Thèse présentée pour obtenir le grade de

DOCTEUR DE L'ÉCOLE POLYTECHNIQUE

Spécialité : Mathématiques Appliquées

par

Zheng QU

Nonlinear Perron-Frobenius theory and max-plus numerical methods for Hamilton-Jacobi equations.

Attenuation of the curse of dimensionality

soutenue le 21 Octobre 2013 devant le jury composé de :

Philippe Bougerol	Université Pierre et Marie Curie	examinateur
Nicole El Karoui	Université Pierre et Marie Curie	examinateur
Maurizio Falcone	Roma 1 La Sapienza	rapporteur
Stéphane Gaubert	INRIA Saclay & CMAP, Polytechnique	directeur
Éric Goubault	CEA	président du jury
William McEneaney	University of Carlifornia, San diego	examinateur
Pierre Rouchon	École des Mines de Paris	rapporteur
Shanjian Tang	Fudan University	codirecteur

Remerciements

Je tiens en premier lieu à exprimer ma profonde reconnaissance envers Stéphane Gaubert dans son rôle de directeur de thèse. J'ai eu la chance de bénéficier pendant trois ans de son fort soutien mathématique, de ses encouragements me motivant à surmonter les difficultés, de sa confiance me laissant une grande liberté et de sa disponibilité malgré un agenda fortement rempli. Ses larges connaissances dans de nombreux domaines des mathématiques, sa passion pour la recherche, sa patience et sa générosité pour ses élèves restent pour moi un exemple à suivre.

Je tiens à remercier également Shanjian Tang, codirecteur de cette thèse. Il a spontanément accepté de m'encadrer et m'a accueillie dans de meilleures conditions pendant chacune de mes visites à Shanghai. Il m'a constamment encouragée, m'a donné de précieuses suggestions et remarques, et il a inspiré une partie importante de ma thèse. Je le remercie aussi de venir de loin pour être membre de mon jury.

Le lecteur remarquera sans peine que la thèse part d'une méthode développée par William McEneaney, qui est à l'origine des premières méthodes numériques max-plus en contrôle optimal. Je lui suis extrêmement redevable et le remercie également de participer à la soutenance.

Je remercie Pierre Rouchon et Maurizio Falcone qui m'ont fait l'honneur d'être rapporteurs de thèse pour le temps qu'ils ont consacré à la lecture et pour leurs commentaires permettant d'améliorer mon manuscrit. Je remercie vivement Philippe Bougerol, Nicole El Karoui, et Éric Goubault, d'avoir accepté de participer à mon jury en tant qu'examinateurs.

Je voudrais aussi remercier les membres de l'équipe "max plus": Marianne Akian, Xavier Allamigeon, Cormac Walsh, Olivier Fercoq, Pascal Benchimol, Andreas Marchesini, pour les discussions et les échanges qui m'ont aidée à progresser dans mon travail. Mes remerciements s'adressent également aux autres doctorants du CMAP, surtout à Anna Kazeykina, Georgios Michailidis, Laetitia Giraldi, Gabriel Delgado, Matteo Santacesaria, Gwenael Mercier, Zixian Jiang, Camille Coron, Laurent Pfeiffer, Xavier Dupuis, avec qui je partage une période importante de ma vie et qui m'ont tous donné un coup de main à un moment ou à un autre.

Un grand merci à Sylvain Ferrand, chargé d'informatique du CMAP, pour ses aides et sa grande patience. Merci à Wallis Filippi, l'ancienne assistante de l'équipe "max plus". Je pense également à toutes nos chères assistantes du CMAP, Nasséra Naar, Alexandra Noiret, Nathalie Hurel et Sandra Schnakenbourg.

Abstract

Dynamic programming is one of the main approaches to solve optimal control problems. It reduces the latter problems to Hamilton-Jacobi partial differential equations (PDE). Several techniques have been proposed in the literature to solve these PDE. We mention, for example, finite difference schemes, the so-called discrete dynamic programming method or semi-Lagrangian method, or the antidiffusive schemes. All these methods are grid-based, i.e., they require a discretization of the state space, and thus suffer from the so-called curse of dimensionality. The present thesis focuses on max-plus numerical solutions and convergence analysis for medium to high dimensional deterministic optimal control problems. We develop here max-plus based numerical algorithms for which we establish theoretical complexity estimates. The proof of these estimates is based on results of nonlinear Perron-Frobenius theory. In particular, we study the contraction properties of monotone or non-expansive nonlinear operators, with respect to several classical metrics on cones (Thompson's metric, Hilbert's projective metric), and obtain nonlinear or non-commutative generalizations of the "ergodicity coefficients" arising in the theory of Markov chains. These results have applications in consensus theory and also to the generalized Riccati equations arising in stochastic optimal control.

Résumé

Une approche fondamentale pour la résolution de problèmes de contrôle optimal est basée sur le principe de programmation dynamique. Ce principe conduit aux équations d'Hamilton-Jacobi, qui peuvent être résolues numériquement par des méthodes classiques comme la méthode des différences finies, les méthodes semi-lagrangiennes, ou les schémas antidiffusifs. À cause de la discrétisation de l'espace d'état, la dimension des problèmes de contrôle pouvant être abordés par ces méthodes classiques est souvent limitée à 3 ou 4. Ce phénomène est appellé "malédiction de la dimension". Cette thèse porte sur les méthodes numériques max-plus en contôle optimal deterministe et ses analyses de convergence. Nous étudions et developpons des méthodes numériques destinées à attenuer la malédiction de la dimension, pour lesquelles nous obtenons des estimations théoriques de complexité. Les preuves reposent sur des résultats de théorie de Perron-Frobenius non linéaire. En particulier, nous étudions les propriétés de contraction des opérateurs monotones et non expansifs, pour différentes métriques de Finsler sur un cône (métrique de Thompson, métrique projective d'Hilbert). Nous donnons par ailleurs une généralisation du "coefficient d'ergodicité de Dobrushin" à des opérateurs de Markov sur un cône général. Nous appliquons ces résultats aux systèmes de consensus ainsi qu'aux équations de Riccati généralisées apparaissant en contrôle stochastique.

Contents

CHAPTER 1

Introduction

1.1 Context and motivations

Optimal control deals with the problem of determining the inputs (control) for a given system in order to maximize a functional of the state trajectory. A typical finite horizon optimal control problem can be described as follows:

$$v(x,T) := \sup_{\mathbf{u}} \int_0^T \ell(\mathbf{x}(s),\mathbf{u}(s))ds + \phi(\mathbf{x}(T)) \ ;$$

$$\dot{\mathbf{x}}(s) = f(\mathbf{x}(s),\mathbf{u}(s)), \quad \mathbf{x}(0) = x, \quad \mathbf{x}(s) \in X \subset \mathbb{R}^d, \mathbf{u}(s) \in U \ . \tag{1.1}$$

Here, $T > 0$ denotes the horizon, X is the state space, U is the control space, $x \in X$ is an initial state, $\mathbf{x}(\cdot) : [0,T] \to \mathbb{R}^d$ is the state trajectory , $\mathbf{u}(\cdot) : [0,T] \to U$ is the control trajectory and $(\mathbf{x}(\cdot),\mathbf{u}(\cdot))$ satisfy the system (1.1). The functions $f : X \times U \to \mathbb{R}^d$, $\ell : X \times U \to \mathbb{R}$ and $\phi : X \to \mathbb{R}$ are called the dynamics, *Lagrangian* and the terminal reward, respectively. The *value* v gives the optimum of the objective as a function of the initial state x and of the horizon T.

The optimal control problems are most often solved numerically. One classical approach is to apply Pontryagin's Maximum Principle [PBGM62]. The latter provides a necessary optimality condition involving a two point boundary problem for an ordinary differential equation (ODE) in the state and costate variables, which can then be solved by various numerical schemes like the shooting method [Mau76, ABM11]. Since the complexity of ODE integration schemes grows polynomially (sometimes even linearly) with the dimension d, this approach can be applied to problems

of large dimension. However, unless the dynamics and the Lagrangian satisfy restrictive structural assumptions leading to a convex optimization problem, Pontryagin's maximum principle (which is only a necessary condition) may not provide the global optimum. An alternative class of methods, also widely used, called "direct methods" [Kra85], consist in discretizing the optimal control problem, leading to nonlinear optimization problem, which can be solved by various numerical algorithms [Bet01, GMSW86, BMG12], some of which run in polynomial time (but are still not guaranteed to provide the global optimum). Note also that with the Pontryagin and direct approaches, the optimal control is not given in feedback form. In particular, if the initial state x changes, we have to solve the problem again.

An alternative class of methods relies on Bellman's Dynamic Programming Principle [Bel52], which leads to the Hamilton-Jacobi Partial Differential Equation (HJ PDE) [FR75]:

$$\begin{cases} \frac{\partial v}{\partial t} - H(x, \frac{\partial v}{\partial x}) = 0, \qquad \forall (x,t) \in X \times (0,T] \ , \\ v(x,0) = \phi(x), \quad \forall x \in X \ . \end{cases} \tag{1.2}$$

where

$$H(x,p) = \sup_{u \in U} p' f(x,u) + \ell(x,u), \ \ x \in X, p \in \mathbb{R}^d$$

denotes the Hamiltonian of the optimal control problem. Under rather general assumptions, the value function is known to be a viscosity solution of the HJ PDE [CL83, LS85]. A first interest of this method is that it leads to the global optimum. A further interest of dynamic programming lies in its generality, as it can be extended to stochastic optimal control problem [Lio89], and to zero-sum game problem [ES84].

In the dynamic programming approach, the HJ PDE must be solved numerically. The need of accurate schemes has motivated the development of several methods. We mention, for example, the finite difference schemes [CL84], the discrete dynamic programming method by Capuzzo Dolcetta [CD83] or the semi-Lagrangian method developed by Falcone, Ferretti and Carlini [Fal87, FF94, CFF04], the high order ENO schemes introduced by Osher, Sethian and Shu [OS88, OS91], the discontinuous Galerkin method by Hu and Shu [HS99], the ordered upwind methods for convex static Hamilton-Jacobi equations by Sethian and Vladimirsky [SV03] which is an extension of the fast marching method for the Eikonal equations [Set99], and the antidiffusive schemes for advection of Bokanowski and Zidani [BZ07]. However, these methods require a discretization of the state space, and so, they are subject to the *curse of dimensionality* (the term was coined by Richard Bellman in [Bel57]). Indeed, a full grid in $\mathbb{R}^d$ with M nodes in every dimension comprises a total of M^d nodes, and so, the execution time of the scheme is exponential in the dimension of the state space d of the controlled dynamical system (1.1).

The question of the attenuation of the curse of dimensionality has received much attention by the numerical optimal control community. We mention the domain decomposition algorithm [CFLS94, FLS94] and the patchy domain decomposition technique [NK07, CCFP12]. In the discrete dynamic programming community, specially in the study of Markov decision processes, various techniques have also been proposed to reduce the curse of dimensionality, including the approximate policy iteration [Ber11], the classification-based policy iteration [LGM10] and the point based value iteration [CLZ97].

Recently numerical methods of new type for solving HJ PDE, which are not grid-based, have been developed after the work of Fleming and McEneaney [FM00]. These methods are referred to as *max-plus basis methods* since they all rely on max-plus or tropical algebra. Their common idea is to approximate the value function by a supremum of finitely many *basis functions* and to propagate the supremum forward in time by exploiting the max-plus linearity of the Lax-Oleinik semigroup,

which is the evolution semigroup of the HJ PDE. The max-plus linearity properties of Lax-Oleinik semigroups were considered previously by several authors, mostly for theoretical purposes [Mas87, KM97, AQV98].

Various max-plus methods have been developed after the initial one [FM00], concerning different optimal control problems. In particular, McEneaney developed in [McE04] a method adapted to eigenvector/ergodic problems, and then in [McE07] a method adapted to switched linear quadratic problems. Akian, Gaubert and Lakhoua [AGL08] developed a max-plus analogue of the classical finite element method, with a control of the error in terms of (non euclidean) projections. The analysis of the max-plus finite element methods also showed connection with Falcone's semi-Lagrangian schemes, as one particular choice of the max-plus finite element yields the simplest method of the latter class, see Section 3.6 of [LAK07] for details. More recent works on max-plus methods include the ones of McEneaney, Deshpande and Gaubert [MDG08], of Sridharan *et al.* [SGJM10] on a quantum control problem, and of Dower and McEneaney [DM11].

The method developed by McEneaney after [McE07], referred to as *curse of dimensionality free method*, is specially appealing since the complexity growth of the algorithm is polynomial (actually only cubic) with respect to the state space dimension d. In its initial form, the method applies to an infinite horizon switched optimal control problem with Hamiltonian written as the supremum of finitely many quadratic forms:

$$H(x,p) = \max_{m\in\mathscr{M}} (A^m x)'p + \frac{1}{2}x'D^m x + \frac{1}{2}p'\Sigma^m p, \quad \forall x,p\in\mathbb{R}^d \quad .$$

Although the complexity is polynomial in d for a fixed required precision, the number of basis functions which are generated is a power of the number of switches $|\mathscr{M}|$, and this power grows quickly as the required precision increases. This is referred to as a *curse of complexity*. The latter can be reduced by applying a pruning procedure, which selects a subset of basis functions contributing most to the approximation. With an SDP-based pruning technique developed in [MDG08], the curse of complexity can be reduced efficiently, allowing one to deal with problems of dimension up to 15 [SGJM10], inaccessible by classical grid based methods.

The analysis of this new class of methods leads to several questions which will be addressed in this thesis. First, the issue of the reduction of the curse of dimensionality by max-plus basis methods can be phrased as an approximation problem: given a value function satisfying certain convexity and regularity properties, what is the minimum number of max-plus basis functions needed to approximate it with a prescribed accuracy? Next, we shall look for tighter error estimates applying to McEneaney's curse of dimensionality free method. Indeed, the known estimates are too conservative, i.e., the efficiency that the method shows in practice is much higher than the one guaranteed by the error bound given by McEneaney and Kluberg [MK10]. Also, the pruning procedure, which is a decisive ingredient in the efficient implementation of the curse of dimensionality free method, should be understood from the theoretical point of view, in order to allow further improvements. Finally, we shall look for new methods of max-plus type, possibly more efficient on some instances, and leading also to an attenuation of the curse of dimensionality.

1.2 Nonlinear Perron-Frobenius theory

Contraction properties play a key role in the error analysis of many approximate dynamic programming algorithms which have been developed so far. Indeed, in the case of the max-plus finite element method [AGL08], the nonexpansiveness or contraction in the sup norm of the Lax-Oleinik

semigroup is used to bound the total error. In the works on approximate value iteration (in which the value function is approximated by a classical linear combination of basis functions), strict contraction properties in the sup-norm, or sometimes in other L^p norms, are also essential to establish the existence of the fixed point and a fast convergence. Then, most known results on approximate value iteration concern discounted problems and require the discount rate to be bounded away from 1, see for example [TVR97, NB03, BY12]. There is, however, no discount factor in the optimal control problem to which McEneaney's curse of dimensionality free method applies, and so, to get tighter error estimates, one should look for contraction properties of a different nature.

Actually, the curse of dimensionality free method relies on the solution of indefinite Riccati differential equations. We shall see that tight error estimates can be derived if the indefinite Riccati flows are local strict contractions in *Thompson's part metric*, which is a metric classically considered in Perron-Frobenius theory. The latter metric can be be thought of as a sup-norm "with log-glasses". It can be defined on any closed convex pointed cone in a Banach space, and it is a Finsler metric. For positive definite matrices A, B, it is simply given by $d_T(A,B) = \log\max(\lambda_{\max}(A^{-1}B), \lambda_{\max}(B^{-1}A))$. It has the property of being invariant by the action of the linear group on positive definite matrices.

A series of results in the theory of Riccati equations concern the contraction properties of the standard Riccati flow with respect to various classical (invariant) Finsler metrics on the cone of positive semidefinite matrices. These include the invariant Riemannian metric, considered by Bougerol [Bou93], and Thompson's part metric, considered by Liverani and Wojtkowski [LW94] and Lawson and Lim [LL07]. General invariant Finsler metrics were considered by Lee and Lim [LL08]. These results, which exploit the symplectic properties of the standard Riccati flow, only apply to the class of Riccati equations arising from deterministic control problem in which the quadratic cost function is positive semidefinite. However, in the study of complexity of McEneaney's curse of dimensionality free method, it is an essential feature that the quadratic cost is indefinite. Moreover, there are other important classes of generalized Riccati equations, arising for instance from stochastic control problems in which the volatility is controlled (with a bilinear term in the control and in the noise), for which the symplectic structure is lost, and it is natural to ask to what extent the known contraction properties carry over to more general Riccati equations.

These motivations led us to study several general questions in nonlinear Perron-Frobenius theory, concerning the contraction properties of linear and nonlinear, sometimes order-preserving, flows with respect to various natural metrics, including Thompson's part metric and Hilbert's projective metric. Further motivations for the present work arise from the generalization of the classical Birkhoff's theorem [Bir57] to nonlinear maps or flows. Birkhoff's theorem characterizes the contraction rate with respect to Hilbert's projective metric of bounded monotone linear operators preserving a cone. It is a fundamental result in the theory of monotone or nonexpansive operators. In particular, a version of the Perron-Frobenius theorem can be deduced from Birkhoff's contraction property. It is therefore interesting to find a general characterization of the contraction rate of nonlinear operators with respect to Hilbert's projective metric. Also, contraction estimates of linear or nonlinear maps on cones appear to be useful in several fields, including classical consensus theory [Mor05, OT09] and quantum information [NC00, SSR10, RKW11], in which "noncommutative consensus" problems arise.

1.3 Contributions

In Chapter 2 we give an explicit (computable) formula for the exponential contraction rate in Thompson's part metric of any order-preserving flow on the interior of a (possibly infinite dimensional) closed convex pointed cone.

As a first application, we show that the contraction results of Liverani and Wojtkowski [LW94] and of Lawson and Lim [LL07] concerning the standard Riccati equation, as well as new contraction results in the indefinite case, can be recovered, or obtained, by an application of our explicit formula. This provides an alternative to the earlier approaches, which rely on the theory of symplectic semigroups. In particular, we establish a necessary condition for indefinite Riccati flows to be local strict contraction, which we recall is the original motivation of this chapter.

As a second application, we show that the flow of the generalized Riccati equation arising in stochastic linear quadratic control is a local contraction on the cone of positive definite matrices and characterize its Lipschitz constant by a matrix inequality. We also show that the same flow is no longer a contraction in other invariant Finsler metrics on this cone, including the standard invariant Riemannian metric.

We make a detailed comparison with Nussbaum's approach [Nus94], which relies on the Finsler structure of Thompson's metric and is widely applicable in its spirit but leads to different technical assumptions, including geodesic convexity, whereas our proof relies on a flow invariance argument. We construct an example in $\mathbb{R}^2$, for which Nussbaum's approach does not imply the nonexpansiveness but our explicit formula leads to establish the strict contraction of the flow and global exponential convergence of the solutions to a fixed point.

In Chapter 3, we consider contraction properties with respect to *Hilbert's seminorm* (which is also known as Hopf oscillation, or as the *diameter* – Tsitsiklis' Lyapunov function in consensus theory). The Hilbert seminorm is the infinitesimal norm associated to Hilbert's projective metric. In $\mathbb{R}^n$ equipped with its usual partial order, it is nothing but the difference between the maximum and minimum of a vector. We consider here an abstract (closed convex pointed) cone in a Banach space, equipped with the order induced by this cone. We give a general characterization of the contraction ratio with respect to Hilbert's seminorm of a bounded linear map, in terms of the extreme points of a certain abstract "simplex" (elements of the dual cone of unit mass). Some ingredients to establish our results include the observation that Hilbert's seminorm is a quotient norm of Thompson's norm (the infinitesimal norm associated to Thompson's part metric) and duality considerations.

The present results generalize classical results concerning the contraction rate of Markov operators. Indeed, when applying our characterization to stochastic matrices (linear operators leaving invariant the standard positive cone of $\mathbb{R}^n$, and preserving the unit vector), we recover the formula of Dobrushin's ergodicity coefficient [Dob56]. This coefficient determines both the contraction rate of a consensus system with respect to the diameter semimetric [MDA05], and the contraction rate of a stochastic matrix acting on the set of probability vectors, equipped with the total variation distance [LPW09]. When applying our result to the space of Hermitian matrices, equipped with the Loewner order, we therefore obtain a noncommutative version of Dobrushin's ergodicity coefficient, which gives the contraction ratio of a Kraus map (representing a quantum channel or a "noncommutative Markov chain") with respect to the diameter of the spectrum. We shall see that it coincides with the contraction ratio of the dual operator with respect to the total variation distance.

Whereas contraction properties are easy to check for stochastic matrices, the verification of their noncommutative analogues require efforts. Using the noncommutative Dobrushin's ergodicity coefficient, we show that a number of decision problems concerning the contraction rate of Kraus maps reduce to finding a rank one matrix in linear spaces satisfying certain conditions. We then show that an irreducible Kraus map is primitive if and only if the associated noncommutative consensus system is globally convergent. We show that this can be checked in polynomial time if the map is irreducible. However, we prove that unlike in the case of standard nonnegative matrices, deciding whether a Kraus map is strictly positive (meaning that it sends the cone to its interior) is NP-hard. We also show that deciding whether the noncommutative Dobrushin's ergodicity coefficientis strictly less

than 1 is equivalent to a finding a clique of cardinality two in a quantum graph.

In Chapter 4, we apply the formula of the contraction ratio in Hilbert's seminorm of linear maps, obtained in Chapter 3, to finite dimensional nonlinear flows. We first deduce a characterization formula for the contraction rate in Hilbert's seminorm of nonlinear flows. Our characterization leads to an explicit computable formula in the case of $\mathbb{R}^n$ equipped with its standard positive cone. In particular, we obtain explicit contraction rate bound for a class of nonlinear consensus protocols [SM03]. The circumstances in which this bound leads to global convergence result are not as general as Moreau's graph connectivity condition [Mor05]. However, our method gives an explicit exponential contraction rate for this class of nonlinear consensus protocols. Using Nussbaum's Finsler approach [Nus94], we also derive from the formula obtained in Chapter 3 a characterization of the contraction rate of a nonlinear flow in Hilbert's projective metric. We apply the general formula to a nonlinear matrix differential equation and obtain an explicit contraction rate bound in Hilbert's projective metric.

In Chapter 5, we first review the general principle of max-plus basis methods. Then, we establish a negative result, showing that some form of curse dimensionality is unavoidable for these methods, but also for more classical approximate dynamic programming methods like stochastic dual dynamic programming [Sha11], in which a convex value function is approximated by a supremum of affine functions. Indeed, we show that asymptotically, the minimal approximation error in the L_1 or L_∞ norm, for a smooth convex function, using at most n affine minorants, is equivalent to $1/n^{2/d}$, as the number of basis functions n goes to infinity. We derive the latter result as an analogue of Gruber's best asymptotic error estimates of approximating a convex body using circumscribed polytopes [Gru93a, Gru93b]. We also give explicit asymptotic constants, respectively for the L_1 or L_∞ norm. Both constants rely on the determinant of the Hessian matrix of the convex function to approximate. We deduce that an attenuation of the curse of dimensionality occurs (fewer basis functions are needed) when the convex function to be approximated is "flat" in some direction, i.e., when its Hessian matrix has some eigenvalues close to zero.

In Chapter 6, we focus on the algorithmic aspects of McEneaney's curse of dimensionality free method introduced in [MDG08] and propose several refinements of the algorithm. We show that the optimal pruning problem, which is a critical step in the implementation of the method, can be formulated as a continuous version of the facility location or k-center combinatorial optimization problems, in which the connection costs arise from a Bregman distance. Hence, we propose several heuristics (combining facility location heuristics and Shor SDP relaxation scheme). Experimental results show that by combining the primal version of the method with improved pruning algorithms, a higher accuracy is reached for a similar running time, by comparison with the results reported in [MDG08].

In Chapter 7, we provide an improved error analysis of McEneaney's curse of dimensionality free method, restricted to the case when the Hamiltonian is the pointwise maximum of pure quadratic forms (without affine terms). We use the contraction result for the indefinite Riccati flow in Thompson's metric, established in Chapter 2, to show that under different technical assumptions, still covering an important class of problems, the error is only of order $O(e^{-\alpha N\tau})+O(\tau)$ for some $\alpha>0$, where τ is the time discretization step and N is the number of iterations. This improves the approximation error bound $O(1/(N\tau))+O(\sqrt{\tau})$ obtained in previous works of McEneaney and Kluberg. Besides, our approach allows to incorporate the pruning error in the analysis and we show that if the pruning error is $O(\tau^2)$, than the same approximation error order holds. This allows us to tune the precision of the pruning procedure, which in practice is a critical element of the method.

In Chapter 8, we develop a new max-plus basis method, called *max-plus randomized algorithm*, for the class of infinite horizon switched optimal control problems with easily computable Hamiltonians. We give a first convergence proof of the method and present some experimental results. We apply

the method to several instances with dimension varying from 4 to 15, and with the number of switches varying from 6 to 50. Experimental results show that the max-plus randomized algorithm can reach the same precision order obtained by the SDP-based method (introduced in [MDG08] and refined in Chapter 6) with a speedup around 10 up to 100 and that the maximal precision order which can be reached by the new algorithm is much better than what can be done by the SDP based algorithm. Besides, with the new randomized algorithm we are now able to deal with instances of more number of switches for which the previous SDP-based curse of dimensionality method can not reduce the initial backsubstitution error in a reasonable running time. This will allow us, in the future work, to consider more general infinite horizon optimal control problems with semiconvex Hamiltonians, because the latter one can be approximated by the supremum of a large number of linear quadratic functions.

