul E. Wright. Convergence Properties of the Nelder–Mead Simplex Method in Low Dimensions. *SIAM Journal on Optimization*, 9(1):112–147, January 1998. (pp. 7 and 235)

[100] John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2012. (pp. 25, 46, and 78)

[101] Johannes Lengler. Drift Analysis. In Benjamin Doerr and Frank Neumann, editors, *Theory of Evolutionary Computation: Recent Developments in Discrete Optimization*, pages 89–131. Springer International Publishing, Cham, 2020. (pp. 16 and 244)

[102] Lennart Ljung. *System Identification*. Citeseer, 1995. (p. 108)

[103] Ilya Loshchilov and Frank Hutter. CMA-ES for Hyperparameter Optimization of Deep Neural Networks. February 2016. (pp. 2 and 230)

[104] Atsuo Maki, Naoki Sakamoto, Youhei Akimoto, Hiroyuki Nishikawa, and Naoya Umeda. Application of optimal control theory based on the evolution strategy (CMA-ES) to automatic berthing. *Journal of Marine Science and Technology*, 25(1):221–233, March 2020. (pp. 2, 58, and 230)

[105] Roy Mathias. Spectral Perturbation Bounds for Positive Definite Matrices. *SIAM Journal on Matrix Analysis and Applications*, 18(4):959–980, October 1997. (p. 108)

[106] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953. (pp. 6, 58, and 235)

[107] S. Meyn and P. Caines. A new approach to stochastic adaptive control. *IEEE Transactions on Automatic Control*, 32(3):220–226, March 1987. (p. 24)

[108] S. P. Meyn and P. E. Caines. Stochastic controllability and stochastic Lyapunov functions with applications to adaptive and nonlinear systems. In *Stochastic Differential Systems*, Lecture Notes in Control and Information Sciences, pages 235–257, Berlin, Heidelberg, 1989. Springer. (p. 24)

[109] S. P. Meyn and P. E. Caines. Asymptotic Behavior of Stochastic Systems Possessing Markovian Realizations. *SIAM Journal on Control and Optimization*, 29(3):535–561, May 1991. (pp. 14, 18, 24, 25, 33, 40, 41, 59, 242, and 247)

[110] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2009. (pp. 10, 11, 12, 13, 14, 15, 24, 27, 30, 32, 36, 37, 44, 45, 65, 68, 69, 91, 92, 109, 126, 129, 130, 195, 201, 208, 210, 218, 238, 239, 240, 241, 242, and 243)

[111] S.P. Meyn and L.J. Brown. Model reference adaptive control of time varying and stochastic systems. *IEEE Transactions on Automatic Control*, 38(12):1738–1753, December 1993. (p. 24)

[112] Abdelkader Mokkadem. *Critères de mélange pour des processus stationnaires. Estimation sous des hypothèses de mélange. Entropie des processus linéaires*. PhD thesis, Université Paris-Sud, September 1987. (p. 24)

[113] Daiki Morinaga, Kazuto Fukuchi, Jun Sakuma, and Youhei Akimoto. Convergence Rate of the (1+1)-ES on Locally Strongly Convex and Lipschitz Smooth Functions. *IEEE Transactions on Evolutionary Computation*, 28(2):501–515, April 2024. (p. 63)

[114] Somabha Mukherjee. A Proof of the Herschel-Maxwell Theorem Using the Strong Law of Large Numbers. *Pi Mu Epsilon Journal*, 14(6):383–387, 2017. (p. 119)

[115] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, January 1965. (pp. 7 and 235)

[116] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2 edition, 2006. (pp. 7, 8, and 236)

[117] Masahiro Nomura and Masashi Shibata. Cmaes : A Simple yet Practical Python Library for CMA-ES, February 2024. (p. 116)

[118] Masahiro Nomura, Shuhei Watanabe, Youhei Akimoto, Yoshihiko Ozaki, and Masaki Onishi. Warm Starting CMA-ES for Hyperparameter Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9188–9196, May 2021. (pp. 2 and 230)

[119] Yann Ollivier, Ludovic Arnold, Anne Auger, and Nikolaus Hansen. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. 2011. (pp. 16 and 244)

[120] Andreas Ostermeier, Andreas Gawelczyk, and Nikolaus Hansen. Step-size adaptation based on non-local use of selection information. In Yuval Davidor, Hans-Paul Schwefel, and Reinhard Männer, editors, *Parallel Problem Solving from Nature — PPSN III*, pages 189–198, Berlin, Heidelberg, 1994. Springer. (p. 58)

[121] Cécile Patte, Pierre-Yves Brillet, Catalin Fetita, Jean-François Bernaudin, Thomas Gille, Hilario Nunes, Dominique Chapelle, and Martin Genet. Estimation of Regional Pulmonary Compliance in Idiopathic Pulmonary Fibrosis Based on Personalized Lung Poromechanical Modeling. *Journal of Biomechanical Engineering*, 144(091008), March 2022. (pp. 2, 25, and 230)