1.4 Organization

The manuscript is divided into two parts. Part I contains all the results on nonlinear Perron-Frobenius theory. Part II contains all the results on max-plus basis methods.

Part I
- In Chapter 2, we establish an explicit formula for the contraction rate in Thompson's metric of arbitrary order-preserving flow on cones;
- In Chapter 3, we characterize the contraction ratio in Hilbert's seminorm of bounded linear maps and study the applications to noncommutative consensus.
- In Chapter 4, we give a characterization of the contraction rate of nonlinear flows in Hilbert's seminorm and in Hilbert's projective metric.

Part 2
- In Chapter 5, we review the general principle of max-plus basis methods and show that the curse of dimensionality is unavoidable for the class of max-plus basis methods in which the value function is smooth, convex and approximated by affine basis functions.
- In Chapter 6, we focus on the algorithmic aspects of McEneaney's curse-of-dimensionality free method introduced in [MDG08] and propose several refinements of the algorithm.
- In Chapter 7, we provide an improved error analysis of McEneaney's curse of dimensionality free method.
- In Chapter 8, we propose a new max-plus basis randomized algorithm for the class of infinite horizon switched optimal control problems.

Appendix A contains some well-known differential calculus formula used in several chapters in the manuscript.

Chapter 2 is based on the preprint [GQ12a], accepted pending minor revision for J. Differential Equations. Chapter 3 is an extended version of an ECC conference article [GQ13]. Chapter 4 is part of the preprint [GQ12b]. Chapter 5 and Chapter 6 are an extended version (with complete proofs) of a CDC conference article [GMQ11]. Chapter 7 is based on the preprint [Qu13a], under revision for SICON. An abridged version of this chapter is included in the ECC conference proceedings [Qu13b].

Part I

Nonlinear Perron-Frobenius theory

CHAPTER 2

The contraction rate in Thompson's part metric of order-preserving flows on a cone

We give a formula for the Lipschitz constant in Thompson's part metric of any order-preserving flow on the interior of a (possibly infinite dimensional) closed convex pointed cone. This shows that in the special case of order-preserving flows, a general characterization of the contraction rate in Thompson metric, given by Nussbaum, leads to an explicit formula. As an application, we show that the flow of the generalized Riccati equation arising in stochastic linear quadratic control is a local contraction on the cone of positive definite matrices and characterize its Lipschitz constant by a matrix inequality. We also show that the same flow is no longer a contraction in other invariant Finsler metrics on this cone, including the standard invariant Riemannian metric. This is motivated by a series of contraction properties concerning the standard Riccati equation, established by Bougerol, Liverani, Wojtkowski, Lawson, Lee and Lim: we show that some of these properties do, and that some other do not, carry over to the generalized Riccati equation.

This chapter is based on the preprint [GQ12a], accepted pending minor revisions in Journal of Differential Equations.

2.1 Introduction

The standard discrete or differential Riccati equation arising in linear-quadratic control or optimal filtering problems has remarkable properties. In particular, Bougerol [Bou93] proved that the standard discrete Riccati operator is non-expansive in the invariant Riemannian metric on the set of positive definite matrices, and that it is a strict contraction under controllability/observability conditions. Liverani and Wojtkowski [LW94] proved that analogous contraction properties hold with respect to Thompson's part metric. These results, which were obtained from algebraic properties of the linear symplectic semigroup associated to a Riccati equation, are reminiscent of Birkhoff's theorem in Perron-Frobenius theory (on the contraction of positive linear operators sending a cone to its interior [Bir57]). Lawson and Lim [LL07] generalized these results to the infinite dimensional setting, and derived analogous contraction properties for the flow of the differential Riccati equation

$$\dot{P} = A'P + PA - P\Sigma P + Q, \qquad P(0) = G \ , \tag{2.1}$$

where A is a square matrix, Σ, Q are positive semidefinite matrices, and G is a positive definite matrix. Moreover, Lee and Lim [LL08] showed that the same contraction properties hold more generally for a family of Finsler metrics invariant under the action of the linear group (the latter metrics arise from symmetric gauge functions).

It is natural to ask whether the contraction properties remain valid for more general equations, like the following constrained differential Riccati equation,

$$\begin{aligned} &\dot{P} = A'P + PA + C'PC + Q \\ &\qquad -(PB + C'PD + L')(R + D'PD)^{-1}(B'P + D'PC + L), \\ &P(0) = G \\ &R + D'PD \quad \text{positive definite,} \end{aligned} \tag{2.2}$$

which has received a considerable attention in stochastic linear quadratic optimal control. The equation (2.2) is known as the *generalized Riccati differential equation* (GRDE) or as the *stochastic Riccati differential equation*. Up to a reversal of time, it is a special case (deterministic matrix coefficients) of the Backward stochastic Riccati differential equation, which has been extensively studied, see in particular [YZ99, CLZ98, RCMZ01]. The reader is referred to the monograph by Yong and Zhou [YZ99] for an introduction. Even for the simpler Riccati equation (2.1), contraction properties have not been established when the matrices Q, Σ are not positive semidefinite, whereas this situation does occur in applications (see Chapter 7).

In this chapter, motivated by the analysis of the generalized Riccati equation, we study the general question of computing the contraction rate in Thompson's part metric of an arbitrary order-preserving (time-dependent) flow defined on a subset of the interior of a closed convex and pointed cone in a possibly infinite dimensional Banach space. Recall that the order associated with such a cone $\mathscr{C}$ is defined by $x \preccurlyeq y \Leftrightarrow y - x \in \mathscr{C}$, and that the Thompson metric can be defined on the interior of $\mathscr{C}$ by the formula

$$d_T(x,y) := \log(\max\{M(x/y), M(y/x)\})$$

where

$$M(x/y) := \inf\{t \in \mathbb{R} : ty \succcurlyeq x\} = \sup_{\psi \in \mathscr{C}^\star} \frac{\psi(x)}{\psi(y)} \ ,$$

and $\mathscr{C}^\star$ denotes the dual cone of $\mathscr{C}$. More background can be found in Section 2.2.1.

Our first main result can be stated as follows.

Theorem 2.1. *Assume that the flow of the differential equation $\dot{x}(t) = \phi(t, x(t))$ is order-preserving with respect to the cone $\mathscr{C}$, and let $\mathscr{U}$ denote an open domain included in the interior of this cone such that $\lambda\mathscr{U} \subset \mathscr{U}$ holds for all $\lambda \in (0,1]$. Then, the contraction rate of the flow over a time interval J, on the domain $\mathscr{U}$, with respect to Thompson metric, is given by the formula*

$$\alpha := - \sup_{s\in J,\, x\in\mathscr{U}} M\big((D\phi_s(x)x - \phi(s,x))/x\big) \ . \tag{2.3}$$

Here, $D\phi_s(x)$ denotes the derivative of the map $(s,x) \mapsto \phi(s,x)$ with respect to the variable x. We make some basic technical assumptions (continuity, Lipschitz character on the function ϕ with respect to the second variable) to make sure that the flow is well defined. We refer the reader to Section 2.3 for more information, and in particular to Theorem 2.5 below, where the definition of the contraction rate can be found. We also show that under additional technical assumptions, the bound (2.3) on the contraction can be refined, the supremum over $s \in J$ being replaced by a mean over $s \in J$, see Theorem 2.7 below.

The idea of the proof is to construct a special flow-invariant set, appealing to a generalization due to Martin [Mar73] of theorems of Bony [Bon69] and Brezis [Bre70] on the geometric characterization of flow invariance. Formula (2.3) should be compared with results of Nussbaum, who studied the more general question of computing the contraction rate of a not necessarily order-preserving flow in Thompson metric [Nus94] , and obtained an explicit formula in the special case of the standard positive cone. The effective evaluation of the contraction rate becomes difficult in the non-order-preserving case, whereas the more special characterization (2.3) is useful from an algorithmic perspective (evaluating the term $M(\cdot/\cdot)$ there reduces to computing the dominant eigenvalue of a positive definite matrix). We also note that Nussbaum's approach, which relies on the Finsler structure of Thompson's part metric, is widely applicable in its spirit but leads to different technical assumptions, including geodesic convexity. See Section 2.7 for a detailed comparison.

As a first illustration, we show, in Section 2.4.2, that the contraction results of Liverani and Wojtkowski [LW94] and of Lawson and Lim [LL07] concerning the standard Riccati equation (2.1) with positive semidefinite matrices Σ, Q, as well as new contraction results in the case when Σ is not positive semidefinite, can be recovered, or obtained, by an application of Formula (2.3). This provides an alternative to the earlier approaches, which relied on the theory of symplectic semigroups. This will allow us to handle as well situations in which the symplectic structure is missing, as it is the case of the generalized Riccati differential equation, see Section 2.5.

Our second main result shows that the flow of the generalized Riccati differential equation is a local contraction in Thompson metric.

Theorem 2.2. *Assume that the coefficients of the generalized Riccati differential equation* (2.2) *are constant, and that the matrix $\left(\begin{smallmatrix} Q & L' \\ L & R \end{smallmatrix}\right)$ is positive definite. Then, the flow of this equation is a strict contraction on the interior of the cone of positive definite matrices, and this contraction is uniform on any subset that is bounded from above in the Loewner order.*

This theorem follows from Theorem 2.11 in Section 2.5, where an explicit bound for the contraction rate on an interval in the Loewner order is given. We shall also see in Section 2.5 that the flow of the generalized Riccati equation is no longer a uniform contraction on the interior of the cone, which reveals a fundamental discrepancy with the case of the standard Riccati equation. Then, motivated by earlier results of Chen, Moore, Ait Rami, and Zhou (see [RCMZ01] and [RZ00]) on the asymptotic behavior of the GRDE, we identify (Theorem 2.13) different assumptions under which a trajectory of the GRDE converges exponentially to a stable solution of the associated Generalized Algebraic Riccati Equation (GARE). We also establish (Section 2.5.4) analogous results concerning the discrete

time case. Then, we give a necessary and sufficient condition (Proposition 2.19) for the generalized discrete Riccati operator to be a strict global contraction.

Finally, in Section 2.6, we establish the following negative result, which shows that the Thompson metric is essentially the only invariant Finsler metric in which the flow of the GRDE is non-expansive for all admissible values of the matrix data.

Theorem 2.3. *The flow of the generalized Riccati differential equation is non-expansive in the invariant Finsler metric arising from a symmetric gauge function, regardless of the matrix parameters (A,B,C,D,L,Q,R), if and only if this symmetric gauge function is a scalar multiple of the sup-norm.*

In particular, the flow of the GRDE is *not* non-expansive in the invariant Riemannian metric, showing that Bougerol's theorem on the contraction of the standard discrete Riccati equation does not carry over to the GRDE.

2.2 Preliminaries

2.2.1 Thompson's part metric

We first recall the definition and basic properties of Thompson's part metric.

Throughout the chapter, $(\mathscr{X}, \|\cdot\|)$ is a real Banach space. Denote by $\mathscr{X}^\star$ the dual space of $\mathscr{X}$. For any $x \in \mathscr{X}$ and $q \in \mathscr{X}^\star$, denote by $\langle q,x\rangle$ the value of $q(x)$. Let $\mathscr{C} \subset \mathscr{X}$ be a closed pointed convex cone, i.e., $\alpha\mathscr{C} \subset \mathscr{C}$ for $\alpha \in \mathbb{R}^+$, $\mathscr{C}+\mathscr{C} \subset \mathscr{C}$ and $\mathscr{C} \cap (-\mathscr{C}) = 0$. The *dual cone* of $\mathscr{C}$ is defined by

$$\mathscr{C}^\star = \{z \in \mathscr{X}^\star : \langle z,x\rangle \geqslant 0 \ \ \forall x \in \mathscr{C}\} \ .$$

We denote by $\mathscr{C}_0$ the interior of $\mathscr{C}$. We define the partial order $\preccurlyeq$ induced by $\mathscr{C}$ on $\mathscr{X}$ by

$$x \preccurlyeq y \Leftrightarrow y - x \in \mathscr{C}$$

so that

$$x \preccurlyeq y \Rightarrow \langle z,x\rangle \leqslant \langle z,y\rangle \,, \ \ \forall z \in \mathscr{C}^\star.$$

We also define the relation $\prec$ by

$$x \prec y \Leftrightarrow y - x \in \mathscr{C}_0 \ .$$

For $x \preccurlyeq y$ we define the order intervals:

$$[x,y] := \{z \in \mathscr{X} | x \preccurlyeq z \preccurlyeq y\}, \qquad (x,y) := \{z \in \mathscr{X} | x \prec z \prec y\}.$$

For $x \in \mathscr{X}$ and $y \in \mathscr{C}_0$, following [Nus88], we define

$$\begin{aligned} M(x/y) &:= \inf\{t \in \mathbb{R} : x \preccurlyeq ty\} \\ m(x/y) &:= \sup\{t \in \mathbb{R} : x \succcurlyeq ty\} \end{aligned} \tag{2.4}$$

Observe that since $y \in \mathscr{C}_0$, and since $\mathscr{C}$ is closed and pointed, the two sets in (2.4) are non-empty, closed, and bounded from below and from above, respectively. In particular, m and M take finite values.

Definition 2.1 ([Tho63])**.** The *Thompson part metric* between two elements x and y of $\mathscr{C}_0$ is

$$d_T(x,y) := \log(\max\{M(x/y), M(y/x)\}) \ . \tag{2.5}$$

It can be verified [Tho63] that $d_T(\cdot,\cdot)$ defines a metric on $\mathscr{C}_0$, namely for any $x,y,z \in \mathscr{C}_0$ we have

$$d_T(x,y) \geqslant 0,\ \ d_T(x,y) = d_T(y,x),\ \ d_T(x,z) \leqslant d_T(x,y) + d_T(y,z),\ \ d_T(x,y) = 0 \Leftrightarrow x = y.$$

The cone $\mathscr{C}$ is normal if there is a constant $K > 0$ such that

$$0 \preccurlyeq x \preccurlyeq y \Rightarrow \|x\| \leqslant K\|y\|.$$

Note that in finite dimensional case, a (closed convex pointed) cone $\mathscr{C}$ is automatically a normal cone. A sufficient condition for $\mathscr{C}_0$ to be complete with respect to $d_T(\cdot,\cdot)$ is that $\mathscr{C}$ is a normal cone, see [Tho63]. We shall consider specially the following two examples: the standard orthant cone and the cone of positive semidefinite matrices.

Example 2.1. We consider the space $\mathscr{X} = \mathbb{R}^n$ and the standard orthant cone $\mathscr{C} = \mathbb{R}^n_+$. We denote by $\operatorname{int}\mathbb{R}^n_+$ the interior of $\mathbb{R}^n_+$. It can be checked that the partial order $\preccurlyeq$ is the pointwise order, i.e., for all $x,y \in \mathbb{R}^n$,

$$x \preccurlyeq y \Leftrightarrow x_i \leqslant y_i,\ \ 1 \leqslant i \leqslant n.$$

Besides, for all $x,y \in \mathscr{C}_0$,

$$M(x/y) = \max_{1\leqslant i\leqslant n} x_i/y_i\ \ ,$$

and Thompson's part metric can be explicitly computed from (2.5).

Example 2.2. Let $\mathscr{X} = \mathrm{S}_n$, the space of Hermitian matrices of dimension n and $\mathscr{C} = \mathrm{S}_n^+ \subset \mathrm{S}_n$ (resp. $\mathscr{C}_0 = \hat{\mathrm{S}}_n^+$), the cone of positive semidefinite matrices (resp. the cone of positive definite matrices). Then the partial order $\preccurlyeq$ is the Loewner order, i.e., for all $A,B \in \mathrm{S}_n$,

$$A \preccurlyeq B \Leftrightarrow x'Ax \leqslant x'Bx,\ \ \forall x \in \mathbb{R}^n\ \ .$$

It can be checked that, for all $A,B \in \mathscr{C}_0$,

$$M(A/B) = \max_{1\leqslant i\leqslant n} \lambda_i\ \ ,$$

where $\lambda_1,\ldots,\lambda_n$ are the eigenvalues of the matrix $B^{-1}A$ (the latter eigenvalues are real and positive) so that Thompson's part metric d_T can be explicitly computed from (2.5).

2.2.2 Characterization of flow invariant sets

We next recall some known results on the characterization of flow-invariant sets in terms of tangent cones, which will be used to characterize order-preserving non-expansive flows in Thompson's part metric.

In the sequel, $J = [0,T) \subset \mathbb{R}$ is a possibly unbounded interval, $\mathscr{D} \subset \mathscr{X}$ is an open set, and $\phi(t,x)$ is a function from $J \times \mathscr{D}$ to $\mathscr{X}$. For $x \in \mathscr{X}$ and $\mathscr{S} \subset \mathscr{X}$ we define the distance function:

$$d(x,\mathscr{S}) = \inf\{|x-y| : y \in \mathscr{S}\}.$$

We study the following Cauchy problem:

$$\begin{cases} \dot{x}(t) = \phi(t,x(t)), \\ x(s) = x_0. \end{cases} \tag{2.6}$$

By a solution of (2.6) on $[s,a) \subset J$ we mean a differentiable function $t \mapsto x(t) : [s,a) \to \mathscr{D}$ such that $x(s) = x_0$ and $\dot{x}(t) = \phi(t,x(t))$ for all $t \in [s,a)$. By an absolutely continuous solution of (2.6) on $[s,a) \subset J$ we mean an absolutely continuous function $t \mapsto x(t) : [s,a) \to \mathscr{D}$ such that $x(s) = x_0$ and $\dot{x}(t) = \phi(t,x(t))$ for almost everywhere $t \in [s,a)$.

Let $\mathscr{S}$ be a closed subset of $\mathscr{X}$. We say that the system $(\mathscr{S} \cap \mathscr{D}, \phi)$ is *flow-invariant* if every absolutely continuous solution of (2.6) leaves $\mathscr{S}$ invariant, in the sense that for any $s \in J$ and $x_0 \in \mathscr{S} \cap \mathscr{D}$, the solution $x(t)$ must be in $\in \mathscr{S} \cap \mathscr{D}$, for all $t \in [s,a)$.

Characterizations of flow invariant sets go back to the works of Bony [Bon69] and Brezis [Bre70]. Several improvements, together with extensions to the infinite dimensional case can be found in [Red72], [Mar73], [Cla75] and [RW75]. We shall actually need here an immediate consequence of a theorem of Martin [Mar73].

Theorem 2.4 (Theorem 1 of [Mar73])**.** *Suppose that the following conditions hold:*

(C1) ψ *is a continuous function on* $J \times \mathscr{D}$*;*

(C2) For every closed bounded set $K \subset \mathscr{D}$*, there is a constant* $L > 0$ *such that*

$$|\psi(t,x) - \psi(t,y)| \leqslant L|x-y|, \quad \forall t \in J, x, y \in K \ ;$$

(C3) For all $t \in J$ *and* $x \in \mathscr{S} \cap \mathscr{D}$*,*

$$\lim_{h \downarrow 0} \frac{d(x + h\psi(t,x), \mathscr{S} \cap \mathscr{D})}{h} = 0 \ .$$

(C4) $\mathscr{S}$ *is convex.*

Then the system $(\mathscr{S} \cap \mathscr{D}, \psi)$ *is flow-invariant.*

It is not difficult to prove that for $x \in \mathscr{S} \cap \mathscr{D}$, $v \in \mathscr{X}$ and sufficiently small $h > 0$,

$$d(x + hv, \mathscr{S}) = d(x + hv, \mathscr{S} \cap \mathscr{D}).$$

Thus, Condition (C3) is equivalent to:

(C5) For all $t \in J$ and $x \in \mathscr{S} \cap \mathscr{D}$,

$$\lim_{h \downarrow 0} \frac{d(x + h\psi(t,x), \mathscr{S})}{h} = 0.$$

Condition (C2) is a local Lipschitz condition for the function ψ, with respect to the second variable. Condition (C3) is a tangency condition (the vector field ψ should not point outward the set $\mathscr{S} \cap \mathscr{D}$).

Definition 2.2 (Tangent cone [Cla75])**.** The *tangent cone* to a closed set $\mathscr{S} \subset \mathscr{X}$ at a point $x \in \mathscr{S}$, written $T_{\mathscr{S}}(x)$, is the set of vectors v such that:

$$\liminf_{h \downarrow 0} \frac{d(x + hv, \mathscr{S})}{h} = 0. \tag{2.7}$$

Remark 2.3. When $\mathscr{S}$ is a closed convex, we know that the limit in (2.7) exists. Thus, Condition (C5), equivalent to (C3) in Theorem 2.4, can be replaced by:

(C6) For all $t \in J$ and $x \in \mathscr{S} \cap \mathscr{D}$,

$$\psi(t,x) \in T_{\mathscr{S}}(x).$$

Besides, this definition coincides with the one in convex analysis, i.e.,

Proposition 2.1 (Proposition 5.5, Exercice 7.2 [CLSW98]). *Let $\mathscr{S}$ be a closed convex set of $\mathscr{X}$, then,*

$$T_{\mathscr{S}}(x) = \mathrm{cl}\{v : \exists \lambda > 0 \text{ with } x + \lambda v \in \mathscr{S}\}, \ \ \forall x \in \mathscr{S}.$$

Now the flow-invariance can be checked by verifying if $\phi(t,x)$ lies in the tangent cone of $\mathscr{S}$. To this end, we need to compute the tangent cone at each point of $\mathscr{S}$. In some cases the tangent cone can be expressed in a simple way:

Lemma 2.2 (Exercise 2.5.3 [CLSW98]). *Let $\mathscr{S}_1, \mathscr{S}_2 \subset \mathscr{X}$ be closed subsets, $x = (x_1, x_2) \in \mathscr{S}_1 \times \mathscr{S}_2$. Then*

$$T_{\mathscr{S}_1 \times \mathscr{S}_2}(x) = T_{\mathscr{S}_1}(x_1) \times T_{\mathscr{S}_2}(x_2).$$

We shall consider specially $\mathscr{S} = \mathscr{C}$. Then, using Proposition 2.1 and the Hahn-Banach separation theorem, one can show that

$$T_{\mathscr{C}}(x) = \{v | \langle q, v \rangle \geqslant 0 \ \text{ if } \ q \in \mathscr{C}^\star \ \text{ and } \ \langle q, x \rangle = 0\}, \ \ x \in \mathscr{C}. \tag{2.8}$$

2.3 Contraction rate in Thompson metric of order-preserving flow

2.3.1 Preliminary results

From now until the end, the function $\phi(t,x)$ is assumed to be continuous on $J \times \mathscr{D}$ and Fréchet differentiable in x. The derivative of ϕ with respect to the second variable at point (t,x) is denoted by $D\phi_t(x)$. We also assume that the derivative is bounded on any closed bounded set, i.e., for any bounded set $K \subset \mathscr{D}$, there is a constant L such that:

$$|D\phi_t(x)| \leqslant L, \quad \forall t \in J, x \in K.$$

Therefore Condition (C1) and Condition (C2) are both satisfied. The existence and uniqueness of the solution of (2.6) follow from the Cauchy-Lipschitz Theorem. We then define the flow $M_\cdot^\cdot(\cdot)$ associated to the system by:

$$M_s^t(x_0) = x(t), \ \ s \in J, \ \ t \in [s,a)$$

where $x(t) : t \in [s,a)$ is the maximal solution of (2.6). (Note that in general the flow is defined only on a subset of $J \times J \times \mathscr{D}$.) For each open subset $\mathscr{U} \subset \mathscr{D}$ and initial value $x_0 \in \mathscr{U}$ we define $t_{\mathscr{U}}(s,x_0)$ as the first time when the trajectory leaves $\mathscr{U}$, i.e.,

$$t_{\mathscr{U}}(s,x_0) = \sup\{b \in (s,a) | M_s^t(x_0) \in \mathscr{U}, \ \ \forall t \in [s,b)\}.$$

When ϕ is independent of time t, we denote simply

$$M_t(x_0) := M_0^t(x_0), \quad t_{\mathscr{U}}(x_0) := t_{\mathscr{U}}(0,x_0), \ \ \forall x_0 \in \mathscr{D}, \mathscr{U} \subset \mathscr{D}.$$

By uniqueness of the solution, the flow has the group property:

$$M_s^t(x_0) = M_s^{t_1}(M_{t_1}^t(x_0)), \quad \forall \ \ 0 \leqslant s \leqslant t_1 \leqslant t < t_{\mathscr{D}}(s,x_0) \ \ .$$

Definition 2.3 (Order-preserving flow)**.** Let $\mathscr{U}$ be an open subset of $\mathscr{D}$. The flow $M_\cdot(\cdot)$ is said to be *order-preserving on* $\mathscr{U}$ if for all $x_1, x_2 \in \mathscr{U}$ such that $x_1 \preccurlyeq x_2$,

$$M_s^t(x_1) \preccurlyeq M_s^t(x_2), \ \ \forall \ 0 \leqslant s \leqslant t < t_{\mathscr{U}}(s,x_1) \wedge t_{\mathscr{U}}(s,x_2).$$

Definition 2.4 (Non-expansiveness and contraction)**.** Suppose that $\mathscr{C}_0 \subset \mathscr{D}$. The flow $M_\cdot(\cdot)$ is said to be *contractive* on $\mathscr{C}_0$ with rate α in Thompson metric if for all $x_1, x_2 \in \mathscr{C}_0$,

$$d_T(M_s^t(x_1), M_s^t(x_2)) \leqslant e^{-\alpha(t-s)} d_T(x_1,x_2), \ \ \forall \ 0 \leqslant s \leqslant t < t_{\mathscr{C}_0}(s,x_1) \wedge t_{\mathscr{C}_0}(s,x_2)$$

If the latter inequality holds with $\alpha = 0$, the flow is said to be *non-expansive*.