[122] A. J. Piergiovanni, Anelia Angelova, and Michael S. Ryoo. Evolving Losses for Unsupervised Video Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 133–142, 2020. (pp. 2 and 230)

[123] Lennart Oswald Purucker and Joeran Beel. CMA-ES for Post Hoc Ensembling in AutoML: A Great Success and Salvageable Failure. In *AutoML Conference 2023*, August 2023. (pp. 2 and 230)

[124] Ingo Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog, Stuttgart, Germany, 1973. (pp. 2, 5, 230, and 234)

[125] S. Surender Reddy, B. K. Panigrahi, Rupam Kundu, Rohan Mukherjee, and Shantanab Debchoudhury. Energy and spinning reserve scheduling for a wind-thermal power system using CMA-ES with mean learning technique. *International Journal of Electrical Power & Energy Systems*, 53:113–122, December 2013. (pp. 2 and 230)

[126] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1(none):20–71, January 2004. (p. 58)

[127] Gareth O. Roberts and Jeffrey S. Rosenthal. Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms. *Journal of Applied Probability*, 44(2):458–475, June 2007. (pp. 10 and 238)

[128] Gareth O. Roberts and Richard L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, March 1996. (pp. 10 and 238)

[129] Maria Rodriguez-Fernandez, Pedro Mendes, and Julio R. Banga. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems*, 83(2):248–265, February 2006. (pp. 2, 25, 58, and 230)

[130] Klaus D. Schmidt. On inequalities for moments and the covariance of monotone functions. *Insurance: Mathematics and Economics*, 55:91–95, March 2014. (p. 133)

[131] M. Schumer and K. Steiglitz. Adaptive step size random search. *IEEE Transactions on Automatic Control*, 13(3):270–276, June 1968. (pp. 2 and 230)

[132] Hans-Paul Schwefel. Evolutionsstrategien für die numerische Optimierung. In Hans-Paul Schwefel, editor, *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie: Mit einer vergleichenden Einführung in die Hill-Climbing- und Zufallsstrategie*, pages 123–176. Birkhäuser, Basel, 1977. (pp. 5 and 234)

[133] Denis Serre. *Matrices: Theory and Applications*, volume 216 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2010. (p. 65)

[134] D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970. (pp. 8 and 236)

[135] G. W. Stewart. *Matrix Algorithms*, volume 2. 2001. (pp. 109, 111, and 141)

[136] Daniel W. Stroock and S. R. S. Varadhan. On the Support of Diffusion Processes with Applications to the Strong Maximum Principle. In *Contributions to Probability Theory*, pages 333–360. University of California Press, December 1972. (p. 24)

[137] Takumi Tanabe, Kazuto Fukuchi, Jun Sakuma, and Youhei Akimoto. Level generation for angry birds with sequential VAE and latent variable evolution. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '21, pages 1052–1060, New York, NY, USA, June 2021. Association for Computing Machinery. (pp. 2, 58, and 230)

[138] R. C. Thompson. The behavior of eigenvalues and singular values under perturbations of restricted rank. *Linear Algebra and its Applications*, 13(1):69–78, January 1976. (p. 108)

[139] Hermann Thorisson. *Coupling, Stationarity, and Regeneration*. Springer New York, 2000. (pp. 133 and 181)

[140] Stephen Tian, Yancheng Cai, Hong-Xing Yu, Sergey Zakharov, Katherine Liu, Adrien Gaidon, Yunzhu Li, and Jiajun Wu. Multi-Object Manipulation via Object-Centric Neural Scattering Functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9021–9031, 2023. (pp. 2 and 230)

[141] Cheikh Toure, Anne Auger, and Nikolaus Hansen. Global linear convergence of evolution strategies with recombination on scaling-invariant functions. *Journal of Global Optimization*, 86(1):163–203, May 2023. (pp. 2, 16, 17, 23, 25, 26, 29, 33, 35, 59, 62, 63, 91, 109, 116, 117, 121, 123, 197, 213, 218, 230, 245, and 246)

[142] Cheikh Toure, Armand Gissler, Anne Auger, and Nikolaus Hansen. Scaling-invariant Functions versus Positively Homogeneous Functions. *Journal of Optimization Theory and Applications*, 191(1):363–383, October 2021. (pp. 17, 26, 59, 63, 122, and 245)

[143] Ninoslav Truhar and Ivan Slapničar. Relative perturbation bound for invariant subspaces of graded indefinite Hermitian matrices. *Linear Algebra and its Applications*, 301(1):171–185, November 1999. (p. 108)

[144] Vanessa Volz, Jacob Schrum, Jialin Liu, Simon M. Lucas, Adam Smith, and Sebastian Risi. Evolving mario levels in the latent space of a deep convolutional generative adversarial network. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '18, pages 221–228, New York, NY, USA, July 2018. Association for Computing Machinery. (pp. 2 and 230)

[145] Carsten Witt. Tight Bounds on the Optimization Time of a Randomized Search Heuristic on Linear Functions. *Combinatorics, Probability and Computing*, 22(2):294–318, March 2013. (pp. 16 and 244)

[146] J.-F. Yao and J.-G. Attali. On stability of nonlinear AR processes with Markov switching. *Advances in Applied Probability*, 32(2):394–407, June 2000. (p. 24)

[147] Serdar Yuksel and Sean P. Meyn. Random-Time, State-Dependent Stochastic Drift for Markov Chains and Application to Stochastic Stabilization Over Erasure Channels. *IEEE Transactions on Automatic Control*, 58(1):47–59, January 2013. (p. 117)

Index of Notations

$\overline{\mathsf{A}}$	Closed hull of a subset A
$\operatorname{Arg\,max}$	Set of global maxima
$\arg\max$	Unique global maximum
$\operatorname{Arg\,min}$	Set of global minima
$\arg\min$	Unique global minimum
$B(x, r)$	Open ball of center x and radius r
$\mathcal{B}(\mathsf{X})$	Borelian σ-field of a topological space X
$\mathbb{C}$	Set of complex numbers
$\mathcal{C}^k$	Set of k times continuously differentiable functions
$\mathcal{C}^\infty$	Set of infinitely many times continuously differentiable
conv	Convex hull
$\overline{\operatorname{conv}}$	Closed convex hull
$\mathcal{D}f$	Differential application of a function f
∂f	Clarke differential of a function f
∇f	Gradient of a function f
$\nabla^2 f$	Hessian of a function f
$\det$	Determinant
$\mathbb{E}$	Expectation
$\exp$	Exponential function
$\inf$	Infimium
int	Interior
L^p	Set of functions that are p times integrable with respect to the Lebesgue measure
$\log$	Logarithm function
$\mathbb{N}$	Set of natural numbers
$\mathbb{N}^*$	Set of positive integers
$\mathcal{N}(m, C)$	Normal distribution with mean m and covariance matrix C
$\mathbb{P}$	Probability
$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+$	Set of nonnegative numbers
$\mathbb{R}_{++}$	Set of positive numbers
rank	Rank of a matrix or a linear operator
$\mathcal{S}^d$	Set of symmetric matrices of size $d \times d$
$\mathcal{S}_+^d$	Set of semidefinite positive matrices of size $d \times d$
$\mathcal{S}_{++}^d$	Set of definite matrices matrices of size $d \times d$

Sp	Spectrum (set of eigenvalues)
$\sup$	Supremum
supp	Support of a function
$T_x\mathsf{X}$	Tangent space of X at x
$\mathbb{Z}$	Set of integers

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre: Convergence linéaire de stratégies d'évolution avec adaptation de matrice de covariance

Mots clés: CMA-ES, Convergence linéaire, Chaînes de Markov, Irréductibilité, Ergodicité, Matrice de covariance

Résumé: En optimisation, l'algorithme CMA-ES est à l'état de l'art des stratégies d'évolution. Bien qu'il y ait eu de nombreuses applications depuis plus de 20 ans, une démonstration de sa convergence est restée une question ouverte. Nous y répondons dans cette thèse. Ainsi, nous donnons des garanties théoriques de convergence linéaire pour des fonctions objectif ellipsoïdales. De plus, nous prouvons que l'information de second ordre de fonctions convexe-quadratiques est apprise puisque la matrice de covariance approche l'inverse de la matrice Hessienne. Afin d'établir nos résultats, nous normalisons les états de CMA-ES pour définir une chaîne de Markov quand la fonction objectif est invariante par changement d'échelle. La stabilité de cette chaîne de Markov permet ensuite de prouver la convergence de CMA-ES. Nous décomposons cette preuve en plusieurs parties. Les deux premiers chapitres présentent une méthodologie pour prouver l'irréductibilité et des propriétés topologiques de chaînes de Markov ainsi que son application à des chaînes normalisées sous-jacentes à CMA-ES. Après les deux chapitres suivants, nous obtenons l'ergodicité géométrique de ces chaînes de Markov pour des problèmes ellipsoïdaux. Le dernier chapitre conclue la démonstration.

Title: Linear convergence of evolution strategies with covariance matrix adaptation

Keywords: CMA-ES, Linear convergence, Markov chains, Irreducibility, Ergodicity, Covariance matrix

Abstract: In optimization, CMA-ES is the state-of-the-art algorithm among the evolution strategies. Although it has found many applications for more than 20 years, a proof of its convergence remained an open question. Solving this problem is the goal of this thesis. More precisely, we provide theoretical guarantees of linear convergence of CMA-ES when minimizing an ellipsoidal objective function. Moreover, we prove that second-order information of a convex-quadratic function is learnt since the covariance matrix approximates the inverse Hessian matrix. To establish our results, we normalize the states of the algorithm and define a Markov chain when the objective function is scaling-invariant. The stability of this Markov chain proves then the convergence of CMA-ES. We decompose this proof in several steps. In the first two chapters, we give a methodology to prove the irreducibility and topological properties of Markov chains that we apply to the normalized Markov chain underlying CMA-ES. After the next two chapters, we obtain that this Markov chain is geometrically ergodic for ellipsoidal problems. The last chapter concludes our proof.

Institut Polytechnique de Paris
91120 Palaiseau, France