In the following, our primary goal is to characterize the best contraction rate for order-preserving flows in Thompson part metric. We shall need the following proposition, which provides a characterization of monotonicity in terms of the function ϕ. The equivalence of the first two assertions was proved in [RW75].

Proposition 2.3 (Compare with Theorem 3 in [RW75])**.** *Let $\mathscr{U}$ be an open subset of $\mathscr{D}$. The following conditions are equivalent:*

(a) The flow $M_\cdot(\cdot)$ is order-preserving on $\mathscr{U}$.

(b) For all $s \in J$ and $x_1, x_2 \in \mathscr{U}$ such that $x_1 \succcurlyeq x_2$, $\phi(s,x_1) - \phi(s,x_2) \in T_{\mathscr{C}}(x_1 - x_2)$.

If $\mathscr{U}$ is convex, then the above conditions are equivalent to:

(c) For all $s \in J$, $x \in \mathscr{U}$ and $v \in \mathscr{C}$,

$$\langle q, D\phi_s(x)v \rangle \geqslant 0, \ \ \forall q \in \{q \in \mathscr{C}^\star : \langle q, v \rangle = 0\}. \tag{2.9}$$

Proof. We only need to prove the equivalence between (b) and (c), since the equivalence between (a) and (b) follows from [RW75]. In view of (2.8), Condition (b) is equivalent to the following:

for all $s \in J$ and $x_1, x_2 \in \mathscr{U}$ such that $x_1 \succcurlyeq x_2$,

$$\langle q, \phi(s,x_1) - \phi(s,x_2) \rangle \geqslant 0, \ \ \forall q \in \{q \in \mathscr{C}^\star : \langle q, x_1 - x_2 \rangle = 0\}.$$

Now suppose that (b) is true. Then for any $s \in J$, $x \in \mathscr{U}$ and any $v \in \mathscr{C}$, there is $\delta > 0$ such that for any $0 \leqslant \varepsilon \leqslant \delta$

$$\langle q, \phi(s,x+\varepsilon v) - \phi(s,x) \rangle \geqslant 0, \ \ \forall q \in \{q \in \mathscr{C}^\star : \langle q, v \rangle = 0\}.$$

Since ϕ is differentiable at point x, dividing by ε the latter inequality, and letting ε tend to 0, we get

$$\langle q, D\phi_s(x)v \rangle \geqslant 0, \ \ \forall q \in \{q \in \mathscr{C}^\star : \langle q, v \rangle = 0\}.$$

Next suppose that Condition (c) holds. Fix any $s \in J$ and $x_1, x_2 \in \mathscr{U}$ such that $x_1 \succcurlyeq x_2$. Fix any $q \in \mathscr{C}^\star$ such that $\langle q, x_1 - x_2 \rangle = 0$. Define the function $g : [0,1] \to \mathbb{R}$ by:

$$g(\lambda) = \langle q, \phi(s, \lambda x_1 + (1-\lambda)x_2) - \phi(s,x_2) \rangle.$$

Then we have $g(0) = 0$ and in view of convexity of $\mathscr{U}$ and (2.9),

$$g'(\lambda) = \langle q, D\phi_s(\lambda x_1 + (1-\lambda)x_2)(x_1 - x_2)) \rangle \geqslant 0, \ \ \forall 0 \leqslant \lambda \leqslant 1.$$

A standard argument establishes that:

$$g(1) = \langle q, \phi(s,x_1) - \phi(s,x_2) \rangle \geqslant 0. \tag{2.10}$$

Since s, x_1, x_2 and q are arbitrary, we deduce Condition (b). □

2.3.2 Characterization of the contraction rate in terms of flow invariant sets

The following is a key technical result in the characterization of the contraction rate of the flow.

Proposition 2.4. *Let $\mathscr{U} \subset \mathscr{D}$ be an open set such that $\lambda \mathscr{U} \subset \mathscr{U}$ for all $\lambda \in (0,1]$. If the flow $M_\cdot^\cdot(\cdot)$ is order-preserving on $\mathscr{U}$, then the following conditions are equivalent:*

(a) For all $x \in \mathscr{U}$ and $\lambda \geqslant 1$ such that $\lambda x \in \mathscr{U}$,

$$M_s^t(\lambda x) \preccurlyeq \lambda^{e^{-\alpha(t-s)}} M_s^t(x), \;\; 0 \leqslant s \leqslant t < t_{\mathscr{U}}(s,x) \wedge t_{\mathscr{U}}(s,\lambda x).$$

(b) For all $s \in J$ and $x \in \mathscr{U}$,

$$D\phi_s(x)x - \phi(s,x) \preccurlyeq -\alpha x.$$

(c) For all $x, y \in \mathscr{U}$ and $\lambda \geqslant 1$ such that $y \preccurlyeq \lambda x$,

$$M_s^t(y) \preccurlyeq \lambda^{e^{-\alpha(t-s)}} M_s^t(x), 0 \leqslant s \leqslant t < t_{\mathscr{U}}(s,x) \wedge t_{\mathscr{U}}(s,y).$$

Proof. Suppose Condition (a) holds. Let x be any point in $\mathscr{U}$. Fix any $\lambda > 1$ such that $\lambda x \in \mathscr{U}$, we must have:

$$M_s^t(\lambda x) \preccurlyeq \lambda^{e^{-\alpha(t-s)}} M_s^t(x), \qquad 0 \leqslant t < t_{\mathscr{U}}(s,x) \wedge t_{\mathscr{U}}(s,\lambda x). \tag{2.11}$$

where it must be the case that $t_{\mathscr{U}}(s,x) \wedge t_{\mathscr{U}}(s,\lambda x) > 0$. Since the terms on both sides of (2.11) coincide when $t = s$, taking the derivative of each of these terms at $t = s$, we obtain

$$\phi(s,\lambda x) \preccurlyeq \lambda \phi(s,x) - \alpha(\lambda \ln \lambda)x \tag{2.12}$$

Since this inequality holds for all $\lambda \geqslant 1$ such that $\lambda x \in \mathscr{U}$, with equality for $\lambda = 1$, the derivation of the two sides of the above inequality at $\lambda = 1$ leads to:

$$D\phi_s(x)x - \phi(s,x) \preccurlyeq -\alpha x$$

for all $x \in \mathscr{U}$. Condition (b) is deduced.

Now suppose that Condition (b) is true. We shall derive Condition (c) by constructing an invariant set. Denote:

$$\tilde{\mathscr{X}} := \mathscr{X} \times \mathscr{X} \times \mathbb{R},$$
$$\tilde{\mathscr{D}} := \mathscr{U} \times \mathscr{U} \times \mathbb{R}^+ \backslash \{0\},$$
$$\mathscr{S} := \{(x_1, x_2, \lambda) \in \tilde{\mathscr{X}} : x_2 \preccurlyeq \lambda x_1, \lambda \geqslant 1\}.$$

Define the differential equation on $\tilde{\mathscr{D}}$:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{\lambda} \end{pmatrix} = \Phi(t, x_1, x_2, \lambda) := \begin{pmatrix} \phi(t,x_1) \\ \phi(t,x_2) \\ -\alpha\lambda \ln \lambda \end{pmatrix} \tag{2.13}$$

It is not difficult to see that Condition (c) is equivalent to the flow-invariance of the system $(\mathscr{S} \cap \tilde{\mathscr{D}}, \Phi)$. It would be natural to show directly the latter flow-invariance by appealing to Theorem 2.4, but the set $\mathscr{S}$ is not convex, making it harder to check the assumptions of this theorem. Therefore, we make a change of variable to replace $\mathscr{S}$ by a convex set.

Define the smooth function $F : \tilde{\mathscr{X}} \to \tilde{\mathscr{X}}$ by:

$$F(x_1,x_2,\lambda) = (x_1, \lambda x_1 - x_2, \lambda - 1), \ \ \forall (x_1,x_2,\lambda) \in \tilde{\mathscr{X}}.$$

Denote

$$\mathscr{S}' = \mathscr{X} \times \mathscr{C} \times \mathbb{R}^+.$$

By Lemma 2.2, for $(y_1,y_2,\kappa) \in \mathscr{S}'$,

$$T_{\mathscr{S}'}(y_1,y_2,\kappa) = \mathscr{X} \times T_{\mathscr{C}}(y_2) \times T_{\mathbb{R}^+}(\kappa).$$

Observe that $\mathscr{S} = \{x \in \tilde{\mathscr{X}} | F(x) \in \mathscr{S}'\}$ and that F has a smooth inverse $G : \tilde{\mathscr{X}} \to \tilde{\mathscr{X}}$ given by:

$$G(y_1,y_2,\kappa) = (y_1, (\kappa+1)y_1 - y_2, \kappa+1).$$

Therefore $F(\tilde{\mathscr{D}}) = G^{-1}(\tilde{\mathscr{D}})$ is an open set. Let $(y_1,y_2,\kappa) = F(x_1,x_2,\lambda)$ and consider the system:

$$(\dot{y}_1,\dot{y}_2,\dot{\kappa})' = \Psi(t,y_1,y_2,\kappa) \tag{2.14}$$

where

$$\Psi(t,y_1,y_2,\kappa) := \begin{pmatrix} \phi(t,y_1) \\ -\alpha(\kappa+1)\ln(\kappa+1)y_1 + (\kappa+1)\phi(t,y_1) - \phi(t,(\kappa+1)y_1 - y_2) \\ -\alpha(\kappa+1)\ln(\kappa+1) \end{pmatrix}$$

One can verify that the invariance of the system $(\mathscr{S} \cap \tilde{\mathscr{D}}, \Phi)$ is equivalent to the invariance of the system $(\mathscr{S}' \cap F(\tilde{\mathscr{D}}), \Psi)$.

Now the function $\Psi : J \times F(\tilde{\mathscr{D}}) \to \tilde{\mathscr{X}}$ defined as above is continuous and differentiable to the second variable with bounded derivative on bounded set. Besides $\mathscr{S}'$ is convex. By applying Theorem 2.4, the system $(\mathscr{S}' \cap F(\tilde{\mathscr{D}}), \Psi)$ is flow-invariant if the following condition is satisfied:

$$\Psi(s,y_1,y_2,\kappa) \in T_{\mathscr{S}'}(y_1,y_2,\kappa), \ \ \forall s \in J, (y_1,y_2,\kappa) \in \mathscr{S}' \cap F(\tilde{\mathscr{D}}). \tag{2.15}$$

That is, for any $(y_1,y_2,\kappa) \in \mathscr{S}' \cap F(\tilde{\mathscr{D}})$ and $s \in J$,

$$\begin{cases} \phi(s,y_1) \in \mathscr{X} \\ -\alpha(\kappa+1)\ln(\kappa+1)y_1 + (\kappa+1)\phi(s,y_1) - \phi(s,(\kappa+1)y_1 - y_2) \in T_{\mathscr{C}}(y_2) \\ -\alpha(\kappa+1)\ln(\kappa+1) \in T_{\mathbb{R}^+}(\kappa) \end{cases}$$

It suffices to check the second condition because the others hold trivially. By applying the bijection F, this condition becomes: for any $s \in J$ and $(x_1,x_2,\lambda) \in \mathscr{S} \cap \tilde{\mathscr{D}}$,

$$-\alpha\lambda \ln \lambda x_1 + \lambda\phi(s,x_1) - \phi(s,x_2) \in T_{\mathscr{C}}(\lambda x_1 - x_2).$$

Let any $s \in J$, $x_1,x_2 \in \mathscr{U}$ and $\lambda \geqslant 1$ such that $x_2 \preccurlyeq \lambda x_1$. Let any $q \in \mathscr{C}^\star$ such that $\langle q, \lambda x_1 - x_2 \rangle = 0$. By (2.8) we only need to prove:

$$\langle q, -\alpha\lambda \ln \lambda x_1 + \lambda\phi(s,x_1) - \phi(s,x_2) \rangle \geqslant 0. \tag{2.16}$$

By the assumptions, we know that $\lambda^{-1}x_2 \in \mathscr{U}$. Then, it suffices to prove: for any $x_1,x_2 \in \mathscr{U}$ such that $x_1 \succcurlyeq x_2$, let $q \in \mathscr{C}^\star$ such that $\langle q, x_1 - x_2 \rangle = 0$, then for any $\lambda \geqslant 1$ such that $\lambda x_2 \in \mathscr{U}$ we have:

$$\langle q, -\alpha\lambda \ln \lambda x_1 + \lambda\phi(s,x_1) - \phi(s,\lambda x_2) \rangle \geqslant 0.$$

Define the function $f:[1,\lambda]\to\mathbb{R}$ by:

$$f(\tau)=\langle q,-\alpha\ln\tau x_1+\phi(s,x_1)-\tau^{-1}\phi(s,\tau x_2)\rangle$$

Notice that the function f is well defined on $[1,\lambda]$. By hypothesis of monotonicity and Proposition 2.3,

$$f(1)=\langle q,\phi(s,x_1)-\phi(s,x_2)\rangle\geqslant 0.$$

Differentiating f gives, for all $\tau\in[1,\lambda]$,

$$\begin{aligned}f'(\tau)&=\langle q,-\tau^{-1}\alpha x_1+\tau^{-2}\phi(s,\tau x_2)-\tau^{-1}D\phi_s(\tau x_2)x_2\rangle\\&\geqslant\langle q,-\tau^{-1}\alpha x_1+\tau^{-1}\alpha x_2\rangle\qquad\text{(by Condition (b))}\\&=0\ .\end{aligned}$$

A standard argument establishes that $f(\lambda)\geqslant 0$, and so (2.16) is proved, whence the flow-invariance of $(\mathscr{S}\cap\tilde{\mathscr{D}},\Phi)$, which is exactly Condition (c). Finally, Condition (a) follows from Condition (c) by considering $y=\lambda x$. □

We next state the main results. Recall that $J=[0,T)\subset\mathbb{R}$.

Theorem 2.5 (Contraction rate)**.** *Assume that ϕ is defined on $J\times\mathscr{U}$ where $\mathscr{U}\subset\mathscr{C}_0$ is an open set in the interior of the cone such that $\lambda\mathscr{U}\subset\mathscr{U}$ for all $\lambda\in(0,1]$. If the flow $M_\cdot(\cdot)$ is order-preserving on $\mathscr{U}$, then the best constant α such that*

$$d_T(M_s^t(x_1),M_s^t(x_2))\leqslant e^{-\alpha(t-s)}d_T(x_1,x_2),\ \ 0\leqslant s\leqslant t<t_{\mathscr{U}}(s,x_1)\wedge t_{\mathscr{U}}(s,x_2)\tag{2.17}$$

holds for all $x_1,x_2\in\mathscr{U}$ is given by

$$\alpha:=-\sup_{s\in J,\,x\in\mathscr{U}}M\big((D\phi_s(x)x-\phi(s,x))/x\big)\ .\tag{2.18}$$

Proof. If (2.17) holds for all $x_1,x_2\in\mathscr{U}$, then Condition (a) in Proposition (2.4) holds. It follows that the constant α must satisfy

$$D\phi_s(x)x-\phi(s,x)\preccurlyeq-\alpha x,\ \ \forall s\in J,x\in\mathscr{U}.\tag{2.19}$$

Now conversely if (2.19) holds. Then Condition (c) in Proposition 2.4 holds. For any $x_1,x_2\in\mathscr{U}$, let $\lambda=e^{d_T(x_1,x_2)}$, then

$$M_s^t(x_1)\preccurlyeq\lambda^{e^{-\alpha(t-s)}}M_s^t(x_2),0\leqslant s\leqslant t<t_{\mathscr{U}}(s,x_2)\wedge t_{\mathscr{U}}(s,x_1)\ .$$

The same is true if we exchange the roles of x_1 and x_2, and so, (2.17) holds for all $x_1,x_2\in\mathscr{U}$. Consequently the best constant α such that (2.17) holds for all $x_1,x_2\in\mathscr{U}$ must be the greatest constant α such that (2.19) holds, which is precisely (2.18). □

Remark 2.4. The reason that we suppose $\lambda\mathscr{U}\subset\mathscr{U}$ for all $\lambda\in(0,1]$ is to have $\lambda^{-1}x_2\in\mathscr{U}$ for all $x_1,x_2\in\mathscr{U}$ and $\lambda\geqslant 1$ such that $x_2\preccurlyeq\lambda x_1$ (see the proofs after (2.16)). Such condition can be weakened in the following way. Let

$$\lambda_0=\sup_{x,y\in\mathscr{U}}M(x/y).$$

Denote

$$\alpha(\mathscr{U},\lambda_0) := -\sup_{\substack{s\in J,\lambda_0^{-1}\leqslant\lambda\leqslant 1\\ x\in\lambda\mathscr{U}}} M((D\phi_s(x)x-\phi(s,x))/x).$$

Suppose that we know a priori that the flow is order-preserving and non-expansive, then for all $x_1,x_2\in\mathscr{U}$,

$$d_T(M_s^t(x_1),M_s^t(x_2))\leqslant e^{-\alpha(\mathscr{U},\lambda_0)(t-s)}d_T(x_1,x_2),\ \ 0\leqslant s\leqslant t<t_{\mathscr{U}}(s,x_1)\wedge t_{\mathscr{U}}(s,x_2)$$

Clearly the assumption $\lambda\mathscr{U}\subset\mathscr{U}$ for all $\lambda\in(0,1]$ is like taking $\lambda_0=+\infty$.

Now we get a direct corollary.

Theorem 2.6. *Suppose that ϕ is defined on $J\times\mathscr{C}_0$. Let $\alpha\in\mathbb{R}$. If the flow is order-preserving on $\mathscr{C}_0$, then the following are equivalent:*

(a) For all $x_1,x_2\in\mathscr{C}_0$:

$$d_T(M_s^t(x_1),M_s^t(x_2))\leqslant e^{-\alpha(t-s)}d_T(x_1,x_2),\ \ 0\leqslant s\leqslant t<t_{\mathscr{C}_0}(s,x_1)\wedge t_{\mathscr{C}_0}(s,x_2).$$

(b) For all $s\in J$ and $x\in\mathscr{C}_0$,

$$D\phi_s(x)x-\phi(s,x)\preccurlyeq -\alpha x.$$

If any of these conditions holds, then the flow leaves $\mathscr{C}_0$ invariant, i.e., for any $s\in J$ and $x\in\mathscr{C}_0$, $t_{\mathscr{C}_0}(s,x)=T$.

Proof. The equivalence between (a) and (b) follows from Theorem 2.5. Now suppose that Condition (b) holds. Let any $s\in J$ and $x_1,x_2\in\mathscr{C}_0$. Let $t_1=t_{\mathscr{C}_0}(s,x_1)$ and $t_2=t_{\mathscr{C}_0}(s,x_2)$. Suppose that $t_1<t_2$. Then it must be the case that $t_1<+\infty$. Thus the set $\{M_s^r(x_2):r\in[s,t_1]\}$ is compact and included in $\mathscr{C}_0$. Denote

$$K=\max\{d_T(M_s^r(x_2),M_s^{t_1}(x_2))|\ \ r\in[s,t_1]\}<+\infty$$

and $K_0=K+\max\{e^{-\alpha(t_1-s)},1\}d_T(x_1,x_2)$. Note that there exists $s<\bar{r}<t_1$ such that

$$d_T(M_s^{\bar{r}}(x_1),M_s^{t_1}(x_2))>K_0,$$

otherwise $t_{\mathscr{C}_0}(s,x_1)>t_1$. But for any $s<r<t_1$,

$$\begin{aligned} d_T(M_s^r(x_1),M_s^{t_1}(x_2)) &\leqslant d_T(M_s^r(x_1),M_s^r(x_2))+d_T(M_s^r(x_2),M_s^{t_1}(x_2))\\ &\leqslant e^{-\alpha(r-s)}d_T(x_1,x_2)+d_T(M_s^r(x_2),M_s^{t_1}(x_2))\\ &\leqslant K_0. \end{aligned}$$

The contradiction implies that $t_1<t_2$ is impossible. We then showed that there exists $\bar{T}\in(0,+\infty]$ such that for any $s\in J$ and $x\in\mathscr{C}_0$, $t_{\mathscr{C}_0}(s,x)=\bar{T}$. From the group property of the flow action, we deduce that $\bar{T}=T$. □

2.3.3 Characterization of a time-dependent contraction rate in terms of flow invariant sets

We next refine the previous estimates of the contraction rate in the time-dependent case: Theorem 2.7 shows that the supremum over time $s\in J$ in the formula of Theorem 2.5 can be replaced by a mean over time. However, in the infinite-dimensional setting, we need to make stronger assumptions to arrive at this tighter estimate.

In particular we shall need to use the following notion. We say that a set S is a *distance set* if for all $x\in\mathscr{X}$, there is $y\in S$ such that $d(x,S)=|x-y|$.

Theorem 2.7. *Let ϕ be defined on $J\times\mathscr{U}$ where $\mathscr{U}\subset\mathscr{C}_0$ is an open set in the interior of the cone such that $\lambda\mathscr{U}\subset\mathscr{U}$ for all $\lambda\in(0,1]$. Assume that the function $\alpha(\cdot,\mathscr{U}):J\to\mathbb{R}$ defined by*

$$\alpha(s,\mathscr{U})=-\sup_{x\in\mathscr{U}}M((D\phi_s(x)x-\phi(s,x))/x),\ \ \forall s\in J$$

is locally integrable. Assume in addition that $\mathscr{C}$ is a distance set. If the flow $M_\cdot(\cdot)$ is order-preserving on $\mathscr{U}$, then for all $x,y\in\mathscr{U}$,

$$d_T(M_s^t(x),M_s^t(y))\leqslant\exp\big(-\int_s^t\alpha(r,\mathscr{U})dr\big)d_T(x,y),\ \ 0\leqslant s<t<t_{\mathscr{U}}(s,x)\wedge t_{\mathscr{U}}(s,y).$$

If we redo the proof of Proposition 2.4, we shall need to prove the invariance of the system $(S'\cap F(\tilde{\mathscr{D}}),\Psi)$ where S', F and $\tilde{\mathscr{D}}$ are the same and Ψ is now defined by:

$$\Psi(t,y_1,y_2,\kappa):=\begin{pmatrix}\phi(t,y_1)\\ -\alpha(t,\mathscr{U})(\kappa+1)\ln(\kappa+1)y_1+(\kappa+1)\phi(t,y_1)-\phi(t,(\kappa+1)y_1-y_2)\\ -\alpha(t,\mathscr{U})(\kappa+1)\ln(\kappa+1)\end{pmatrix}$$

Note that Ψ may not be continuous with respect to time so that Theorem 2.4 is no longer directly applicable to show the invariance. In fact, in this case we only have the existence of an absolutely continuous solution of system (2.13) and thus of System (2.14), the constant α being replaced by a time-dependent function $\alpha(\cdot,\mathscr{U})$. In order to prove the invariance of $(S'\cap F(\tilde{\mathscr{D}}),\Psi)$, we shall use the following invariance characterization:

Theorem 2.8 ([RW75])**.** *Suppose that the following conditions hold:*

(C1) For every closed bounded set $K\subset\mathscr{D}$, there is a locally integrable function $w(\cdot):J\to\mathbb{R}$ such that

$$|\psi(t,x)-\psi(t,y)|\leqslant w(t)|x-y|,\quad\forall t\in J,x,y\in K\ ;$$

(C2) For all $t\in J$ and $x\in\mathscr{S}\cap\mathscr{D}$,

$$\lim_{h\downarrow0}\frac{d(x+h\psi(t,x),\mathscr{S}\cap\mathscr{D})}{h}=0\ .$$

(C3) $\mathscr{S}$ is a distance set.

Then the system $(\mathscr{S}\cap\mathscr{D},\psi)$ is flow-invariant.

Theorem 2.8 follows from Theorem 2 in [RW75]. Assumption (C1) corresponds to the uniqueness condition (U_2) there. Assumption (C3) requires the distance set assumption. Note that Redheffer and Walter considered invariance for solutions in the strong sense. It is easy to check that the proofs are also valid for absolutely continuous solutions.

Once we have Theorem 2.8, Theorem 2.7 can be proved in the same way as Theorem 2.5. Therefore we omit the detail of the proofs.

Remark 2.5. In a reflexive Banach space, every closed convex set is a distance set [FHH$^+$01, 3.104].

2.3.4 Convergence rate characterization to a fixed point

In the sequel we suppose that the dynamics ϕ are independent of time and study the convergence of an orbit of the flow to a fixed point in the interior of the cone. Let $\bar{x} \in \mathscr{C}_0$ be such that $\phi(\bar{x}) = 0$. Let $\mu > 1$. Denote by $\mathscr{U}$ the open interval $(\mu^{-1}\bar{x}, \mu\bar{x})$. We look for the best constant $\alpha \in \mathbb{R}$ such that:

$$d_T(M_t(x), \bar{x}) \leqslant e^{-\alpha t} d_T(x, \bar{x}), \ \ \forall x \in \mathscr{U}, 0 \leqslant t < t_{\mathscr{U}}(x). \tag{2.20}$$

Theorem 2.9 (Convergence rate)**.** *We assume that ϕ is independent of time, defined on $\mathscr{C}_0$ and such that the flow is order-preserving on $\mathscr{C}_0$. Let $\bar{x} \in \mathscr{C}_0$ be a zero point of ϕ. Then the best constant α such that (2.20) holds is given by*

$$\alpha = \inf_{\mu^{-1} < \lambda < \mu} m\big((-(\lambda \ln \lambda)^{-1} \phi(\lambda \bar{x}))/\bar{x}\big) \ . \tag{2.21}$$

Moreover, if the latter α is non-negative, then for all $x \in [\mu^{-1}\bar{x}, \mu\bar{x}]$,

$$d_T(M_t(x), \bar{x}) \leqslant e^{-\alpha t} d_T(x, \bar{x}), \ \ \forall t \geqslant 0. \tag{2.22}$$

Proof. Suppose that α satisfies (2.20). Let any $\lambda \in (1, \mu)$. Then $\lambda\bar{x} \in \mathscr{U}$ and

$$M_t(\lambda \bar{x}) \preccurlyeq \lambda^{e^{-\alpha t}} \bar{x}, \ \ 0 \leqslant t < t_{\mathscr{U}}(\lambda \bar{x}).$$

Since $t_{\mathscr{U}}(\lambda\bar{x}) > 0$ and both sides of the former inequality coincide when $t = 0$, we get the inequality for the derivative at $t = 0$:

$$\phi(\lambda \bar{x}) \preccurlyeq -\alpha \lambda (\ln \lambda) \bar{x} \ , \tag{2.23}$$

and so

$$\alpha \bar{x} \preccurlyeq -(\lambda \ln \lambda)^{-1} \phi(\lambda \bar{x}), \ \ \forall 1 < \lambda < \mu. \tag{2.24}$$

Similarly, for $\lambda \in (\mu^{-1}, 1)$,

$$\lambda^{e^{-\alpha t}} \bar{x} \preccurlyeq M_t(\lambda \bar{x}), \ \ 0 \leqslant t < t_{\mathscr{U}}(\lambda \bar{x}),$$

thus

$$-\alpha \lambda \ln \lambda \bar{x} \preccurlyeq \phi(\lambda \bar{x}) \tag{2.25}$$

leading to

$$\alpha \bar{x} \preccurlyeq -(\lambda \ln \lambda)^{-1} \phi(\lambda \bar{x}), \ \ \forall \mu^{-1} < \lambda < 1. \tag{2.26}$$

It follows that α is bounded above by the expression in (2.21). To prove that conversely, (2.20) holds when α is given by (2.21), we use an invariance argument as in the proof of Proposition 2.4. Denote:

$$\tilde{\mathscr{X}} := \mathscr{X} \times \mathbb{R},$$

$$\mathscr{D} := \mathscr{U} \times (1, \mu),$$

$$\mathscr{S}_1 := \{(x, \lambda) \in \tilde{\mathscr{X}} : x \preccurlyeq \lambda \bar{x}\},$$

$$\mathscr{S}_2 := \{(x, \lambda) \in \tilde{\mathscr{X}} : \bar{x} \preccurlyeq \lambda x\},$$

and define the differential equation:

$$\begin{pmatrix} \dot{x} \\ \dot{\lambda} \end{pmatrix} = \Phi(x,\lambda) := \begin{pmatrix} \phi(x) \\ -\alpha\lambda\ln\lambda \end{pmatrix}. \tag{2.27}$$

Then (2.20) holds if $(\mathscr{S}_1 \cap \mathscr{D}, \Phi)$ and $(\mathscr{S}_2 \cap \mathscr{D}, \Phi)$ are invariant systems. Given the convexity of $\mathscr{S}_1$, we can directly apply Theorem 2.4 to prove the invariance of the system $(\mathscr{S}_1 \cap \mathscr{D}, \Phi)$. The tangent cone of $\mathscr{S}_1$ at point $(x,\lambda) \in \mathscr{S}_1$ is given by:

$$T_{\mathscr{S}_1}(x,\lambda) = \{(z,\eta) : \langle q, \eta\bar{x} - z\rangle \geqslant 0, \forall q \in \mathscr{C}^*, \langle q, \lambda\bar{x} - x\rangle = 0\}.$$

For any $q \in \mathscr{C}^*$ such that $\langle q, \lambda\bar{x} - x\rangle = 0$, by the order-preserving assumption and Proposition 2.3,

$$\langle q, \phi(\lambda\bar{x})\rangle \geqslant \langle q, \phi(x)\rangle.$$

Now, using the expression of α in (2.21),

$$\langle q, -\alpha\lambda\ln\lambda\bar{x} - \phi(x)\rangle \geqslant \langle q, -\alpha\lambda\ln\lambda\bar{x} - \phi(\lambda\bar{x})\rangle \geqslant 0.$$

This shows that

$$\Phi(x,\lambda) \in T_{\mathscr{S}_1}(x,\lambda), \ \ \forall (x,\lambda) \in \mathscr{D} \cap \mathscr{S}_1,$$

whence the invariance of $(\mathscr{S}_1 \cap \mathscr{D}, \Phi)$. For the invariance of system $(\mathscr{S}_2 \cap \mathscr{D}, \Phi)$, we define a bijection on $\mathscr{D}$:

$$F(x,\lambda) = (\lambda x - \bar{x}, \lambda)$$

whose inverse is:

$$G(y,\kappa) = (\kappa^{-1}(\bar{x} + y), \kappa).$$

If $(x(\cdot),\lambda(\cdot)) \in \mathscr{D}$ follows the dynamics of (2.27), then $(y(\cdot),\kappa(\cdot)) = F(x(\cdot),\lambda(\cdot))$ is the solution of the following differential equation:

$$\begin{pmatrix} \dot{y} \\ \dot{\kappa} \end{pmatrix} = \Psi(y,\kappa) = \begin{pmatrix} -\alpha\ln\kappa(\bar{x}+y) + \kappa\phi(\kappa^{-1}(\bar{x}+y)) \\ -\alpha\kappa\ln\kappa \end{pmatrix}. \tag{2.28}$$

Thus the invariance of system $(F(\mathscr{D}) \cap F(\mathscr{S}_2), \Psi)$ implies the invariance of system $(\mathscr{D} \cap \mathscr{S}_2, \Phi)$. Note that $F(\mathscr{S}_2) = \mathscr{C} \times \mathbb{R}$. Therefore by Theorem 2.4 the system $(F(\mathscr{D}) \cap F(\mathscr{S}_2), \Psi)$ is invariant if

$$\Psi(y,\kappa) \in T_{F(\mathscr{S}_2)}(y,\kappa), \ \ \forall (y,\kappa) \in F(\mathscr{D}) \cap F(\mathscr{S}_2).$$

The tangent cone of $F(\mathscr{S}_2)$ at point $(y,\kappa) \in F(\mathscr{S}_2)$ is given by:

$$T_{F(\mathscr{S}_2)}(y,\kappa) = \{z : \langle q, z\rangle \geqslant 0, \forall q \in \mathscr{C}^*, \langle q, y\rangle = 0\} \times \mathbb{R}.$$

Again by the order-preserving assumption, for any $q \in \mathscr{C}^*$ such that $\langle q, y\rangle = 0$,

$$\langle q, \phi(\kappa^{-1}(\bar{x}+y))\rangle \geqslant \langle q, \phi(\kappa^{-1}(\bar{x}))\rangle$$

Using again the expression of α in (2.21),

$$\langle q, \kappa\phi(\kappa^{-1}(\bar{x}))\rangle \geqslant \langle q, (\alpha\ln\kappa)\bar{x}\rangle$$

because $\kappa \in (1,\mu)$. Therefore

$$\langle q, -(\alpha \ln \kappa)(\bar{x}+y) + \kappa\phi(\kappa^{-1}(\bar{x}+y))\rangle \geqslant 0,$$

which implies

$$\Psi(y,\kappa) \in T_{F(\mathscr{S}_2)}(y,\kappa), \ \ \forall (y,\kappa) \in F(\mathscr{D}) \cap F(\mathscr{S}_2),$$

whence the invariance of $(F(\mathscr{S}_2)\cap F(\mathscr{D}),\Psi)$ and that of $(\mathscr{S}_2 \cap \mathscr{D}, \Phi)$.

Finally, if $\alpha \geqslant 0$, then the set $\mathscr{U}$ is invariant (by (2.20)). Thus $t_{\mathscr{U}}(x) = +\infty$ for all $x \in \mathscr{U}$. Since the closure $[\mu^{-1}\bar{x}, \mu\bar{x}]$ of $\mathscr{U}$ is in the interior of the cone, we conclude that the relation (2.20) holds as well for $x \in [\mu^{-1}\bar{x}, \mu\bar{x}]$. □

2.3.5 The discrete time case

For completeness, we give in this section the results analogous to Proposition 2.3 and Theorem 2.5 for discrete operators. These results are of a simpler character. In this section we consider a differentiable map $F : \mathscr{C}_0 \to \mathscr{C}_0$. The first proposition characterizes order-preserving maps, its elementary proof is left to the reader.

Proposition 2.5. *Let $\mathscr{U} \subset \mathscr{C}_0$ be any open convex set. Then F is order-preserving on $\mathscr{U}$ if and only if*

$$DF(P)\cdot Z \geqslant 0, \ \ \forall P \in \mathscr{U}, Z \in \mathscr{C}$$

Let $\mathscr{G} \subset \mathscr{C}_0$. The Lipschitz constant of F on $\mathscr{G}$, denoted by $\mathrm{Lip}(F;\mathscr{G})$, is defined as:

$$\mathrm{Lip}(F;\mathscr{G}) := \sup_{P_1,P_2\in\mathscr{G}} \frac{d_T(F(P_1),F(P_2))}{d_T(P_1,P_2)}. \tag{2.29}$$

Proposition 2.6. *Let $\mathscr{G} \subset \mathscr{C}_0$ be a set such that $t\mathscr{G} \subset \mathscr{G}$ for any $t \geqslant 1$. If F is order-preserving on $\mathscr{G}$, then*

$$\mathrm{Lip}(F;\mathscr{G}) = \inf\{\alpha : DF(P)\cdot P \preccurlyeq \alpha F(P), \ \ \forall P \in \mathscr{G}\}.$$

Proof. It suffices to prove the equivalence between the following two conditions:

(a) $d_T(F(P_1),F(P_2)) \leqslant \alpha d_T(P_1,P_2), \ \ \forall P_1,P_2 \in \mathscr{G}$

(b) $DF(P)\cdot P \preccurlyeq \alpha F(P), \ \ \forall P \in \mathscr{G}$

As was pointed out in Remark 1.9 [Nus94], if F is order-preserving, then Condition (a) is true if and only if:

$$\lambda^{-\alpha}F(\lambda P) \preccurlyeq F(P), \forall P \in \mathscr{G}, \lambda \geqslant 1.$$

Condition (b) is a necessary condition (differentiate the above inequality at $\lambda = 1$). For the sufficiency, note that the derivative of the left-hand side is

$$\lambda^{-\alpha-1}(DF(\lambda P)\cdot(\lambda P) - \alpha F(\lambda P))$$

which is always negative semidefinite given that Condition (b) is true. □

Remark 2.6. Nussbaum treated the discrete case in [Nus94], as an intermediate step before considering differential equations. Corollary 1.3 there shows that for any open subset $\mathscr{G} \subset \mathscr{C}_0$ such that for all $u, v \in \mathscr{G}$ there exists a piecewise $\mathscr{C}^1$ minimal geodesic contained in $\mathscr{G}$ (geodesic convexity assumption), the Lipschitz constant of the map F on $\mathscr{G}$ satisfies :

$$\mathrm{Lip}(F;\mathscr{G}) = \inf\{\alpha : -\alpha F(P) \preccurlyeq DF(P)\cdot Z \preccurlyeq \alpha F(P), \ \ \forall P \in \mathscr{G}, -P \preccurlyeq Z \preccurlyeq P\} \tag{2.30}$$

Thus, when the map F is order-preserving, a variant of Proposition 2.6, in which the domain $\mathscr{G}$ satisfies the previous geodesic convexity assumption can be easily obtained as a corollary of this result.

2.4 First applications and illustrations

In this section, we show that several known contraction results, which were originally obtained in [LW94] and [LL07] by means of symplectic semigroups, as well as new ones concerning the standard Riccati equation with indefinite coefficients, can be obtained readily from Theorem 2.5. The extension of these results to the generalized Riccati equation will be dealt with in Section 2.5.

2.4.1 Contraction rate of order-preserving flows on the standard positive cone

Let us consider the standard cone $\mathscr{C} = \mathbb{R}^n_+$ in $\mathscr{X} = \mathbb{R}^n$ (see Example 2.1) and an order-preserving flow $M_\cdot(\cdot)$ associated to a differentiable function $\phi : J \times \mathscr{C} \to \mathbb{R}^n$. For a subset $\mathscr{U} \subset \mathscr{C}_0$, define the best contraction rate on $\mathscr{U}$ to be the greatest value of α satisfying:

$$d_T(M^t_s(x_1), M^t_s(x_2)) \leqslant e^{-\alpha(t-s)} d_T(x_1,x_2), \forall x_1,x_2 \in \mathscr{U}, 0 \leqslant s \leqslant t < t_{\mathscr{U}}(s,x_1) \wedge t_{\mathscr{U}}(s,x_2) \tag{2.31}$$

A direct application of Theorem 2.5 is the following:

Corollary 2.7 (Compare with [Nus94, Th. 3.10])**.** *Let $\mathscr{U} \subset \mathscr{C}_0$ be an open set satisfying $\lambda \mathscr{U} \subset \mathscr{U}$ for all $\lambda \in (0,1]$. For $s \in J$ and $x \in \mathscr{U}$ define $g_i(s,x)$ by:*

$$g_i(s,x) = -x_i^{-1}[\sum_{j=1}^n \frac{\partial \phi_i}{\partial x_j}(s,x)x_j - \phi_i(s,x)] \tag{2.32}$$

then the best contraction rate on $\mathscr{U}$ is given by:

$$\alpha = \inf\{g_i(s,x) : 1 \leqslant i \leqslant n, x \in \mathscr{U}, s \in J\} \tag{2.33}$$

Nussbaum [Nus94] showed that a modification of this formula, with an absolute value enclosing each term $\frac{\partial \phi_i}{\partial x_j}(s,x)$ for $i \neq j$, holds for a not necessarily order-preserving flow. We defer a detailed comparison to Section 2.7.

2.4.2 Standard Riccati flow

One major application of the above analysis is the Riccati operator, arising from the Linear Quadratic (LQ) control problem. Let $\mathbb{E}$ be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$. The set of bounded linear operators on $\mathbb{E}$ is denoted by $\mathrm{End}(\mathbb{E})$. For $A \in \mathrm{End}(\mathbb{E})$, let A' denote the adjoint of A. The set of symmetric bounded linear operators is denoted by $\mathrm{Sym}(\mathbb{E})$. A symmetric bounded linear operator A is positive semidefinite if $\langle x, Ax \rangle \geqslant 0$ for all $x \in \mathbb{E}$. Following [LL07], let $\mathrm{Sym}^+(\mathbb{E})$ (resp. $\mathrm{Sym}^+_0(\mathbb{E})$) be the set of positive semidefinite (resp. positive semidefinite invertible) bounded symmetric linear operators of $\mathbb{E}$. Then $\mathrm{Sym}^+(\mathbb{E})$ is a convex closed pointed cone with interior $\mathrm{Sym}^+_0(\mathbb{E})$ and induces the *Loewner* order '$\preccurlyeq$' on $\mathrm{Sym}(\mathbb{E})$:

$$P \preccurlyeq Q \Longleftrightarrow Q - P \in \mathrm{Sym}^+(\mathbb{E}).$$

Then we may define Thompson's part metric on $\mathrm{Sym}^+_0(\mathbb{E})$. This is of course a special case of the definition in Section 2.2.1. Note that equipped with the operator norm, the cone $\mathrm{Sym}^+(\mathbb{E})$ is normal [Nus88]. Therefore the metric space $(\mathrm{Sym}^+_0(\mathbb{E}); d_T)$ is complete [Tho63].

Consider the Riccati differential equation defined on $\mathrm{Sym}(\mathbb{E})$:

$$\dot{P}(t) = \phi(t,P) := A(t)'P(t) + P(t)A(t) + D(t) - P(t)\Sigma(t)P(t). \tag{2.34}$$

where $A:\mathbb{R}\to \mathrm{End}(\mathbb{E})$, $D:\mathbb{R}\to \mathrm{Sym}(\mathbb{E})$, $\Sigma:\mathbb{R}\to \mathrm{Sym}(\mathbb{E})$ are assumed to be continuous and bounded maps. First we check the order-preserving property of the flow associated to (2.34) on $\mathrm{Sym}_0^+(\mathbb{E})$. The latter property is a standard result for finite-dimensional Riccati equations. For infinite-dimensional case, we found a statement of this property in [EM82, Thm 2.4] without proof. For completeness we give here a short elementary proof, following Coppel [Cop71, Prop. 6, Page 50].

Lemma 2.8. *The flow associated to* (2.34) *is order-preserving on* $\mathrm{Sym}_0^+(\mathbb{E})$.

Proof. Let $P_1(t),P_2(t)$ be two solutions of (2.34) on an interval $[0,t_0]$. Let $V(t)=P_1(t)-P_2(t)$ and $U(t)=(P_1(t)+P_2(t))/2$. Then

$$\begin{aligned}\dot V(t) &= A(t)'V(t)+V(t)A(t)-U(t)\Sigma(t)V(t)-V(t)\Sigma(t)U(t)\\ &= (A(t)-\Sigma(t)U(t))'V(t)+V(t)(A(t)-\Sigma(t)U(t)),\ \ t\in[0,t_0].\end{aligned}$$

Hence, defining $X(t)$ as the solution of $\dot X(t)=X(t)(A(t)-\Sigma(t)U(t))$ with the initial condition $X(0)=I$, we get

$$V(t)=X(t)'V(0)X(t),\ \ t\in[0,t_0].$$

Therefore if $P_1(0)\succcurlyeq P_2(0)$, then $P_1(t)\succcurlyeq P_2(t)$ for all $t\in[0,t_0]$. □

The least contraction rate of the flow on $\mathrm{Sym}_0^+(\mathbb{E})$ is the best constant α such that for all $P_1,P_2\in \mathrm{Sym}_0^+(\mathbb{E})$ and $s\geqslant 0$,

$$d_T(M_s^t(P_1),M_s^t(P_2))\leqslant e^{-\alpha(t-s)}d_T(P_1,P_2),\ \ \forall s\leqslant t<t_{\mathrm{Sym}_0^+(\mathbb{E})}(s,P_1)\wedge t_{\mathrm{Sym}_0^+(\mathbb{E})}(s,P_2). \tag{2.35}$$

An immediate consequence of Theorem 2.6 is:

Theorem 2.10. *The least contraction rate defined as in* (2.35) *satisfies:*

$$\alpha=\sup\{\beta\in\mathbb{R}: P\Sigma(t)P+D(t)\succcurlyeq\beta P,\ \ \forall t\geqslant 0, P\in\mathrm{Sym}_0^+(\mathbb{E})\} \tag{2.36}$$

Remark 2.7. Even if in the statement of Theorem 2.10 we do not require Σ and D to be positive semidefinite, the set to which applies the supremum in (2.36) is easily seen to be empty as soon as Σ or D are not positive semidefinite. Hence, the finiteness of the constant α in Theorem 2.10 does require Σ and D to be positive semidefinite and then we must have $\alpha\geqslant 0$. This shows a dichotomy: either the flow is non-expansive, or it is not uniformly Lipschitz.

Corollary 2.9 (Theorem 8.5 [LL07])**.** *We suppose that* $D(t),\Sigma(t)\in\mathrm{Sym}^+(\mathbb{E})$ *for all* $t\geqslant 0$*. Then the least contraction rate is given by:*

$$\alpha=2\inf_{t\geqslant 0}\sqrt{m((\Sigma(t)^{1/2}D(t)\Sigma(t)^{1/2})/I)}$$

Proof. The best contraction rate is given by:

$$\alpha=\sup\{\beta\geqslant 0: P\Sigma(t)P+D(t)\succcurlyeq\beta P,\ \ \forall t\geqslant 0, P\in\mathrm{Sym}_0^+(\mathbb{E})\}$$

Consider all $P=\lambda I$, then

$$\alpha\leqslant\sup\{\beta\geqslant 0:\lambda^2\Sigma(t)\succcurlyeq\beta\lambda I-D(t),\ \ \forall t\geqslant 0,\lambda>0\}$$

If $\Sigma(t) \in \mathrm{Sym}^+(\mathbb{E})$ is not invertible, then $m(\Sigma/I) = 0$. Thus

$$\alpha \leqslant \sup\{\beta \geqslant 0 : 0 \geqslant \beta\lambda + m(-D(t)/I),\ \ \forall t \geqslant 0, \lambda > 0\} = 0$$

Now suppose that $\Sigma(t) \in \mathrm{Sym}_0^+(\mathbb{E})$, $\forall t \geqslant 0$. In that case, $P\Sigma(t)P + D(t) \succcurlyeq \beta P$ if and only if

$$\begin{aligned}
&\Sigma(t)^{\frac{1}{2}} P\Sigma(t)P\Sigma(t)^{\frac{1}{2}} + \Sigma(t)^{\frac{1}{2}} D(t)\Sigma(t)^{\frac{1}{2}} - \beta\Sigma(t)^{\frac{1}{2}} P\Sigma(t)^{\frac{1}{2}} \\
&= (\Sigma(t)^{\frac{1}{2}} P\Sigma(t)^{\frac{1}{2}})^2 - \beta\Sigma(t)^{\frac{1}{2}} P\Sigma(t)^{\frac{1}{2}} + \Sigma(t)^{\frac{1}{2}} D(t)\Sigma(t)^{\frac{1}{2}} \\
&= (\Sigma(t)^{\frac{1}{2}} P\Sigma(t)^{\frac{1}{2}} - \frac{\beta}{2} I)^2 + \Sigma(t)^{\frac{1}{2}} D(t)\Sigma(t)^{\frac{1}{2}} - \frac{\beta^2}{4} I \succcurlyeq 0
\end{aligned}$$

Therefore,

$$\begin{aligned}
\alpha &= \sup\{\beta \geqslant 0 : \beta \leqslant 2\sqrt{m((\Sigma(t)^{1/2} D(t)\Sigma(t)^{1/2})/I)},\ \ \forall t \geqslant 0\} \\
&= 2 \inf_{t \geqslant 0} \sqrt{m((\Sigma(t)^{1/2} D(t)\Sigma(t)^{1/2})/I)}.
\end{aligned}$$

□

The above theorem was proved by Lawson and Lim in [LL07], Theorem 8.5, using a Birkhoff contraction formula of the fractional transformation on symmetric cones. Their approach requires the coefficients $\Sigma(t)$ and $D(t)$ to be positive semidefinite. By Remark 2.7, this condition is also necessary to the existence of a global contraction rate. However, a local contraction may occur even the coefficients are not positive semidefinite.

2.4.3 Indefinite Riccati flow

In this section, we consider the finite dimensional case when $\mathbb{E} = \mathbb{R}^n$ and $\mathrm{Sym}(\mathbb{E}) = \mathrm{S}_n$. The cone is the set of positive semidefinite matrices, i.e., $\mathrm{Sym}^+(\mathbb{E}) = \mathrm{S}_d^+$. We consider the time-independent matrix coefficients (A, D, Σ). Let

$$\Phi(P) = A'P + PA + D - P\Sigma P \ .$$

The time-invariant Riccati equation is

$$\dot{P} = \Phi(P) \ .$$

The following lemma is standard and can be proved directly by considering the corresponding linear quadratic optimal control problem.

Lemma 2.10. *If $D \succcurlyeq 0$, then S_d^+ is invariant by the Riccati flow associated to function Φ.*

Lemma 2.11. *Let $0 \preccurlyeq P_0 \preccurlyeq Q_0$. If $\Phi(P_0) \succcurlyeq 0$ and $\Phi(Q_0) \preccurlyeq 0$, then the interval $[P_0, Q_0]$ is invariant by the Riccati flow.*

Proof. It is sufficient to remark that if $P(\cdot) : [0,T] \to \mathrm{S}_d$ is a solution of

$$\dot{P}(t) = \Phi(P(t)),\ \ t \in [0,T] \ ,$$

then $P(t) - P_0 : [0,T] \to \mathrm{S}_d$ is a solution of

$$\dot{Q}(t) = \Psi(Q(t)),\ \ t \in [0,T] \ ,$$

where

$$\Psi(Q) = (A' - P_0\Sigma)Q + Q(A - \Sigma P_0) - Q\Sigma Q + \Phi(P_0) \ .$$

Hence by Lemma 2.10, if $\Phi(P_0) \succcurlyeq 0$, then the interval $[P_0, \infty)$ is invariant by the Riccati flow. The same we can prove that if $\Phi(Q_0) \preccurlyeq 0$, then the interval $(-\infty, Q_0]$ is invariant by the Riccati flow. □

The common fixed point of the flow M_t for all t must satisfy the algebraic Riccati equation(ARE) equation:

$$A'P + PA + D - P\Sigma P = 0$$

If $\Sigma, D \in \hat{\mathbf{S}}_d^+$, by Corollary 2.9 and the completeness of the metric space $(\hat{\mathbf{S}}_d^+; d_T)$ we know that the solution of ARE exists and is unique. We next give sufficient conditions for the existence of solutions of ARE even when Σ is not positive semidefinite. Below is a direct consequence of Theorem 2.5.

Corollary 2.12. *Let $P_0 \succ 0$ and $\alpha \in \mathbb{R}$. The following are equivalent:*

(a) For all $P_1, P_2 \in (0, P_0)$,

$$d_T(M_t(P_1), M_t(P_2)) \leqslant e^{-\alpha t} d_T(P_1, P_2), \ \ \forall t < t_{(0,P_0)}(P_1) \wedge t_{(0,P_0)}(P_2).$$

(b) For all $P \in (0, P_0)$,

$$D + P\Sigma P \succcurlyeq \alpha P.$$

In particular, this corollary allows to prove the local contraction property of the Riccati equation (2.34) when Σ is not positive definite. Let $c_A, c_D, m_D, c_\Sigma \in \mathbb{R}$ such that:

$$A + A' \preccurlyeq -2c_A I, \ \ m_D I \preccurlyeq D \preccurlyeq c_D I, \ \ -\Sigma \preccurlyeq c_\Sigma I.$$

The situation considered in the next corollary is motivated by the analysis of a method of reduction of the curse of dimensionality introduced by McEneaney [McE07]. This method applies to a control problem in which one can switch between several linear-quadratic models, see Chapter 6 for an introduction to the method. We will see in Chapter 7 that the next corollary is crucial to an improved convergence bound.

Corollary 2.13. *Suppose that $c_A, c_D > 0, m_D, c_\Sigma > 0$ and*

$$c_A^2 \geqslant c_D c_\Sigma, \ \ c_\Sigma m_D > (c_A - \sqrt{c_A^2 - c_D c_\Sigma})^2,$$

then for any $\lambda \in [\frac{c_A - \sqrt{(c_A^2 - c_D c_\Sigma)}}{c_\Sigma}, \sqrt{\frac{m_D}{c_\Sigma}})$, there is $\alpha \geqslant (m_D - c_\Sigma \lambda^2)/\lambda$ such that for all $P_1, P_2 \in (0, \lambda I]$

$$d_T(M_t(P_1), M_t(P_2)) \leqslant e^{-\alpha t} d_T(P_1, P_2), \quad \forall t \geqslant 0.$$

In particular, there exists a unique solution $\bar{P}$ to ARE in $(0, \lambda I]$ and for any $P \in (0, \lambda I]$,

$$d_T(M_t(P), \bar{P}) \leqslant e^{-\alpha t} d_T(P, \bar{P}), \ \ \forall t \geqslant 0.$$

Proof. Let any $\lambda \in [\frac{c_A - \sqrt{(c_A^2 - c_D c_\Sigma)}}{c_\Sigma}, \sqrt{\frac{m_D}{c_\Sigma}})$. Since

$$\phi(\lambda I) = \lambda(A + A') + D - \lambda^2 \Sigma \preccurlyeq (-2\lambda c_A + c_D + \lambda^2 c_\Sigma) I \preccurlyeq 0.$$

we deduce from Lemma 2.11 that the closed set $(0, \lambda I]$ is invariant by the Riccati flow. It is not difficult to show that given $\lambda_0 \in (\lambda, \sqrt{\frac{m_D}{c_\Sigma}})$ there is $\alpha \geqslant (m_D - c_\Sigma \lambda_0^2)/\lambda_0$ such that

$$D + P\Sigma P \succcurlyeq \alpha P, \ \ \forall P \in (0, \lambda_0 I).$$

Indeed, note that a sufficient condition would be:

$$m_D I - c_\Sigma P^2 \succcurlyeq \alpha P, \quad \forall P \in (0, \lambda_0 I)$$

which is equivalent to:

$$m_D - c_\Sigma \lambda_0^2 \geqslant \alpha \lambda_0.$$

By Corollary 2.12, for any $P_1, P_2 \in (0, \lambda I] \subset (0, \lambda_0 I)$,

$$d_T(M_t(P_1), M_t(P_2)) \leqslant e^{-\alpha t} d_T(P_1, P_2), \ \ \forall t \geqslant 0$$

Since the metric space $((0, \lambda I]; d_T)$ is complete, we deduce that there is a unique fixed point $\bar{P} \in (0, \lambda I]$ and all solutions with initial value in $(0, \lambda I]$ converge exponentially to $\bar{P}$ with rate α. □

Another interesting case is when $\Sigma \succcurlyeq 0$ not invertible. In that case, Corollary 2.9 tells that the least contraction rate on S_d^+ is 0. However, using Corollary 2.12 we can say something more about the asymptotic behavior of the trajectories.

Corollary 2.14. *Suppose that $c_A > 0, m_D > 0$ and $c_\Sigma = 0$. Then for any $\lambda \geqslant \frac{c_D}{c_A}$, there is $\alpha > 0$ such that the flow is α-contractive on the set $(0, \lambda I]$. In particular, the existence and uniqueness of solution $\bar{P} \in \hat{\mathrm{S}}_d^+$ to ARE is insured and for any $P \in \hat{\mathrm{S}}_d^+$,*

$$d_T(M_t(P), \bar{P}) \leqslant e^{-\alpha t} d_T(P, \bar{P}), \ \ \forall t \geqslant 0.$$

where $\alpha = \min(m(I/P), \frac{c_A}{c_D}) m_D$.

We leave the proof to the reader, which is similar to the one of Corollary 2.13.

2.5 Application to stochastic Riccati differential equations

In the sequel, we apply the previous results to the cone of positive semidefinite matrices, i.e. $\mathscr{X} = \mathrm{S}_n$, $\mathscr{C} = \mathrm{S}_n^+$ and $\mathscr{C}_0 = \hat{\mathrm{S}}_n^+$ (see Example 2.2). Note that here $\mathscr{C}^\star = \mathscr{C}$. We shall use the notation $\succcurlyeq$ (and $\succ$) for the (strict) Loewner order, and d_T for Thompson's part metric induced by S_n^+ (see Section 2.2.1).

2.5.1 Stochastic LQ problem and GRDE

Consider the following stochastic linear quadratic optimal control problem:

$$\begin{aligned} v(s,y) &= \min_{u(\cdot)} \mathbb{E} \int_s^T [x(t)' Q(t) x(t) + 2u'(t) L(t) x(t) + u(t)' R(t) u(t)] dt + \mathbb{E}[x(T)' G x(T)] \\ \text{s.t.} \quad &\begin{cases} dx(t) = (A(t)x(t) + B(t)u(t))dt + (C(t)x(t) + D(t)u(t))dW(t), & \forall t \in [s,T], \\ x(s) = y. \end{cases} \end{aligned}$$

where the functions appearing above satisfy:

$$\begin{cases} A(\cdot), C(\cdot) \in L^\infty \cap C^0(0,T; \mathbb{R}^{n \times n}), \quad B(\cdot), D(\cdot), L(\cdot) \in L^\infty \cap C^0(0,T; \mathbb{R}^{n \times k}), \\ Q(\cdot) \in L^\infty \cap C^0(0,T; \mathrm{S}_n), \quad R(\cdot) \in L^\infty \cap C^0(0,T; \mathrm{S}^k). \end{cases}$$

Here W is a standard Brownian motion defined on a complete probability space. We refer the reader to [YZ99] Chapter 6, for the precise definition of this control problem. In [YZ99], the above functions are only assumed to be bounded. In our case, the continuity is necessary to apply the previous results.

The above stochastic LQ control problem over the time interval $[s,T]$ is solvable, i.e., admits an optimal control for all $y \in \mathbb{R}^n$ if the solution of the following constrained differential matrix equation exists:

$$\begin{cases} \dot{P}+A'P+PA+C'PC+Q= \\ \qquad (PB+C'PD+L')(R+D'PD)^{-1}(B'P+D'PC+L), \qquad t\in[s,T] \\ P(T)=G \\ R(t)+D(t)'P(t)D(t)\succ 0, \qquad t\in[s,T] \end{cases} \tag{2.37}$$

which we refer to as generalized Riccati differential equation (GRDE). In that case, the value function of the optimal control problem is given by

$$v(s,y)=y'P(s)y. \tag{2.38}$$

2.5.2 GRDE with semidefinite weighting matrices

The solvability of the GRDE (2.37) with indefinite matrix coefficients has been treated by Chen, Moore, Ait Rami, and Zhou in [RCMZ01]. In order to apply our previous results, we only consider the case:

$$\begin{pmatrix} Q(t) & L(t)' \\ L(t) & R(t) \end{pmatrix} \succcurlyeq 0, \;\; \ker R(t)\cap\ker D(t)=\{0\}, \;\; \forall t\in[0,T], \tag{2.39}$$

so that the function

$$\begin{aligned} \phi(t,P)= \;& PA+A'P+C'PC+Q- \\ & (B'P+D'PC+L)'(R+D'PD)^{-1}(B'P+D'PC+L) \end{aligned} \tag{2.40}$$

is well defined on $[0,+\infty)\times\hat{\mathrm{S}}_n^+$ and satisfies the assumptions made at the beginning of section 2.3. We are going to apply the preceding results to show the monotonicity and the non-expansiveness of the GRDE differential equation defined on $\hat{\mathrm{S}}_n^+$:

$$\begin{cases} \dot{P}=\phi(t,P), \\ P(0)=G \end{cases} \tag{2.41}$$

Proposition 2.15. *Assume that* (2.39) *holds. Then the flow associated to (2.41) is order-preserving and non-expansive on* $\hat{\mathrm{S}}_n^+$.

Proposition 2.15 could be derived by exploiting the relation between the solution of the Riccati equation and the value function of the stochastic control problem (see (2.38)). Here we choose to prove it from the infinitesimal characterizations of Proposition 2.3 and Theorem 2.6.

Proof. By Proposition 2.3, if suffices to prove that for any $P\in\hat{\mathrm{S}}_n^+$, any $Q,Z\in\mathrm{S}_n^+$ such that $\langle Q,Z\rangle=0$:

$$\langle Q, D\phi_t(P)Z\rangle \geqslant 0.$$

Indeed,

$$\begin{aligned} D\phi_t(P)Z= \;& ZA(t)+A(t)'Z+C(t)'ZC(t)-(B(t)'Z+D(t)'ZC(t))'N_t(P) \\ & -N_t(P)'(B(t)'Z+D(t)'ZC(t))+N_t(P)'D(t)'ZD(t)N_t(P) \end{aligned}$$

where $N_t(P) = (R(t) + D(t)'PD(t))^{-1}(B(t)'P + D(t)'PC(t) + L(t))$. Remark that if $Q, Z \in \mathrm{S}_n^+$ and $\langle Q, Z\rangle = 0$ then $QZ = 0$. Therefore,

$$\begin{aligned}\langle Q, D\phi_t(P)Z\rangle &= \langle Q, C(t)'ZC(t) - C(t)'ZD(t)N_t(P) - N_t(P)'D(t)'ZC(t) \\ &\quad + N_t(P)'D(t)'ZD(t)N_t(P)\rangle \\ &= \langle Q, \big(C(t) - D(t)N_t(P)\big)'Z\big(C(t) - D(t)N_t(P)\big)\rangle \geqslant 0.\end{aligned}$$

Now for non-expansiveness, by Theorem 2.6 it remains to verify that for any $P \in \hat{\mathrm{S}}_n^+$ and any $t \in [0, T]$,

$$D\phi_t(P)P - \phi(t, P) \preccurlyeq 0.$$

Indeed,

$$\begin{aligned}&D\phi_t(P)P - \phi(t, P) \\ &= -Q(t) + N_t(P)'L(t) + L'(t)N_t(P) - N_t(P)'R(t)N_t(P) \\ &= H_t(P)' \begin{pmatrix} -Q(t) & -L(t)' \\ -L(t) & -R(t) \end{pmatrix} H_t(P) \preccurlyeq 0\end{aligned} \tag{2.42}$$

where $H_t(P)' = \begin{pmatrix} I & -N_t(P)' \end{pmatrix}$. □

Remark 2.8. A fundamental discrepancy with the standard Riccati equation is that the flow of the generalized Riccati equation is not a global contraction. This is because there is no $\alpha > 0$ such that the condition

$$D\phi_t(P)P - \phi(t, P) \preccurlyeq -\alpha P, \ \ \forall P \in \hat{\mathrm{S}}_n^+,$$

which by Theorem 2.6 is necessary to the global contraction property of the flow, is satisfied. However, we shall see in the next section that a local contraction property does hold.

2.5.3 Asymptotic behavior of GRDE

We are going to investigate the behavior of the GRDE flow as time horizon goes to infinity. All the matrices A, B, C, D, L, Q, R are assumed to be constant. First we show a local contraction property under the condition

$$\begin{pmatrix} Q & L' \\ L & R \end{pmatrix} \succ 0. \tag{2.43}$$

More precisely,

Theorem 2.11. *Assume that* (2.43) *holds. Let $\mathscr{U} \subset \hat{\mathrm{S}}_n^+$ be an open set such that $\lambda\mathscr{U} \subset \mathscr{U}$ for all $\lambda \in (0, 1]$. Assume that there is $P_0 \in \hat{\mathrm{S}}_n^+$ such that $\mathscr{U} \subset (0, P_0]$ and let $\alpha = m(Q - L'R^{-1}L/P_0)$, then for all $P_1, P_2 \in \mathscr{U}$,*

$$d_T(M_t(P_1), M_t(P_2)) \leqslant e^{-\alpha t} d_T(P_1, P_2), \ \ 0 \leqslant t < t_{\mathscr{U}}(P_1) \wedge t_{\mathscr{U}}(P_2)$$

Proof. By applying Theorem 2.5, we need to prove

$$D\phi(P)P - \phi(P) \preccurlyeq -\alpha P, \quad \forall P \in \mathscr{U}$$

Indeed, for all $P \in \mathscr{U}$,

$$Q - \alpha P - L'R^{-1}L \succcurlyeq Q - \alpha P_0 - L'R^{-1}L \succcurlyeq 0.$$

Besides, the previous calculation yields

$$\begin{aligned} &D\phi(P)P-\phi(P)+\alpha P \\ &= H(P)' \begin{pmatrix} -Q+\alpha P & -L' \\ -L & -R \end{pmatrix} H(P) \end{aligned} \tag{2.44}$$

where $H(P)' = \begin{pmatrix} I & -N(P)' \end{pmatrix}$ and $N(P) = (R+D'PD)^{-1}(B'P+D'PC+L)$. By Schur's complement lemma [BTEGN09, Lemma 6.3.4], we get

$$D\phi(P)P-\phi(P) \preccurlyeq -\alpha P, \quad \forall P \in \mathscr{U}.$$

□

The fixed point of the GRDE flow associated to (2.41), if it exists, satisfies the so-called general algebraic Riccati equation (GARE):

$$\begin{cases} \phi(P) = 0. \\ R+D'PD \succ 0 \end{cases} \tag{2.45}$$

where $\phi(P) := A'P+PA+C'PC+Q-(B'P+D'PC+L)'(R+D'PD)^{-1}(B'P+D'PC+L)$. The existence of solutions of GARE and the asymptotic behavior of the GRDE flow have been studied in [RCMZ01] and [RZ00]. The authors assumed the following mean-square stabilizability condition:

Definition 2.5 (Definition 4.1 [RZ00])**.** The system of matrices (A,B,C,D) is said to be mean-square stabilizable if there exists a control law of feedback form

$$u(t) = Kx(t),$$

where K is a constant matrix, such that for every initial (t_0,x_0), the closed loop system

$$\begin{cases} dx(t) = (A+BK)x(t)dt + (C+DK)x(t)dW(t) \\ x(0) = x_0 \end{cases}$$

satisfies

$$\lim_{t\to+\infty} \mathbb{E}[x(t)'x(t)] = 0$$

Under the mean-square stabilizability assumption, they established a necessary and sufficient condition for the existence of a solution. To make a comparison, let us first quote their theorem:

Theorem 2.12 (Theorem 4.1 [RCMZ01])**.** *Under the mean-square stabilizability assumption, there exists a solution of the GARE (2.45) if and only if there exists $P_0 \in \mathrm{S}_n$ such that*

$$\phi(P_0) \succcurlyeq 0, \;\; R+D'P_0D \succcurlyeq 0 \;\; .$$

Moreover, for any such P_0, the solution $P(t)$ of (2.41) with initial condition $P(0) = P_0$ converges to a solution to the GARE as $t \to \infty$.

It follows directly from the above theorem that under the mean-square stabilizability assumption, if (2.43) is true, then there must be a solution to the GARE (2.45). We next show a necessary and sufficient condition for the existence of a stable solution without the mean-square stabilizability assumption.

Theorem 2.13. *Assume that the condition* (2.43) *holds. Then, the GARE admits a solution* $\bar{P} \in \hat{\mathrm{S}}_n^+$ *if and only if there exists* $P_0 \in \hat{\mathrm{S}}_n^+$ *such that:*

$$\phi(P_0) \preccurlyeq 0. \tag{2.46}$$

In that case, for any $P \in \hat{\mathrm{S}}_n^+$*:*

$$d_T(M_t(P), \bar{P}) \leqslant e^{-\alpha t} d_T(P, \bar{P}), \ \ \forall t \geqslant 0,$$

where

$$\alpha \geqslant \frac{1 - e^{-d_T(P,\bar{P})}}{d_T(P,\bar{P})} m((Q - L'R^{-1}L)/\bar{P}) > 0. \tag{2.47}$$

In particular, the solution is unique in $\hat{\mathrm{S}}_n^+$.

Proof. If $\bar{P} \in \hat{\mathrm{S}}_n^+$ is a solution of the GARE, then (2.46) is satisfied by considering $P_0 = \bar{P}$. Conversely, note that if $\phi(P_0) \preccurlyeq 0$ for some $P_0 \in \hat{\mathrm{S}}_n^+$, then $(0, P_0]$ is an invariant set. Consider the open set $\mathscr{U} = (0, P_0 + I)$. By Theorem 2.11, there is $\alpha > 0$ such that for all $P_1, P_2 \in (0, P_0] \subset \mathscr{U}$, we have:

$$d_T(M_t(P_1), M_t(P_2)) \leqslant e^{-\alpha t} d_T(P_1, P_2), \ \ \forall 0 \leqslant t \leqslant t_{\mathscr{U}}(P_1) \wedge t_{\mathscr{U}}(P_2),$$

Since $[0, P_0] \subset \mathscr{U}$ is invariant, we have that $t_{\mathscr{U}}(P_1), t_{\mathscr{U}}(P_2) = +\infty$. Thus the flow M_t is contractive in the complete metric space $((0, P_0], d_T)$. There must be a unique fixed point $\bar{P} \in (0, P_0]$ such that $\phi(\bar{P}) = 0$. Next, assuming the existence of a solution $\bar{P} \in \mathscr{C}_0$ to the GARE, we apply Theorem 2.9 to obtain the rate of convergence. A basic calculation yields:

$$\begin{aligned}
&\lambda^{-1}\phi(\lambda\bar{P}) \\
&= (B'\bar{P} + D'\bar{P}\tilde{C})'((R + D'\bar{P}D)^{-1} - (\lambda^{-1}R + D'\bar{P}D)^{-1})(B'\bar{P} + D'\bar{P}\tilde{C}) + (\lambda^{-1} - 1)\tilde{Q}
\end{aligned}$$

where $\tilde{C} = C - DR^{-1}L$ and $\tilde{Q} = Q - L'R^{-1}L$. Therefore, if $\lambda \geqslant 1$, then

$$\lambda^{-1}\phi(\lambda\bar{P}) \preccurlyeq (\lambda^{-1} - 1)\tilde{Q} \tag{2.48}$$

and

$$\lambda\phi(\lambda^{-1}\bar{P}) \succcurlyeq (\lambda - 1)\tilde{Q}. \tag{2.49}$$

Now for any $P \in \hat{\mathrm{S}}_n^+ \neq \bar{P}$, let $\mu = e^{d_T(P,\bar{P})}$ and $\alpha = \frac{1-\mu^{-1}}{\ln\mu} m(\tilde{Q}/\bar{P}) > 0$. Then

$$(\lambda^{-1} - 1)\tilde{Q} \preccurlyeq -\alpha(\ln\lambda)\bar{P}, (\lambda - 1)\tilde{Q} \succcurlyeq \alpha(\ln\lambda)\bar{P}, \ \ \forall \lambda \in (1, \mu)$$

and (2.48) and (2.49) lead to:

$$\alpha\ln(\lambda)\bar{P} \preccurlyeq \lambda\phi(\lambda^{-1}\bar{P}), \ \ \alpha\ln(\lambda)\bar{P} \preccurlyeq -\lambda^{-1}\phi(\lambda P), \ \ \forall \lambda \in (1, \mu).$$

Thus,

$$0 < \alpha \leqslant \inf_{\mu^{-1} < \lambda < \mu} m\big((-(\lambda\ln\lambda)^{-1}\phi(\lambda\bar{P}))/\bar{P}\big) \ . \tag{2.50}$$

By virtue of (2.50) and Theorem 2.9, we have

$$d_T(M_t(P), \bar{P}) \leqslant e^{-\alpha t} d_T(P, \bar{P}), \ \ \forall t \geqslant 0.$$

□

2.5.4 Discrete Generalized Riccati operator

The linear quadratic stochastic control problem has a discrete time analogue [ARCZ01], which leads to the generalized discrete Riccati operator $F : \mathrm{S}_n \to \mathrm{S}_n$:

$$F(P) = A'PA + C'PC + Q - (B'PA + D'PC)'(R + B'PB + D'PD)^{-1}(B'PA + D'PC) \tag{2.51}$$

where $A, C \in \mathbb{R}^{n\times n}$, $B, D \in \mathbb{R}^{n\times m}$ and $Q, R \in \mathrm{S}_n$. We assume that $Q \succ 0$ and $R \succ 0$. Then by applying the Schur complement condition for positive definiteness, one can prove that F sends $\hat{\mathrm{S}}_n^+$ to itself. Note that when $C = D = 0$, we recover the standard Riccati operator:

$$T(P) = A'PA + Q - A'PB(R + B'PB)^{-1}B'PA. \tag{2.52}$$

The object of this section is to get the Lipschitz constant of F on $\hat{\mathrm{S}}_n^+$ (see (2.29)). First we show that this operator is order-preserving on $\hat{\mathrm{S}}_n^+$.

Proposition 2.16. *The operator F is order-preserving on $\hat{\mathrm{S}}_n^+$.*

Proof. Let any $P \in \hat{\mathrm{S}}_n^+$ and $Z \in \mathrm{S}_n^+$. A simple calculation show that:

$$DF(P)\cdot Z = (A - BN)'Z(A - BN) + (C - DN)'Z(C - DN) \succcurlyeq 0$$

where $N = (R + B'PB + D'PD)^{-1}(B'PA + D'PC)$. By Proposition 2.3, F is order-preserving on $\hat{\mathrm{S}}_n^+$. □

Next we apply Proposition 2.6 to get:

$$\mathrm{Lip}(F; \hat{\mathrm{S}}_n^+) = \inf\{\alpha \geqslant 0 : DF(P)\cdot P \preccurlyeq \alpha F(P), \forall P \in \hat{\mathrm{S}}_n^+\}. \tag{2.53}$$

The following two lemmas will be useful.

Lemma 2.17. *Let $\left(\begin{smallmatrix} B \\ D \end{smallmatrix}\right) = \left(\begin{smallmatrix} \bar{B} \\ \bar{D} \end{smallmatrix}\right) w$ be a rank factorization (so that the last two factors have maximal column and row rank, respectively). Then the operator F defined in (2.51) satisfies:*

$$F(P) = A'PA + C'PC + Q - (\bar{B}'PA + \bar{D}'PC)'(\bar{R} + \bar{B}'P\bar{B} + \bar{D}'P\bar{D})^{-1}(\bar{B}'PA + \bar{D}'PC) \tag{2.54}$$

where $\bar{R} = (WR^{-1}W')^{-1}$.

Proof. To simplify the notation, denote $X(P) = \bar{B}P\bar{B} + \bar{D}'P\bar{D}$. Notice that since the matrix $\left(\begin{smallmatrix} \bar{B} \\ \bar{D} \end{smallmatrix}\right)$ is of full column rank, $X(P)$ is invertible for all $P \in \hat{\mathrm{S}}_n^+$. It follows from (2.51) that:

$$F(P) = A'PA + C'PC + Q - (\bar{B}'PA + \bar{D}'PC)'W(R + W'X(P)W)^{-1}W'(\bar{B}'PA + \bar{D}'PC)$$

Now appealing to the Woodbury matrix identity [Mey00, Sec 3.8], we obtain:

$$\begin{aligned} W(R + X(P)W)^{-1}W' &= W(R^{-1} - R^{-1}W'(X(P)^{-1} + WR^{-1}W')^{-1}WR^{-1})W' \\ &= WR^{-1}W' - WR^{-1}W'(X(P)^{-1} + WR^{-1}W')^{-1}WR^{-1}W' \\ &= ((WR^{-1}W')^{-1} + X(P))^{-1} \end{aligned} \tag{2.55}$$

from which we get (2.54). □

Lemma 2.18. *Let $\delta \geqslant 2$, then*

$$X - X(R + X)^{-1}(\delta R + X)(R + X)^{-1}X \preccurlyeq \frac{R}{4(\delta - 1)}, \quad \forall X \in \mathrm{S}_n^+ \tag{2.56}$$

Proof. Let any $X \in \mathrm{S}_n^+$. Since X commutes with I, we have that:

$$\begin{aligned}
&X - X(I+X)^{-1}(\delta I + X)(I+X)^{-1}X \\
&= (I+X)^{-1}(X(I+X)^2 - X^2(\delta I + X))(I+X)^{-1} \\
&= (I+X)^{-1}((2-\delta)X^2 + X - \tfrac{1}{4(\delta-1)}(I+X)^2)(I+X)^{-1} + \tfrac{1}{4(\delta-1)}I \\
&= -(I+X)^{-1}((2\delta-3)X - I)^2(I+X)^{-1} + \tfrac{1}{4(\delta-1)}I \\
&\preccurlyeq \tfrac{1}{4(\delta-1)}I.
\end{aligned}$$

To obtain (2.56), it suffices to notice that:

$$\begin{aligned}
&R^{-\frac{1}{2}}(X - X(R+X)^{-1}(\delta R + X)(R+X)^{-1}X)R^{-\frac{1}{2}} \\
&= Y - Y(I+Y)^{-1}(\delta I + Y)(I+Y)^{-1}Y
\end{aligned}$$

where $Y = R^{-\frac{1}{2}}XR^{-\frac{1}{2}}$. □

Proposition 2.19. *The operator F is non-expansive:* $\mathrm{Lip}(F;\hat{\mathrm{S}}_n^+) \leqslant 1$. *Let*

$$\begin{pmatrix} B \\ D \end{pmatrix} = \begin{pmatrix} \bar{B} \\ \bar{D} \end{pmatrix} W$$

be a rank factorization. Then a necessary and sufficient condition to have $\mathrm{Lip}(F;\hat{\mathrm{S}}_n^+) < 1$ *is that there is a matrix S such that:*

$$\begin{pmatrix} A \\ C \end{pmatrix} = \begin{pmatrix} \bar{B} \\ \bar{D} \end{pmatrix} S. \tag{2.57}$$

In that case,

$$\mathrm{Lip}(F;\hat{\mathrm{S}}_n^+) \leqslant \frac{M(S'\bar{R}S/Q)}{(1+\sqrt{1+M(S'\bar{R}S/Q)})^2} < 1$$

where $\bar{R} = (WR^{-1}W')^{-1}$.

Proof. Lemma 2.17 implies that it is sufficient to prove the proposition for the case $W = I$, i.e. when $\binom{B}{D}$ is of full column rank. A simple calculation shows that:

$$\begin{aligned}
DF(P)\cdot P - \alpha F(P) = \quad & (1-\alpha)(A'PA + C'PC) - \alpha Q \\
& -(1-\alpha)N(P)'(R+X(P))^{-1}N(P) \\
& -N(P)'(R+X(P))^{-1}R(R+X(P))^{-1}N(P)
\end{aligned}$$

where

$$N(P) = B'PA + D'PC, \quad X(P) = B'PB + D'PD.$$

Then it is evident that $\mathrm{Lip}(F;\hat{\mathrm{S}}_n^+) \leqslant 1$. Now let $S \in \mathbb{R}^{n\times m}$ such that (2.57) holds. Then $N(P) = X(P)S$, $A'PA + C'PC = S'X(P)S$ and

$$\begin{aligned}
DF(P)\cdot P - \alpha F(P) = \quad & (1-\alpha)S'X(P)S - \alpha Q \\
& -(1-\alpha)S'X(P)(R+X(P))^{-1}X(P)S \\
& -S'X(P)'(R+X(P))^{-1}R(R+X(P))^{-1}X(P)S
\end{aligned}$$

To simplify the notation, let $X := X(P)$ and $\delta := \frac{2-\alpha}{1-\alpha}$, then

$$DF(P)\cdot P - \alpha F(P) = (1-\alpha)S'(X - X(R+X)^{-1}(\delta R + X)(R+X)^{-1}X)S - \alpha Q.$$

By Lemma 2.18:

$$X - X(R+X)^{-1}(\delta R+X)(R+X)^{-1}X \preccurlyeq \tfrac{1}{4(\delta-1)}R = \tfrac{1-\alpha}{4}R, \ \ \forall X \in \mathrm{S}_n^+ .$$

Therefore

$$(1-\alpha)S'(X - X(R+X)^{-1}(\delta R+X)(R+X)^{-1}X)S \preccurlyeq \frac{(1-\alpha)^2}{4}S'RS.$$

Consequently if α is such that: $\frac{4\alpha}{(1-\alpha)^2} = M(S'RS/Q)$, then

$$DF(P)\cdot P - \alpha F(P) \preccurlyeq 0, \ \ \forall P \in \mathrm{S}_n^+ .$$

Together with (2.53) this shows that

$$\mathrm{Lip}(F;\hat{\mathrm{S}}_n^+) \leqslant \frac{M(S'RS/Q)}{(1+\sqrt{1+M(S'RS/Q)})^2}.$$

Next we prove the necessity of condition (2.57). Remember that since the matrix $\binom{B}{D}$ has full rank, $X(P)$ is always invertible for $P \in \hat{\mathrm{S}}_n^+$. Besides, there is $\alpha < 1$ such that

$$DF(P)\cdot P - \alpha F(P) \preccurlyeq 0, \ \ \forall P \in \hat{\mathrm{S}}_n^+$$

if and only if for any $P \in \hat{\mathrm{S}}_n^+$,

$$(A'PA + C'PC) - \tfrac{\alpha}{1-\alpha}Q - N(P)'(R+X(P))^{-1}(\tfrac{2-\alpha}{1-\alpha}R + X(P))(R+X(P))^{-1}N(P) \preccurlyeq 0.$$

That is, for any $P \in \hat{\mathrm{S}}_n^+$ and $\lambda > 0$,

$$(A'PA + C'PC) - \tfrac{\alpha\lambda^{-1}}{1-\alpha}Q - N(P)'(\tfrac{1}{\lambda}R + X(P))^{-1}(\tfrac{2-\alpha}{\lambda(1-\alpha)}R + X(P))(\tfrac{1}{\lambda}R + X(P))^{-1}N(P) \preccurlyeq 0.$$

Letting λ go to infinity, by continuity, we obtain that:

$$(A'PA + C'PC) - N(P)'X(P))^{-1}N(P) \preccurlyeq 0.$$

The above expression is the Schur complement of the positive semidefinite matrix

$$\begin{pmatrix} B'PB+D'PD & B'PA+D'PC \\ A'PB+C'PD & A'PA+C'PC \end{pmatrix} = \begin{pmatrix} B' \\ A' \end{pmatrix} P \begin{pmatrix} B & A \end{pmatrix} + \begin{pmatrix} D' \\ C' \end{pmatrix} P \begin{pmatrix} D & C \end{pmatrix}.$$

Therefore for any $x \in \mathbb{R}^n$ there is $u \in \mathbb{R}^m$ such that

$$\left\langle \begin{pmatrix} u \\ x \end{pmatrix}, \begin{pmatrix} B'PB+D'PD & B'PA+D'PC \\ A'PB+C'PD & A'PA+C'PC \end{pmatrix} \begin{pmatrix} u \\ x \end{pmatrix} \right\rangle = 0. \tag{2.58}$$

That is, for any $x \in \mathbb{R}^n$ there is $u \in \mathbb{R}^m$ such that:

$$\begin{pmatrix} B & A \\ D & C \end{pmatrix} \begin{pmatrix} u \\ x \end{pmatrix} = 0.$$

This is equivalent to say that there is $S \in \mathbb{R}^{m\times n}$ such that:

$$\begin{pmatrix} A \\ C \end{pmatrix} = \begin{pmatrix} BS \\ DS \end{pmatrix}.$$

□

The contraction rate of the standard discrete Riccati operator $T : \hat{\mathrm{S}}_n^+ \to \hat{\mathrm{S}}_n^+$ can now be recovered as a corollary:

Corollary 2.20 (Compare with [LL08])**.** *The standard Riccati operator T defined in* (2.52) *is non-expansive:* $\mathrm{Lip}(T;\hat{\mathrm{S}}_n^+) \leqslant 1$. *A necessary and sufficient condition to have the strict contraction property is that the matrix B is of full row rank. In that case, let $B = \bar{B}W$ be a rank factorization, then*

$$\mathrm{Lip}(T;\hat{\mathrm{S}}_n^+) \leqslant \frac{M(S'\bar{R}S/Q)}{(1+\sqrt{1+M(S'\bar{R}S/Q)})^2} < 1$$

where $S = \bar{B}^{-1}A$ and $\bar{R} = (WR^{-1}W')^{-1}$.

Remark 2.9. Condition (2.57) leads to a formal argument explaining why strict global contraction cannot be hoped for the GRDE flow. Indeed, we can approximate the continuous-time LQ control problem in Section 2.5 over a small time horizon ε by the following one-step discrete-time stochastic linear quadratic control problem:

$$\begin{aligned} &\min_{u\in\mathbb{R}^m} \mathbb{E}(\langle x_0, \varepsilon Q x_0\rangle + \langle u, \varepsilon R u\rangle + \langle x_\varepsilon, G x_\varepsilon\rangle) \\ &\quad s.t.\ \ x_\varepsilon = (I+\varepsilon A)x_0 + \varepsilon B u + (\sqrt{\varepsilon}Cx_0 + \sqrt{\varepsilon}Du)w \end{aligned}$$

where $w \sim \mathcal{N}(0,1)$. Without loss of generality, we suppose that $\binom{B}{D}$ is of full column rank. If a strict contraction result was valid for the continuous time system, we would expect the same to be true for its discrete approximation if ε is sufficiently small. However, the strict global contraction condition requires the existence of S such that:

$$\begin{pmatrix} I+\varepsilon A \\ \sqrt{\varepsilon}C \end{pmatrix} = \begin{pmatrix} \varepsilon BS \\ \sqrt{\varepsilon}DS \end{pmatrix},$$

which can not hold for a set of ε converging to 0 if C and D are not zero.

2.6 Loss of non-expansiveness of the GRDE flow in other invariant Finsler metrics

The standard Riccati flow is known to be a contraction in the standard Riemannian metric [Bou93], and more generally in any invariant Finsler metric (with the same bound on the contraction rate) [LL08]. We next construct an explicit counter example showing that Thompson's part metric is essentially the only invariant Finsler metric in which the GRDE Riccati flow is non-expansive.

2.6.1 Preliminary results

We first recall the definition of symmetric gauge functions and of the associated invariant Finsler metrics on the interior of the cone of positive definite matrices. Then, we will show some conditions that are necessary for an order-preserving flow to be non-expansive in a given metric of this kind.

Definition 2.6 (Symmetric gauge function)**.** A symmetric gauge function $\nu : \mathbb{R}^n \to \mathbb{R}$ is a convex, positively homogeneous of degree 1 function such that for any permutation σ,

$$\nu(\lambda_1,\cdots,\lambda_n) = \nu(|\lambda_{\sigma(1)}|,\cdots,|\lambda_{\sigma(n)}|),\ \ \forall\ \ \lambda = (\lambda_1,\cdots,\lambda_n) \in \mathbb{R}^n.$$

The next lemma collects several useful properties of subdifferentials of symmetric gauge function (see [Roc70] for more background on subdifferentials). The straightforward proof is left to the reader.

Lemma 2.21. *Let $\nu : \mathbb{R}^n \to \mathbb{R}$ be a symmetric gauge function. The following properties hold:*

1 For all $\lambda \in \mathbb{R}^n$ and $\mu \in \partial\nu(\lambda)$,

$$\mu_i\lambda_i \geqslant 0, \quad \forall i = 1, \cdots, n.$$

2 For all $\lambda \in \mathbb{R}^n$ and $\mu \in \partial\nu(\lambda)$,

$$\langle \mu, \lambda \rangle = \nu(\lambda).$$

3 For all $\lambda, \lambda' \in \mathbb{R}^n$ and $\mu \in \partial\nu(\lambda), \mu' \in \partial\nu(\lambda')$,

$$\langle \mu - \mu', \lambda \rangle \geqslant 0.$$

For every symmetric gauge function ν, we define a spectral function $\hat{\nu} : \mathrm{S}_n \to \mathbb{R}$:

$$\hat{\nu}(P) = \nu(\lambda(P)).$$

where $\lambda(P)$ is the vector of eigenvalues of P.

Theorem 2.14 ([Lew96])**.** *If ν is a symmetric gauge function, then $\hat{\nu}$ is a convex function on S_n. Moreover, $Z \in \partial\hat{\nu}(P)$ if and only if there exists $y \in \partial\nu(\lambda(P))$ such that:*

$$Z = V \operatorname{diag}(y) V^T,$$

where V is the unitary matrix such that $P = V \operatorname{diag}(\lambda(P)) V^T$.

Following [Bha03], [LL08] and [ACS00], we define a metric on $\hat{\mathrm{S}}_n^+$ as follows,

$$d_\nu(P, Q) = \hat{\nu}(\log(P^{-1/2} Q P^{-1/2})).$$

It coincides with the Finsler metric obtained by thinking of $\hat{\mathrm{S}}_n^+$ as a manifold and taking

$$\|dQ\|_P = \hat{\nu}(P^{-\frac{1}{2}}(dQ)P^{-\frac{1}{2}})$$

as the length of an infinitesimal displacement in the tangent space at point P. This metric is invariant by the canonical action on the linear group on $\hat{\mathrm{S}}_n^+$.

We shall consider specially, as in [ACS00], the p-norm function:

$$\nu(\lambda) = \|\lambda\|_p = (\sum_{i=1}^n |\lambda_i|^p)^{1/p},$$

so that the metric d_ν is Thompson's part metric for $p = +\infty$ and the Riemannian metric for $p = 2$.

Lemma 2.22. *Let ν be a symmetric gauge function and d_ν be the associated metric on $\hat{\mathrm{S}}_n^+$. Let $M : \mathrm{S}_n^+ \to \mathrm{S}_n^+$ be a differentiable function such that:*

$$d_\nu(M(P), M(Q)) \leqslant d_\nu(P, Q), \ \ \forall \ P, Q \in \hat{\mathrm{S}}_n^+, \tag{2.59}$$

then

$$\hat{\nu}(M(P)^{-1/2}(DM(P) \cdot Z)M(P)^{-1/2}) \leqslant \hat{\nu}(P^{-1/2} Z P^{-1/2}), \ \ \forall \ P \in \hat{\mathrm{S}}_n^+, Z \in \mathrm{S}_n\,.$$

Proof. Let any $P \in \hat{\mathrm{S}}_n^+$ and $Z \in \mathrm{S}_n$. There exists $\delta > 0$ such that for any $0 \leqslant \varepsilon \leqslant \delta$, $P + \varepsilon Z \in \hat{\mathrm{S}}_n^+$. By (2.59) and the definition of d_ν:

$$\hat{\nu}\log(M(P)^{-1/2}M(P+\varepsilon Z)M(P)^{-1/2}) \leqslant \hat{\nu}\log(P^{-1/2}(P+\varepsilon Z)P^{-1/2}).$$

Divide the two sides by ε and take the limit:

$$\lim_{\varepsilon\to 0}\frac{\hat{\nu}\log(M(P)^{-1/2}M(P+\varepsilon Z)M(P)^{-1/2})}{\varepsilon} \leqslant \lim_{\varepsilon\to 0}\frac{\hat{\nu}\log(P^{-1/2}(P+\varepsilon Z)P^{-1/2})}{\varepsilon}$$

In view of homogeneity and continuity of the function $\hat{\nu}$,

$$\hat{\nu}(\lim_{\varepsilon\to 0}\frac{\log(M(P)^{-1/2}M(P+\varepsilon Z)M(P)^{-1/2})}{\varepsilon}) \leqslant \hat{\nu}(\lim_{\varepsilon\to 0}\frac{\log P^{-1/2}(P+\varepsilon Z)P^{-1/2}}{\varepsilon})$$

The matrix function log is differentiable at I:

$$\lim_{\|U\|\to 0}\frac{\log(I+U)-U}{\|U\|} = 0. \tag{2.60}$$

Hence by chain rule:

$$\hat{\nu}(M(P)^{-1/2}(DM(P)\cdot Z)M(P)^{-1/2}) \leqslant \hat{\nu}(P^{-1/2}ZP^{-1/2})$$

□

We consider the following time independent differential equation:

$$\begin{cases} \dot{x}(t) = \Phi(x(t)), \\ x(s) = x_0. \end{cases} \tag{2.61}$$

where Φ is differentiable on $\hat{\mathrm{S}}_n^+$. We assume that the associated flow $M_\cdot(\cdot) : (0,+\infty)\times\hat{\mathrm{S}}_n^+ \to \mathrm{S}_n$ leaves $\hat{\mathrm{S}}_n^+$ invariant and is globally defined.

Lemma 2.23. *Let ν be a symmetric gauge function. If there exists $\varepsilon > 0$ such that for any $0 \leqslant t \leqslant \varepsilon$,*

$$\hat{\nu}(M_t(I)^{-1/2}(\mathrm{D}M_t(I)\cdot Z)M_t(I)^{-1/2}) \leqslant \hat{\nu}(Z), \ \ \forall\ Z \in \mathrm{S}_n\,. \tag{2.62}$$

then

$$\langle \mathrm{diag}(\mu), \mathrm{D}\Phi(I)\cdot\mathrm{diag}(\lambda) - \mathrm{diag}(\lambda)\Phi(I)\rangle \leqslant 0, \ \ \forall\lambda \in \mathbb{R}^n, \ \ \mu \in \partial\nu(\lambda).$$

Proof. Let any $Z \in \mathrm{S}_n$. For readability, denote

$$P_t := M_t(I), \ \ H_t := P_t^{1/2}, \ \ Q_t := P_t^{-1/2}, \ \ G_t := P_t^{-1},$$

and

$$U_t := \mathrm{D}M_t(I)\cdot Z, \ \ J_t := U_tG_t, \ \ K_t = Q_tJ_tH_t.$$

The derivative of J_t with respect to t is:

$$\begin{aligned} \dot{J}_t &= \dot{U}_tG_t - U_tG_t\dot{P}_tG_t \\ &= (\mathrm{D}\Phi(P_t)\cdot U_t)G_t - U_tG_t\Phi(P_t)G_t. \end{aligned} \tag{2.63}$$

The derivative of K_t with respect to t is:

$$\begin{aligned}\dot{K}_t &= Q_t\dot{J}_tH_t + \dot{Q}_tJ_tH_t + Q_tJ_t\dot{H}_t\\ &= Q_t\dot{J}_tH_t - Q_t\dot{H}_tQ_tJ_tH_t + Q_tJ_t\dot{H}_t\\ &= Q_t\dot{J}_tH_t - Q_t\dot{H}_tK_t + K_tQ_t\dot{H}_t\end{aligned}$$

Hence,

$$\dot{J}_t|_{t=0} = \mathrm{D}\Phi(I)\cdot Z - Z\Phi(I),$$

and

$$\dot{K}_t|_{t=0} = \mathrm{D}\Phi(I)\cdot Z - Z\Phi(I) - (\dot{H}_t|_{t=0})Z + Z(\dot{H}_t|_{t=0}). \tag{2.64}$$

By Theorem 2.14, the right derivative of the function $\hat{\nu}(K_t)$ with respect to t exists:

$$\hat{\nu}(K_t)'_+ = \sup_{y\in\partial\hat{\nu}(K(t))}\langle y,\dot{K}_t\rangle = \sup_{\substack{\mu\in\partial\nu(\lambda),VV'=I\\ V'K_tV=\mathrm{diag}(\lambda)}}\langle V\,\mathrm{diag}(\mu)V',\dot{K}_t\rangle.$$

Since

$$\hat{\nu}(K_t)\leqslant\hat{\nu}(K_0),\ \ t\in[0,\delta),$$

the right derivative at $t=0$ must be negative:

$$\hat{\nu}(K_t)'_+|_{t=0}\leqslant 0.$$

Namely,

$$\sup_{\substack{\mu\in\partial\nu(\lambda),VV'=I\\ V'ZV=\mathrm{diag}(\lambda)}}\langle V\,\mathrm{diag}(\mu)V',\mathrm{D}\Phi(I)\cdot Z - Z\Phi(I) - (\dot{H}_t|_{t=0})Z + Z(\dot{H}_t|_{t=0})\rangle\leqslant 0.$$

Note that for any unitary matrix V such that $V'ZV=\mathrm{diag}(\lambda)$, we have

$$\langle V\,\mathrm{diag}(\mu)V',(\dot{H}_t|_{t=0})Z\rangle = \langle V\,\mathrm{diag}(\mu)V',Z(\dot{H}_t|_{t=0})\rangle$$

Hence by taking $Z=\mathrm{diag}(\lambda)$ and $V=I$, we obtain a necessary condition of (2.62):

$$\langle\mathrm{diag}(\mu),\mathrm{D}\Phi(I)\cdot\mathrm{diag}(\lambda)-\mathrm{diag}(\lambda)\Phi(I)\rangle\leqslant 0$$

for all $\lambda\in\mathbb{R}^n$ and $\mu\in\partial\nu(\lambda)$. □

The above two lemmas lead to the following conclusion:

Proposition 2.24. *If the flow $M.(\cdot):(0,+\infty)\times\hat{\mathrm{S}}_n^+\to\hat{\mathrm{S}}_n^+$ is non-expansive in the metric d_V, then,*

$$\langle\mathrm{diag}(\mu),\mathrm{D}\Phi(I)\cdot\mathrm{diag}(\lambda)-\mathrm{diag}(\lambda)\Phi(I)\rangle\leqslant 0 \tag{2.65}$$

for all $\lambda\in\mathbb{R}^n$ and $\mu\in\partial\nu(\lambda)$.

2.6.2 The counter example

We finally arrive at the announced counter example: we give a system of matrix parameters (A,B,C,D,L,Q,R) such that the corresponding Φ of GRDE does not satisfy the necessary condition (2.65) of non-expansiveness in any Finsler metric other than Thompson's part metric.

Recall that

$$\Phi(P) = A'P + PA + C'PC + Q - (B'P + D'PC + L)'(R + D'PD)^{-1}(B'P + D'PC + L).$$

Let I_n denote the n-dimensional identity matrix and $e = (e_1, \cdots, e_{n-1})' \in \mathbb{R}^{n-1}$ be a vector. The parameters are chosen as follows:

$$A = I_n,\ \ B = \begin{pmatrix} (\varepsilon - \sqrt{1-\varepsilon})I_{n-1} & 0 \\ -(\sqrt{1-\varepsilon})e' & \varepsilon \end{pmatrix},\ \ C = \begin{pmatrix} (1+\sqrt{1-\varepsilon})I_{n-1} & e \\ 0 & \sqrt{1-\varepsilon} \end{pmatrix},$$

and

$$D = (\sqrt{1-\varepsilon})I_n,\ \ L = 0,\ \ R = \varepsilon I_n,\ \ Q = \varepsilon I_n$$

to make

$$R + D'D = I_n,\ \ B' + D'C = I_n,\ \ C - D = \begin{pmatrix} I_{n-1} & e \\ 0 & 0 \end{pmatrix}$$

An elementary calculation yields

$$\begin{aligned}
&D\Phi(I)\cdot Z - Z\Phi(I) \\
&= A'Z + ZA + C'ZC - (B'Z + D'ZC)'(R + D'D)^{-1}(B' + D'C) \\
&\quad -(B' + D'C)(R + D'D)^{-1}(B'Z + D'ZC) \\
&\quad +(B' + D'C)(R + D'D)^{-1}D'ZD(R + D'D)^{-1}(B' + D'C) \\
&\quad -Z(A' + A + C'C + Q - (B' + D'C)'(R + D'D)^{-1}(B' + D'C)) \\
&= 2Z + C'ZC - (B'Z + D'ZC)' - (B'Z + D'ZC) + D'ZD - Z - ZC'C - Z'Q \\
&= Z + (C - D)'Z(C - D) - B'Z - ZB' - ZC'C - ZQ
\end{aligned}$$

Now let any $\lambda = (\lambda_1, \cdots, \lambda_n) \in \mathbb{R}^n$ and $\mu \in \partial\nu(\lambda)$. Then

$$\begin{aligned}
&\langle \mathrm{diag}(\mu), \mathrm{D}\Phi(I)\cdot\mathrm{diag}(\lambda) - \mathrm{diag}(\lambda)\Phi(I)\rangle \\
&= -2\varepsilon\langle \mu, \lambda\rangle + \mu_n(-\lambda_n|e|^2 + \textstyle\sum_{i=1}^{n-1}\lambda_i e_i^2)
\end{aligned}$$

Recall that

$$\langle \mu, \lambda\rangle = \nu(\lambda), \quad \forall \mu \in \partial\nu(\lambda).$$

So if there is any $\lambda \in \mathbb{R}^n$ and $\mu \in \partial\nu(\lambda)$ such that

$$\mu_n(-\lambda_n|e|^2 + \sum_{i=1}^{n-1}\lambda_i e_i^2) > 0,$$

then there always exists $\varepsilon \in (0,1)$ such that

$$\langle \mathrm{diag}(\mu), \mathrm{D}\Phi(I)\cdot\mathrm{diag}(\lambda) - \mathrm{diag}(\lambda)\Phi(I)\rangle > 0.$$

Finally we need a lemma to conclude:

Lemma 2.25. *If for all $\lambda \in \mathbb{R}^n$, $\mu \in \partial v(\lambda)$ and $e \in \mathbb{R}^{n-1}$ we have*

$$\mu_n(-\lambda_n\|e\|^2 + \sum_{i=1}^{n-1} \lambda_i e_i^2) \leqslant 0,$$

then

$$v(\lambda_1, \cdots, \lambda_n) = c \max_i |\lambda_i|$$

for some constant $c > 0$.

Proof. First consider $e = e_i$ the i-th standard basis vector of $\mathbb{R}^{n-1}$ for all $i = 1, \cdots, n-1$. We see that

$$\mu_n(-\lambda_n + \lambda_i) \leqslant 0, \quad \forall i = 1, \cdots, n-1$$

for all $\lambda = (\lambda_1, \cdots, \lambda_n) \in \mathbb{R}^n$ and $\mu = (\mu_1, \cdots, \mu_n) \in \partial v(\lambda)$. By the symmetric property of v, this implies actually

$$\mu_j(-\lambda_j + \lambda_i) \leqslant 0, \quad \forall i, j = 1, \cdots, n \tag{2.66}$$

Therefore, for any $\lambda \neq 0$ if $\lambda_j = 0$ then $\mu_j = 0$ for all $\mu \in \partial v(\lambda)$. Next, let any $i \in \{1, \cdots, n\}$, consider the following set

$$\Lambda_i := \{\lambda \neq 0 : \lambda_1 = \lambda_2 = \cdots = \lambda_i > \lambda_{i+1} \geqslant \ldots \lambda_n \geqslant 0\}.$$

Let any $\lambda \in \Lambda_i$ and $\mu \in \partial v(\lambda)$. By Property 1 in Lemma 2.21, $\mu \geqslant 0$. Using (2.66), we know that:

$$\mu_j \leqslant 0, \quad \forall j = i+1, \ldots, n.$$

Hence,

$$\mu_j = 0, \quad \forall j = i+1, \ldots, n.$$

Now let any $\lambda^1, \lambda^2 \in \Lambda_i$ and $\mu^1 \in \partial v(\lambda^1)$, $\mu^2 \in \partial v(\lambda^2)$. By Property 3 in Lemma 2.21,

$$\langle \mu^1 - \mu^2, \lambda^1 \rangle \geqslant 0.$$

It follows that

$$\sum_{j=1}^{i} \mu_j^1 \geqslant \sum_{j=1}^{i} \mu_j^2.$$

We deduce that $\sum_{j=1}^{i} \mu_j^1 = \sum_{j=1}^{i} \mu_j^2$. Hence there is a constant $c_i \geqslant 0$ such that

$$v(\lambda) = \langle \mu, \lambda \rangle = \sum_{i=1}^{j} \mu_j \lambda_j = \lambda_1 \sum_{i=1}^{j} \mu_j = c_i \lambda_1, \quad \forall \lambda \in \Lambda_i, \mu \in \partial v(\lambda).$$

It remains to prove that $c_i = c_1$ for all $i = 1, \ldots, n$. To see this, again we use Property 3 in Lemma 2.21. First consider $\lambda = (1, \ldots, 1, 0, \ldots, 0) \in \Lambda_i$ and any $\mu \in \partial v(\lambda)$, then

$$\langle \mu - (c_1, 0, \ldots, 0)', \lambda \rangle = \sum_{j=1}^{i} \mu_j - c_1 = c_i - c_1 \geqslant 0.$$

On the other hand, for all $\lambda^1 \in \Lambda_1$

$$\langle (c_1, 0, \ldots, 0)' - \mu, \lambda^1 \rangle = (c_1 - \mu_1)\lambda_1^1 - \sum_{j=2}^{i} \mu_j \lambda_j^1 \geqslant 0$$

This implies $c_1 = \sum_{j=1}^{i} \mu_j = c_i$ for all $i = 1, \ldots, n$. □

The proof of Theorem 2.3 is now complete.

2.7 Comparison with a theorem of Nussbaum

We next recall a characterization of the contraction rate in Thompson's part metric of a not necessarily order-preserving flow established by Nussbaum [Nus94]. The result there is established in the finite dimensional setting. We slightly modified the statement of [Nus94] in order to unify the notation and thus to make easier the comparison with our results.

Theorem 2.15 (Thm 3.9 in [Nus94])**.** *Consider a finite dimensional vector space $\mathscr{X}$. Suppose that the dynamics $\phi(\cdot,\cdot)$ is defined on $J\times\mathscr{C}_0$. Let $D_0\subset\mathscr{C}_0$ be a compact set. Let $0\leqslant t_0<t_1$ such that the flow is defined on $[t_0,t_1]\times[t_0,t_1]\times D_0$. Define $D_1\subset\mathscr{C}_0$ by*

$$D_1=\{M^t_{t_0}(x):x\in D_0,t_0\leqslant t\leqslant t_1\}.$$

Let D_2 be a compact set satisfying the following property: for all $x,y\in D_1$, there exists a piecewise C^1 minimal geodesic (with respect to the part metric) $\varphi:[0,1]\to\mathscr{C}_0$ with $\varphi(0)=x$, $\varphi(1)=y$ and $\varphi(t)\in D_2$ for $0\leqslant t\leqslant 1$. Then for any points $x_1,x_2\in D_0$, we have

$$d_T(M^t_{t_0}(x_1),M^t_{t_0}(x_2))\leqslant\exp\big(\int_{t_0}^t k(s,D_2)dr\big)d_T(x_1,x_2),\ \ t_0\leqslant t\leqslant t_1$$

where

$$k(s,D_2):=\limsup_{\Delta\to0^+}\frac{c(s,\Delta,D_2)-1}{\Delta},\tag{2.67}$$

and

$$c(s,\Delta,D_2):=\sup_{x\in D_2}\inf\{\lambda>0:-\lambda(x+\Delta\phi(s,x))\preccurlyeq z+\Delta D\phi_s(x)z\preccurlyeq\lambda(x+\Delta\phi(s,x)),\forall -x\preccurlyeq z\preccurlyeq x\}.$$

Nussbaum's proofs rely on the Finsler nature of Thompson's part metric. In the case of a finite-dimensional order-preserving flow, we deduce from his result the following corollary.

Corollary 2.26. *Consider a finite dimensional vector space $\mathscr{X}$. Let ϕ be defined on $J\times\mathscr{U}$ where $\mathscr{U}\subset\mathscr{C}_0$ is an open set. Let $\bar{\mathscr{U}}\supset\mathscr{U}$ be a closed set satisfying the following property: for all $x,y\in\mathscr{U}$, there exists a piecewise C^1 minimal geodesic (with respect to the part metric) $\varphi:[0,1]\to\mathscr{C}_0$ with $\varphi(0)=x$, $\varphi(1)=y$ and $\varphi(t)\in\bar{\mathscr{U}}$ for $0\leqslant t\leqslant 1$. Assume that the function $\alpha(\cdot,\bar{\mathscr{U}}):J\to\mathbb{R}$ defined by*

$$\alpha(s,\bar{\mathscr{U}})=-\sup_{x\in\bar{\mathscr{U}}}M((D\phi_s(x)x-\phi(s,x))/x),\ \ s\in J$$

is locally integrable. If the flow $M_\cdot(\cdot)$ is order-preserving on $\mathscr{U}$, then for all $x_1,x_2\in\mathscr{U}$,

$$d_T(M^t_s(x_1),M^t_s(x_2))\leqslant\exp\big(\int_s^t-\alpha(r,\bar{\mathscr{U}})dr\big)d_T(x_1,x_2),\ \ 0\leqslant s\leqslant t<t_{\mathscr{U}}(s,x_1)\wedge t_{\mathscr{U}}(s,x_2).\tag{2.68}$$

Proof. Let $x_1,x_2\in\mathscr{U}$ and $0\leqslant t_0<t_1<t_{\mathscr{U}}(t_0,x_1)\wedge t_{\mathscr{U}}(t_0,x_2)$. Define

$$D_0=\{x_1,x_2\},\quad D_1=\{M^t_{t_0}(x):x\in D_0,t_0\leqslant t\leqslant t_1\}.$$

Then by the closure of $\bar{\mathscr{U}}$ and the boundedness of D_1, there is a compact $D_2\subset\bar{\mathscr{U}}$ satisfying the geodesic condition in Theorem 2.15. Therefore,

$$d_T(M^t_{t_0}(x_1),M^t_{t_0}(x_2))\leqslant\exp\big(\int_{t_0}^t k(s,D_2)ds\big)d_T(x_1,x_2),\ \ t_0\leqslant t\leqslant t_1.$$

We next show that

$$k(s,D_2)=\sup_{x\in D_2} M((D\phi_s(x)-\phi(s,x))/x),\ \ t_0\leqslant s\leqslant t_1.$$

By definition, for a $\Delta\geqslant 0$ sufficiently small,

$$c(s,\Delta,D_2)=\sup_{x\in D_2}\sup_{-x\preccurlyeq z\preccurlyeq x}\sup_{\substack{q\in\mathscr{C}^*\\ |q|^*=1}}\frac{|\,\langle q,z+\Delta D\phi_s(x)z\rangle\,|}{\langle q,x+\Delta\phi(s,x)\rangle} \tag{2.69}$$

$$=\sup_{x\in D_2}\sup_{-x\preccurlyeq z\preccurlyeq x}\sup_{\substack{q\in\mathscr{C}^*\\ |q|^*=1}}\frac{\langle q,z+\Delta D\phi_s(x)z\rangle}{\langle q,x+\Delta\phi(s,x)\rangle} \tag{2.70}$$

For fixed s and D_2, the function $c(s,\cdot,D_2):\mathbb{R}_+\to\mathbb{R}$ is a subsmooth function (see Appendix A). Then we apply (A.3) to calculate its one-side derivative at point 0:

$$\begin{aligned}&\lim_{\Delta\to 0^+}\frac{c(s,\Delta,D_2)-1}{\Delta}\\ &=\sup_{x\in D_2}\sup_{-x\preccurlyeq z\preccurlyeq x}\sup_{\substack{\langle q,x-z\rangle=0\\ q\in\mathscr{C}^*,|q|^*=1}}\lim_{\Delta\to 0^+}\frac{\langle q,z+\Delta D\phi_s(x)z\rangle/\langle q,x+\Delta\phi(s,x)\rangle-1}{\Delta}\end{aligned}$$

Since the flow is order-preserving, by Proposition 2.3, for all $q\in\mathscr{C}^\star, z\preccurlyeq x$ such that $\langle q,x\rangle=\langle q,z\rangle$ we must have

$$\langle q,z+\Delta D\phi_s(x)z\rangle\leqslant\langle q,x+\Delta D\phi_s(x)x\rangle.$$

Therefore,

$$\begin{aligned}k(s,D_2)&=\lim_{\Delta\to 0^+}\frac{c(s,\Delta,D_2)-1}{\Delta}\\ &\leqslant\sup_{x\in D_2}\sup_{\substack{q\in\mathscr{C}^*\\ |q|^*=1}}\lim_{\Delta\to 0^+}\frac{\langle q,x+\Delta D\phi_s(x)x\rangle/\langle q,x+\Delta\phi(s,x)\rangle-1}{\Delta}\\ &=\sup_{x\in D_2}\sup_{\substack{q\in\mathscr{C}^*\\ |q|^*=1}}\frac{\langle q,D\phi_s(x)x-\phi(s,x)\rangle}{\langle q,x\rangle}\\ &=\sup_{x\in D_2}M((D\phi_s(x)x-\phi(s,x))/x).\end{aligned}$$

Define

$$\tilde{c}(s,\Delta,D_2):=\sup_{x\in D_2}\inf\{\lambda>0: x+\Delta D\phi_s(x)x\preccurlyeq\lambda(x+\Delta\phi(s,x))\}.$$

It is clear that

$$c(s,\Delta,D_2)\geqslant\tilde{c}(s,\Delta,D_2).$$

Besides,

$$\begin{aligned}\lim_{\Delta\to 0^+}\frac{\tilde{c}(s,\Delta,D_2)-1}{\Delta}&\geqslant\sup_{x\in D_2}\lim_{\Delta\to 0^+}\frac{M((x+\Delta D\phi_s(x)x)/(x+\Delta\phi(s,x)))-1}{\Delta}\\ &\geqslant\sup_{x\in D_2}\sup_{\substack{q\in\mathscr{C}^*\\ |q|^*=1}}\lim_{\Delta\to 0^+}\frac{\langle q,x+\Delta D\phi_s(x)x\rangle/\langle q,x+\Delta\phi_s(x)\rangle-1}{\Delta}\\ &=\sup_{x\in D_2}M((D\phi_s(x)x-\phi(x))/x).\end{aligned}$$

Therefore,

$$k(s,D_2)=\lim_{\Delta\to 0^+}\frac{c(s,\Delta,D_2)-1}{\Delta}\geqslant\sup_{x\in D_2}M((D\phi_s(x)x-\phi(x))/x)$$

Hence we proved that

$$k(s,D_2)=-\alpha(s,D_2)\leqslant-\alpha(s,\bar{\mathscr{U}}),\ \ t_0\leqslant s\leqslant t_1.$$

It follows that

$$d_T(M^t_{t_0}(x_1),M^t_{t_0}(x_2))\leqslant\exp\big(\int_{t_0}^t-\alpha(s,\bar{\mathscr{U}})ds\big)d_T(x_1,x_2),\ \ t_0\leqslant t\leqslant t_1.$$

Since $0\leqslant t_0<t_1<t_{\mathscr{U}}(t_0,x_1)\wedge t_{\mathscr{U}}(t_0,x_2)$ are arbitrary, we deduce (2.68). □

Remark 2.10. The formula (A.3) used in the proof to get the directional derivative of $c(s,\cdot,D_2)$ at point 0 requires that the supremum of (2.69) be taken over compact sets. In an infinite-dimensional setting, the interval $[-x,x]$ for some $x\in D_2$ is in general not a compact set. In that case, one may need other techniques to obtain (2.68) as a corollary of Nussbaum's result.

Remark 2.11. For a finite dimensional order-preserving flow, the main difference between Corollary 2.26 and Theorem 2.7 is the following. Let $\mathscr{U}\subset\mathscr{C}_0$ be an arbitrary open set. Theorem 2.7 shows that

$$d_T(M^t_s(x_1),M^t_s(x_2))\leqslant\exp\big(\int_s^t-\alpha(r,\tilde{\mathscr{U}})dr\big)d_T(x_1,x_2),\ \ 0\leqslant s\leqslant t<t_{\mathscr{U}}(s,x_1)\wedge t_{\mathscr{U}}(s,x_2).$$

where $\tilde{\mathscr{U}}$ is defined to be the following "radial closure" of $\mathscr{U}$:

$$\tilde{\mathscr{U}}:=\{\lambda\mathscr{U}:\lambda\in(0,1]\}.$$

Similarly, we may find a set $\bar{\mathscr{U}}\supset\mathscr{U}$ ("geodesic convex hull") satisfying the geodesic constraint in Corollary 2.26, and then, Corollary 2.26, obtained through Nussbaum's theorem (Theorem 2.15), shows that

$$d_T(M^t_s(x_1),M^t_s(x_2))\leqslant\exp\big(\int_s^t-\alpha(r,\bar{\mathscr{U}})dr\big)d_T(x_1,x_2),\ \ 0\leqslant s\leqslant t<t_{\mathscr{U}}(s,x_1)\wedge t_{\mathscr{U}}(s,x_2).$$

In a number of concrete examples, the geodesic convexity condition of Nussbaum is satisfied ($\mathscr{U}=\bar{\mathscr{U}}$) and leads to optimal estimates. In fact in the previous applications to the deterministic and stochastic Riccati equations the domains considered are all geodesically convex. However, there are examples of domains and maps for which our approach yields a tighter estimate of the contraction rate, because the "geodesic convex hull" $\bar{\mathscr{U}}$ is too large. Below is an example for which $\alpha(s,\tilde{\mathscr{U}})>0$ but $\alpha(s,\bar{\mathscr{U}})<0$.

Example 2.12. Consider the following differential equation defined on $\mathrm{int}(\mathbb{R}^2_+)$:

$$\begin{cases}\dot{X}=X\phi_1(\ln X,\ln Y)\\ \dot{Y}=Y\phi_2(\ln X,\ln Y)\end{cases}\tag{2.71}$$

where $\phi_1,\phi_2:\mathbb{R}^2\to\mathbb{R}$ are defined by:

$$\begin{array}{l}\phi_1(x,y)=x\arctan x-\frac{1}{2}\ln(1+x^2)+\arctan y-(\frac{1+\pi}{2})x-\frac{\pi}{2}\\ \phi_2(x,y)=\frac{x}{8}-\frac{y}{4}\end{array}$$

It is clear that $(X(\cdot),Y(\cdot)):[0,T)\to \mathrm{int}(\mathbb{R}^2_+)$ is a solution of (2.71) if and only if $(\ln(X(\cdot)),\ln(Y(\cdot))):[0,T)\to\mathbb{R}^2$ is a solution of

$$\begin{cases} \dot{x}=\phi_1(x,y)\\ \dot{y}=\phi_2(x,y)\end{cases} \tag{2.72}$$

Since ϕ_1 and ϕ_2 are globally Lipschitz functions, we know that the flow associated to (2.72) exists globally. Hence, the flow associated to (2.71) leaves $\mathrm{int}(\mathbb{R}^2_+)$ invariant and also exists globally. Since $\ln(\cdot)$ is an increasing function, the flow of (2.71) is order-preserving if and only if the one of (2.72) is. We verify the Kamke condition [HS05] for System (2.72):

$$\frac{\partial\phi_1}{\partial y}(x,y)=\frac{1}{1+y^2}\geqslant 0,\quad \frac{\partial\phi_2}{\partial x}(x,y)=\frac{1}{8}\geqslant 0$$

Hence the system (2.71) is order-preserving on $\mathrm{int}(\mathbb{R}^2_+)$. Let $h=3/2$. Now consider the following open set

$$\mathscr{U}=\{(X,Y)\in\mathrm{int}(\mathbb{R}^2_+):hY>X \ \text{ or } \ Y^2<hX^{-1}\}$$

Interested reader can verify that for any $\delta\in[1,h)$, the closed set

$$\{(X,Y)\in\mathrm{int}(\mathbb{R}^2_+):\delta Y\geqslant X \ \text{ or } \ Y^2\leqslant\delta X^{-1}\}\subset\mathscr{U}$$

is invariant with respect to the flow of (2.71). Therefore, for any $Z=(X,Y)\in\mathscr{U}$, the leaving time $t_{\mathscr{U}}(Z)$ equals to $+\infty$.

It is clear that $\lambda\mathscr{U}\subset\mathscr{U}$ for all $\lambda\in(0,1]$. Now we calculate $\alpha(\mathscr{U})$. By definition,

$$-\alpha(\mathscr{U})=\sup_{(X,Y)\in\mathscr{U}}\max(\frac{\partial\phi_1}{\partial x}(\ln X,\ln Y)+\frac{\partial\phi_1}{\partial y}(\ln X,\ln Y),\frac{\partial\phi_2}{\partial x}(\ln X,\ln Y)+\frac{\partial\phi_2}{\partial y}(\ln X,\ln Y)).$$

Denote

$$\mathscr{V}=\{(x,y)\in\mathbb{R}^2:(e^x,e^y)\in\mathscr{U}\}=\{(x,y)\in\mathbb{R}^2:y>x-\ln h \ \text{ or } \ y<-\frac{x-\ln h}{2}\}.$$

Then

$$-\alpha(\mathscr{U})=\sup_{(x,y)\in\mathscr{V}}\max(\frac{\partial\phi_1}{\partial x}(x,y)+\frac{\partial\phi_1}{\partial y}(x,y),\frac{\partial\phi_2}{\partial x}(x,y)+\frac{\partial\phi_2}{\partial y}(x,y)).$$

We have:

$$\frac{\partial\phi_1}{\partial x}(x,y)+\frac{\partial\phi_1}{\partial y}(x,y)=\arctan x+\frac{1}{1+y^2}-\frac{1+\pi}{2},$$

and

$$\frac{\partial\phi_2}{\partial x}(x,y)+\frac{\partial\phi_2}{\partial y}(x,y)=-\frac{1}{8}.$$

Next we show that every set $\bar{\mathscr{U}}\supset\mathscr{U}$ satisfying the geodesic constraint in Corollary 2.26 contains $\mathrm{int}(\mathbb{R}^2_+)$. Note that the minimal geodesics with respect to Thompson's part metric in $\mathrm{int}\,\mathbb{R}^2_+$ are in one to one correspondence (by a logarithmic transformation) with the minimal geodesics with respect to the sup-norm in $\mathbb{R}^2$. The unique minimal geodesic with respect to the sup-norm between two points $(a,b)\in\mathbb{R}^2$ and $(c,d)\in\mathbb{R}^2$ such that $\mid a-c\mid=\mid b-d\mid$ is the straight line. Hence, for any $(X,Y)\in$

$\mathrm{int}(\mathbb{R}^2_+)\setminus\mathscr{U}$, the minimal geodesic (with respect to Thompson's part metric) from $(\frac{\sqrt{XY}}{2}, 2\sqrt{XY})\in\mathscr{U}$ to $(2X^2Y^2, \frac{1}{2XY})\in\mathscr{U}$ is unique and passes through (X,Y). Therefore,

$$-\alpha(\bar{\mathscr{U}}) = -\alpha(\mathrm{int}(\mathbb{R}^2_+)) = \frac{1}{2}.$$

Hence Corollary 2.26 yields that for any two solutions $Z_1(\cdot), Z_2(\cdot) : [0,+\infty)\to\mathscr{U}$ of (2.71):

$$d_T(Z_1(t),Z_2(t)) \leqslant e^{\frac{t-s}{2}} d_T(Z_1(s),Z_2(s)),\ \ 0\leqslant s<t<+\infty.$$

However, it can be checked that the level line

$$\{(x,y) : \arctan x + \frac{1}{1+y^2} = \frac{\pi}{2}+\frac{3}{8}\}$$

does not intersect the boundary of $\mathscr{V}$:

$$\{(x,y) : x\geqslant 0, y = x-\ln h \ \text{ or } \ y = -\frac{x-\ln h}{2}\}.$$

Therefore,

$$\alpha(\mathscr{U}) \geqslant \frac{1}{8},$$

and Theorem 2.5 implies that the system (2.71) is a strict contraction on the domain $\mathscr{U}$ with a contraction rate at least equal to $1/8$. That is, for any two solutions $Z_1(\cdot), Z_2(\cdot) : [0,+\infty)\to\mathscr{U}$ of (2.71) we have:

$$d_T(Z_1(t),Z_2(t)) \leqslant e^{-\frac{t-s}{8}} d_T(Z_1(s),Z_2(s)),\ \ 0\leqslant s<t<+\infty.$$

We deduce from the latter formula the existence and uniqueness of a fixed point in the domain $\mathscr{U}$ and the exponential convergence of all the trajectories to that fixed point with a uniform rate at least equal to $1/8$.

Remark 2.13. Although the estimation of the global contraction rate in Theorem 2.5 can be recovered as a corollary of Nussbaum's theorem when the domain is geodesically convex, invariance arguments used in the proof lead to tighter estimates of the convergence rate to a fixed point (Theorem 2.9). More precisely, the convergence rate (2.47) in Theorem 2.13 follows from Theorem 2.9 and is tighter than the one obtained by applying Theorem 2.5, which is:

$$\alpha \geqslant e^{-d_T(P,\bar{P})} m((Q-L'R^{-1}L)/\bar{P}).$$

CHAPTER 3

Dobrushin ergodicity coefficient for consensus operators on cones

In the previous chapter, we determined the contraction rate of nonlinear order-preserving maps or flows in Thompson's metric. A related problem is to characterize the contraction rate of nonlinear maps or flows in Hilbert's projective metric, which is a weak Finsler metric. In the present chapter we consider contraction properties with respect to *Hilbert's seminorm*, which is the infinitesimal distance associated to Hilbert's projective metric. We give a characterization of the contraction ratio of bounded linear maps in Banach space with respect to Hilbert's seminorm, in terms of the extreme points of a certain abstract "simplex". The formula is then applied to abstract consensus operators defined on arbitrary cones, which extend the row stochastic matrices acting on the standard positive cone and the completely positive unital maps acting on the cone of positive semidefinite matrices. When applying our characterization to a stochastic matrix, we recover the formula of Dobrushin's ergodicity coefficient. When applying our result to a completely positive unital map, we therefore obtain a noncommutative version of Dobrushin's ergodicity coefficient, which gives the contraction ratio of the map (representing a quantum channel or a "noncommutative Markov chain") with respect to the diameter of the spectrum. The contraction ratio of the dual operator (Kraus map) with respect to the total variation distance will be shown to be given by the same coefficient.

We finally consider some complexity issues for Kraus maps. Whereas contraction properties are easy to check for stochastic matrices, the verification of their noncommutative analogues require efforts. Using the noncommutative Dobrushin's ergodicity coefficient, we show that a number of decision problems concerning the contraction rate of Kraus maps reduce to finding a rank one matrix

in linear spaces satisfying certain conditions. We then show that an irreducible Kraus map is primitive if and only if the associated noncommutative consensus system is globally convergent, which can be checked in polynomial time if the map is irreducible. However, we prove that unlike in the case of standard nonnegative matrices, deciding whether a Kraus map is strictly positive (meaning that it sends the cone to its interior) is NP-hard.

This chapter is an extended version of an ECC conference article [GQ13].

3.1 Introduction

3.1.1 Motivation: from Birkhoff's theorem to consensus dynamics

Hilbert's projective metric d_H on the interior of a (closed, convex, and pointed) cone $\mathscr{C}$ in a Banach space $\mathscr{X}$ can be defined by:

$$d_H(x,y) := \log\inf\{\frac{\beta}{\alpha} :\ \alpha,\beta>0,\ \alpha x \preccurlyeq y \preccurlyeq \beta x\},$$

where $\preccurlyeq$ is the partial order induced by $\mathscr{C}$. Birkhoff [Bir57] characterized the contraction ratio with respect to d_H of a linear map T preserving the interior $\mathscr{C}^0$ of the cone $\mathscr{C}$,

$$\sup_{x,y\in\mathscr{C}^0}\frac{d_H(Tx,Ty)}{d_H(x,y)} = \tanh(\frac{\operatorname{diam} T(\mathscr{C}^0)}{4}), \qquad \operatorname{diam} T(\mathscr{C}^0) := \sup_{x,y\in\mathscr{C}^0} d_H(Tx,Ty)\ .$$

This fundamental result, which implies that a linear map sending the cone $\mathscr{C}$ into its interior is a strict contraction in Hilbert's metric, can be used to derive the Perron-Frobenius theorem from the Banach contraction mapping theorem, see [Bus73, KP82, EN95] for more information.

Hilbert's projective metric is related to the following family of seminorms. To any point $\mathbf{e}\in\mathscr{C}^0$ is associated the seminorm

$$x \mapsto \omega(x/\mathbf{e}) := \inf\{\beta-\alpha : \alpha\mathbf{e} \preccurlyeq x \preccurlyeq \beta\mathbf{e}\}$$

which is sometimes called *Hopf's oscillation* [Hop63, Bus73] or *Hilbert's seminorm* [GG04]. Nussbaum [Nus94] showed that d_H is precisely the weak Finsler metric obtained when taking $\omega(\cdot/\mathbf{e})$ to be the infinitesimal distance at point $\mathbf{e}$. In other words,

$$d_H(x,y) = \inf_\gamma \int_0^1 \omega(\dot\gamma(s)/\gamma(s))ds$$

where the infimum is taken over piecewise C^1 paths $\gamma:[0,1]\to\mathscr{C}^0$ such that $\gamma(0)=x$ and $\gamma(1)=y$. He deduced that the contraction ratio, with respect to Hilbert's projective metric, of a nonlinear map $f:\mathscr{C}^0\to\mathscr{C}^0$ that is positively homogeneous of degree 1 (i.e. $f(\lambda x)=\lambda f(x)$ for all $\lambda>0$), can be expressed in terms of the Lipschitz constants of the linear maps $Df(x)$ with respect to a family of Hopf's oscillation seminorms:

$$\sup_{x,y\in U}\frac{d_H(f(x),f(y))}{d_H(x,y)} = \sup_{x\in U}\sup_{\substack{z\in\mathscr{X}\\ \omega(z/x)\neq 0}}\frac{\omega(Df(x)z/f(x))}{\omega(z/x)}\ . \tag{3.1}$$

Hence, to arrive at an explicit formula for the contraction rate of nonlinear maps in Hilbert's projective metric , a basic issue is to determine the Lipschitz constant of a bounded linear map $T : \mathscr{X} \to \mathscr{X}$ with respect to Hopf's oscillation seminorm, i.e.,

$$\|T\|_H := \sup_{z\in\mathscr{X},\, \omega(z/\mathbf{e})\neq 0} \frac{\omega(T(z)/T(\mathbf{e}))}{\omega(z/\mathbf{e})} \quad . \tag{3.2}$$

The problem of computing the contraction rate (3.2) also arises in the study of consensus algorithms. A *consensus operator* is a linear map T which preserves the positive cone $\mathscr{C}$ and fixes a unit element $\mathbf{e} \in \mathscr{C}^0$:

$$T(\mathbf{e}) = \mathbf{e} \ .$$

A discrete time consensus system can be described by

$$x_{k+1} = T_{k+1}(x_k), \quad k \in \mathbb{N}, \tag{3.3}$$

where $T_1, T_2, \dots$ is a sequence of consensus operators preserving the same unit element $\mathbf{e}$. The main concern of consensus theory is the convergence of the orbit x_k to a *consensus state*, which is represented by a scalar multiple of the unit element.

This model includes in particular the classical linear consensus system case when $\mathscr{X} = \mathbb{R}^n$, $\mathscr{C} = \mathbb{R}^n_+$, $\mathbf{e} = (1, \cdots, 1)^\top$ and

$$x_{k+1} = Ax_k, \ \ k \in \mathbb{N} \ , \tag{3.4}$$

where A is a row stochastic matrix. This has been studied in the field of communication networks, control theory and parallel computation [Hir89, BT89, BGPS06, Mor05, VJAJ05, OT09, AB09]. A widely used Lyapunov function for the consensus dynamics, first considered by Tsitsiklis (see [TBA86]), is the "diameter" of the state x defined as

$$\Delta(x) = \max_{1\leqslant i,j\leqslant n} (x_i - x_j),$$

which is precisely Hopf's oscillation seminorm $\omega(x/\mathbf{e})$. It turns out that the latter seminorm can still be considered as a Lyapunov function for a consensus operator T, with respect to an arbitrary cone. When $\mathscr{C} = \mathbb{R}^n_+$, it is well known that if the contraction ratio of the stochastic matrix A with respect to the diameter is strictly less than one, then the orbits of the consensus dynamics (3.4) converge exponentially to a consensus state. We shall see here that the same remains true in general (Theorem 3.4). For time-dependent consensus systems, a common approach is to bound the contraction ratio of every product of p consecutive operators $T_{i+p} \circ \cdots \circ T_{i+1}$, $i = 1, 2, \dots$, for a fixed p, see for example [Mor05]. Moreover, if $\{T_k : k \geqslant 1\}$ is a stationary ergodic random process, then the almost sure convergence of the orbits of (3.3) to a consensus state can be deduced by showing that $\mathbb{E}[\log \|T_{1+p} \dots T_1\|_H] < 0$ for some $p > 0$, see Bougerol [Bou93]. Hence, in consensus applications, a central issue is again to compute the contraction ratio (3.2).

3.1.2 Main results

Our first result characterizes the contraction ratio (3.2), in a slightly more general setting. We consider a bounded linear map T from a Banach space $\mathscr{X}_1$ to a Banach space $\mathscr{X}_2$. The latter are equipped with normal cones $\mathscr{C}_i \subset \mathscr{X}_i$, and *unit* elements $\mathbf{e}_i \in \mathscr{C}_i^0$.

Theorem 3.1 (Contraction rate in Hopf's oscillation seminorm)**.** *Let $T : \mathscr{X}_1 \to \mathscr{X}_2$ be a bounded linear map such that $T(\mathbf{e}_1) \in \mathbb{R}\mathbf{e}_2$. Then*

$$\sup_{\substack{z\in\mathscr{X}_1\\ \omega(z/\mathbf{e}_1)\neq 0}} \frac{\omega(T(z)/\mathbf{e}_2)}{\omega(z/\mathbf{e}_1)} = \frac{1}{2} \sup_{\substack{\nu,\pi\in\operatorname{extr}\mathscr{P}(\mathbf{e}_2)\\ \nu\perp\pi}} \|T^\star(\nu)-T^\star(\pi)\|_T^\star = \sup_{\substack{\nu,\pi\in\operatorname{extr}\mathscr{P}(\mathbf{e}_2)\\ \nu\perp\pi}} \sup_{x\in[0,\mathbf{e}_1]} \langle \nu-\pi, T(x)\rangle.$$

The notations and notions used in this theorem are detailed in Section 3.5. In particular, we denote by the same symbol $\preccurlyeq$ the order relations induced by the two cones $\mathscr{C}_i$, $i=1,2$; $\mathscr{P}(\mathbf{e}_2) = \{\mu \in \mathscr{C}_2^\star : \langle \mu, \mathbf{e}_2\rangle = 1\}$ denotes the abstract *simplex* of the dual Banach space $\mathscr{X}_2^\star$ of $\mathscr{X}_2$, where $\mathscr{C}_2^\star$ is the dual cone of $\mathscr{C}_2$; extr denotes the extreme points of a set; $\perp$ denotes a certain *disjointness* relation, which will be seen to generalize the condition that two measures have disjoint supports; and $T^\star$ denotes the adjoint of T. We shall make use of the following norm, which we call *Thompson's norm*,

$$\|z\|_T = \inf\{\alpha > 0 : -\alpha\mathbf{e}_1 \preccurlyeq z \preccurlyeq \alpha\mathbf{e}_1\}$$

on the space $\mathscr{X}_1$, and denote by $\|\cdot\|_T^\star$ the dual norm.

When $\mathscr{C} = \mathbb{R}^n_+$, and $T(z) = Az$ for some stochastic matrix A, we shall see that the second supremum in Theorem 3.1 is simply

$$\frac{1}{2}\max_{i<j} \sum_{1\leqslant k\leqslant n} |A_{ik} - A_{jk}| = \frac{1}{2}\max_{i<j} \|A_{i\cdot} - A_{j\cdot}\|_{\ell_1} \ ,$$

where $A_{i\cdot}$ denotes the ith row of the matrix A. This quantity is called *Doeblin contraction coefficient* in the theory of Markov chains; it is known to determine the contraction rate of the adjoint $T^\star$ with respect to the ℓ_1 (or total variation) metric, see [LPW09]. Moreover, the last supremum in Theorem 3.1 can be rewritten more explicitly as

$$1 - \min_{i<j} \sum_{s=1}^n \min(A_{is}, A_{js}) \ ,$$

a term which is known as *Dobrushin's ergodicity coefficient* [Dob56]. Note that in general, the norm $\|\cdot\|_T^\star$ can be thought of as an abstract version of the ℓ_1 or total variation norm.

When specializing to a unital completely positive map T on the cone of positive semidefinite matrices, representing a quantum channel [SSR10, RKW11], we shall see that the last supremum in Theorem 3.1 coincides with the following expression, which provides a non commutative analogue of Dobrushin's ergodicity coefficient (see Corollary 3.9):

$$1 - \min_{\substack{X=(x_1,\ldots,x_n)\\ XX^*=I_n}} \min_{\substack{u,v:u^*v=0\\ u^*u=v^*v=1}} \sum_{i=1}^n \min\{u^*T(x_ix_i^*)u, v^*T(x_ix_i^*)v\} \ .$$

We finally address some complexity issues using the latter noncommutative version of Dobrushin's ergodicity coefficient. More precisely we address three decision problems for Kraus map, all straightforward for classical stochastic matrix. First, we study the complexity of checking irreducibility and the primitivity of a Kraus map (Section 3.8.2). We show that the global convergence of a noncommutative consensus system is equivalent to the existence of a rank one matrix in certain matrix subspace (Theorem 3.6). It follows from this characterization a noncommutative extension of the property that a Markov matrix is primitive if and only if it is irreducible and aperiodic (Proposition 3.13), as in the classical stochastic matrix case. This result implies that deciding if a (rational) noncommutative

consensus system is globally convergent can be done in polynomial time if the consensus operator is irreducible (Corollary 3.15). Secondly we consider the complexity of deciding the strict positivity of a Kraus map (Section 3.8.3). We show that deciding if there is a rank one matrix orthogonal to the space generated by the Kraus operators is NP-hard, by reducing a 3SAT problem to it (Theorem 3.8), which implies that deciding if a completely positive unital map is strictly positive is NP-hard (Corollary 3.20). Finally we deduce from Corollary 3.10 the equivalent between the strict contraction of a Kraus map and a quantum clique problem.

3.2 Thompson's norm and Hilbert's seminorm

We start by some preliminary results. Some of the notations are already introduced in Section 2.2.1.

In the whole chapter, $(\mathscr{X}, \|\cdot\|)$ is a real Banach space with dual space $\mathscr{X}^\star$. Let $\mathscr{C} \subset \mathscr{X}$ be a closed pointed convex cone with non empty interior $\mathscr{C}_0$ and $\preccurlyeq$ be the partial order induced by $\mathscr{C}$. For $x \in \mathscr{X}$ and $y \in \mathscr{X}_0$, we call *oscillation* [Bus73] the difference between $M(x/y)$ and $m(x/y)$:

$$\omega(x/y) := M(x/y) - m(x/y).$$

Let $\mathbf{e}$ denote a distinguished element in $\mathscr{C}_0$, which we shall call a *unit*. For $x \in \mathscr{X}$, define

$$\|x\|_T := \max(M(x/\mathbf{e}), -m(x/\mathbf{e}))$$

which we call *Thompson's norm*, with respect to the element $\mathbf{e}$, and

$$\|x\|_H := \omega(x/\mathbf{e})$$

which we call *Hilbert's seminorm* with respect to the element $\mathbf{e}$.

Remark 3.1. These terminologies are motivated by the fact that Thompson's part metric and Hilbert's projective metric are Finsler metrics (see [Nus94]) for which the infinitesimal distances at the point $\mathbf{e} \in \mathscr{C}^0$ are respectively given by $\|\cdot\|_T$ and $\|\cdot\|_H$. The seminorm $\|\cdot\|_H$ is also called Hopf's oscillation seminorm [Bus73].

We assume that the cone is normal. It is known that under this assumption the two norms $\|\cdot\|$ and $\|\cdot\|_T$ are equivalent, see [Nus94]. Therefore the space $\mathscr{X}$ equipped with the norm $\|\cdot\|_T$ is a Banach space. Since Thompson's norm $\|\cdot\|_T$ is defined with respect to a particular element $\mathbf{e}$, we write $(\mathscr{X}, \mathbf{e}, \|\cdot\|_T)$ instead of $(\mathscr{X}, \|\cdot\|_T)$. By the definition and (2.4), Thompson's norm can be rewritten by:

$$\|x\|_T = \sup_{z \in \mathscr{C}^\star} \frac{|\langle z, x\rangle|}{\langle z, \mathbf{e}\rangle} \quad . \tag{3.5}$$

Example 3.2. We consider the finite dimensional vector space $\mathscr{X} = \mathbb{R}^n$, the standard orthant cone $\mathscr{C} = \mathbb{R}^n_+$ and the unit vector $\mathbf{e} = \mathbf{1} := (1, \ldots, 1)^T$. It can be checked that Thompson's norm with respect to $\mathbf{1}$ is nothing but the sup norm

$$\|x\|_T = \max_i |x_i| = \|x\|_\infty,$$

whereas Hilbert's seminorm with respect to $\mathbf{1}$ is the so called *diameter*:

$$\|x\|_H = \max_{1 \leqslant i,j \leqslant n} (x_i - x_j) = \Delta(x).$$

Example 3.3. Let $\mathscr{X} = \mathrm{S}_n$, the space of Hermitian matrices of dimension n and $\mathscr{C} = \mathrm{S}_n^+$, the cone of positive semidefinite matrices. Let the identity matrix I_n be the unit element: $\mathbf{e} = I_n$. Then Thompson's norm with respect to I_n is nothing but the sup norm of the spectrum of X, i.e.,

$$\|X\|_T = \max_{1 \leqslant i \leqslant n} \lambda_i(X) = \|\lambda(X)\|_\infty,$$

where $\lambda(X) := (\lambda_1(X), \ldots, \lambda_n(X))$, is the vector of ordered eigenvalues of X, counted with multiplicities, whereas Hilbert's seminorm with respect to I_n is the diameter of the spectrum:

$$\|X\|_H = \max_{1 \leqslant i,j \leqslant n} (\lambda_i(X) - \lambda_j(X)) = \Delta(\lambda(X)).$$

3.3 Abstract simplex in the dual space and dual unit ball

We denote by $(\mathscr{X}^\star, \mathbf{e}, \|\cdot\|_T^\star)$ the dual space of $(\mathscr{X}, \mathbf{e}, \|\cdot\|_T)$ where the dual norm $\|\cdot\|_T^\star$ of a continuous linear functional $z \in \mathscr{X}^\star$ is defined by:

$$\|z\|_T^\star := \sup_{\|x\|_T = 1} \langle z, x \rangle.$$

We define the *abstract simplex* in the dual space by:

$$\mathscr{P}(\mathbf{e}) := \{\mu \in \mathscr{C}^\star \mid \langle \mu, \mathbf{e} \rangle = 1\} \ . \tag{3.6}$$

Remark 3.4. For the standard orthant cone (Example 3.2, $\mathscr{X} = \mathbb{R}^n$, $\mathscr{C} = \mathbb{R}^n_+$ and $\mathbf{e} = \mathbf{1}$), the dual space $\mathscr{X}^\star$ is $\mathscr{X} = \mathbb{R}^n$ itself and the dual norm $\|\cdot\|_T^\star$ is the ℓ_1 norm:

$$\|x\|_T^\star = \sum_i |x_i| = \|x\|_1.$$

The abstract simplex $\mathscr{P}(\mathbf{1})$ is the standard simplex in $\mathbb{R}^n$:

$$\mathscr{P}(\mathbf{1}) = \{\nu \in \mathbb{R}^n_+ : \sum_i \nu_i = 1\},$$

i.e., the set of probability measures on the discrete space $\{1, \ldots, n\}$.

Remark 3.5. For the cone of semidefinite matrices (Example 3.3, $\mathscr{X} = \mathrm{S}_n$, $\mathscr{C} = \mathrm{S}_n^+$ and $\mathbf{e} = I_n$), the dual space $\mathscr{X}^\star$ is $\mathscr{X} = \mathrm{S}_n$ itself and the dual norm $\|\cdot\|_T^\star$ is the trace norm:

$$\|X\|_T^\star = \sum_{1 \leqslant i \leqslant n} |\lambda_i(X)| = \|X\|_1, \quad X \in \mathrm{S}_n$$

The simplex $\mathscr{P}(I_n)$ is the set of positive semidefinite matrices with trace 1:

$$\mathscr{P}(I_n) = \{\rho \in \mathrm{S}_+^n : \mathrm{trace}(\rho) = 1\}.$$

The elements of this set are called *density matrices* in quantum physics. They are thought of as noncommutative analogues of probability measure.

We denote by $B_T^\star(\mathbf{e})$ the dual unit ball:

$$B_T^\star(\mathbf{e}) = \{x \in \mathscr{X}^\star \mid \|x\|_T^\star \leqslant 1\} \ .$$

We denote by $\mathrm{conv}(S)$ the convex hull of a set S. The next lemma relates the abstract simplex $\mathscr{P}(\mathbf{e})$ to the dual unit ball $B_T^\star(\mathbf{e})$.

Lemma 3.1. *The dual unit ball $B_T^\star(\mathbf{e})$ of the space $(\mathscr{X}^\star, \mathbf{e}, \|\cdot\|_T^\star)$, satisfies*

$$B_T^\star(\mathbf{e}) = \mathrm{conv}(\mathscr{P}(\mathbf{e}) \cup -\mathscr{P}(\mathbf{e})) \ . \tag{3.7}$$

Proof. For simplicity we write $\mathscr{P}$ instead of $\mathscr{P}(\mathbf{e})$ and $B_T^\star$ instead of $B_T^\star(\mathbf{e})$ in the proof. It follows from (3.5) that

$$\|x\|_T = \sup_{\mu \in \mathscr{P}} |\langle \mu, x\rangle| = \sup_{\mu \in \mathscr{P} \cup -\mathscr{P}} \langle \mu, x\rangle \ . \tag{3.8}$$

Hence $\|z\|_T^\star \leqslant 1$ if and only if,

$$\langle z, x\rangle \leqslant \|x\|_T = \sup_{\mu \in \mathscr{P} \cup -\mathscr{P}} \langle \mu, x\rangle, \ \ \forall x \in \mathscr{X} \ . \tag{3.9}$$

By the strong separation theorem [FHH+01, Thm 3.18], if z did not belong to the closed convex hull $\overline{\mathrm{conv}}(\mathscr{P} \cup -\mathscr{P})$, the closure being understood in the weak star topology of $\mathscr{X}^\star$, there would exist a vector $x \in \mathscr{X}$ and a scalar γ such that

$$\langle z, x\rangle > \gamma \geqslant \langle \mu, x\rangle, \ \ \forall \mu \in \mathscr{P} \cup -\mathscr{P} \ ,$$

contradicting (3.9). Hence,

$$B_T^\star = \overline{\mathrm{conv}}(\mathscr{P} \cup -\mathscr{P}) \ .$$

We claim that the latter closure operation can be dispensed with. Indeed, by the Banach Alaoglu theorem, $B_T^\star$ is weak-star compact. Hence, its subset $\mathscr{P}$, which is weak-star closed, is also weak-star compact. If $\mu \in B_T^\star$, by the characterization of $B_T^\star$ above, μ is a limit, in the weak star topology, of a net $\{\mu_a = s_a \nu_a - t_a \pi_a : a \in \mathscr{A}\}$ with $s_a + t_a = 1$, $s_a, t_a \geqslant 0$ and $\nu_a, \pi_a \in \mathscr{P}$ for all $a \in \mathscr{A}$. By passing to a subnet we can assume that $\{s_a, t_a : a \in \mathscr{A}\}$ converge respectively to $s, t \in [0,1]$ such that $s + t = 1$ and $\{\nu_a, \pi_a : a \in \mathscr{A}\}$ converge respectively to $\nu, \pi \in \mathscr{P}$. It follows that $\mu = s\nu - t\pi \in \mathrm{conv}(\mathscr{P} \cup -\mathscr{P})$. □

Remark 3.6. We make a comparison with [RKW11]. In a finite dimensional setting, Reeb, Kastoryano, and Wolf defined a *base* $\mathscr{B}$ of a proper cone $\mathscr{K}$ in a vector space $\mathscr{V}$ to be a cross section of this cone, i.e., $\mathscr{B}$ is the intersection of the cone $\mathscr{K}$ with a hyperplane given by a linear functional in the interior of the dual cone $\mathscr{K}^\star$. Their vector space $\mathscr{V}$ corresponds to our dual space $\mathscr{X}^\star$, and, since $\mathscr{V}$ is of finite dimension, their dual space $\mathscr{V}^\star$ corresponds to our primal space $\mathscr{X}$. Our cone $\mathscr{C} \subset \mathscr{X}$ corresponds to their dual cone $\mathscr{K}^\star$. Modulo this identification, the base $\mathscr{B}$ can be written precisely as

$$\mathscr{B} = \{\mu \in \mathscr{K} \mid \langle \mu, \mathbf{e}\rangle = 1\} \ ,$$

for some $\mathbf{e}$ in the interior of $\mathscr{K}^\star$, so that the base $\mathscr{B}$ coincides with our abstract simplex $\mathscr{P}(\mathbf{e})$. They defined the *base norm* of $\mu \in \mathscr{V}$ with respect to $\mathscr{B}$ by:

$$\|\mu\|_{\mathscr{B}} = \inf\{\lambda \geqslant 0 : \mu \in \lambda\, \mathrm{conv}(\mathscr{B} \cup -\mathscr{B})\}.$$

They also defined the *distinguishability norm* of $\mu \in \mathscr{V}$ by:

$$\|\mu\|_{\tilde{M}} = \sup_{0 \preccurlyeq x \preccurlyeq \mathbf{e}} \langle \mu, 2x - \mathbf{e} \rangle. \tag{3.10}$$

And Theorem 14 in their paper [RKW11] states that the distinguishability norm is equal to the base norm:

$$\|\mu\|_{\tilde{M}} = \|\mu\|_{\mathscr{B}} \ . \tag{3.11}$$

In a finite dimensional setting, Lemma 3.1 is equivalent to the duality result (3.11) of Reeb et al. and the two approaches are dual to each other.

3.4 Characterization of extreme points of the dual unit ball

The next lemma states that Hilbert's seminorm coincides with the quotient norm on the quotient Banach space $\mathscr{X}/\mathbb{R}\mathbf{e}$.

Lemma 3.2. *For all $x \in \mathscr{X}$, we have:*

$$\|x\|_H = 2 \inf_{\lambda \in \mathbb{R}} \|x + \lambda \mathbf{e}\|_T$$

Proof. The expression

$$\|x + \lambda \mathbf{e}\|_T = \max(M(x/\mathbf{e}) + \lambda, -m(x/\mathbf{e}) - \lambda)$$

is minimal when $M(x/\mathbf{e}) + \lambda = -m(x/\mathbf{e}) - \lambda$. Substituting the value of λ obtained in this way in $\|x + \lambda \mathbf{e}\|_T$, we arrive at the announced formula. □

A standard result [Con90, P.88] of functional analysis shows that if $\mathscr{W}$ is a closed subspace of a Banach space $(\mathscr{X}, \|\cdot\|)$, then the quotient space $\mathscr{X}/\mathscr{W}$ is complete. Besides, the dual of the quotient space $\mathscr{X}/\mathscr{W}$ can be identified isometrically to the space of continuous linear forms on $\mathscr{X}$ that vanish on $\mathscr{W}$, equipped with the dual norm $\|\cdot\|^\star$ of $\mathscr{X}^\star$. Specializing this result to $\mathscr{W} = \mathbb{R}\mathbf{e}$, we get:

Lemma 3.3. *The quotient normed space $(\mathscr{X}/\mathbb{R}\mathbf{e}, \|\cdot\|_H)$ is a Banach space. Its dual is $(\mathscr{M}(\mathbf{e}), \|\cdot\|_H^\star)$ where*

$$\mathscr{M}(\mathbf{e}) := \{\mu \in \mathscr{X}^\star | \langle \mu, \mathbf{e} \rangle = 0\},$$

and

$$\|\mu\|_H^\star := \frac{1}{2}\|\mu\|_T^\star, \ \ \forall \mu \in \mathscr{M}(\mathbf{e}). \tag{3.12}$$

The above lemma implies that the unit ball of the space $(\mathscr{M}(\mathbf{e}), \|\cdot\|_H^\star)$, denoted by $B_H^\star(\mathbf{e})$, satisfies:

$$B_H^\star(\mathbf{e}) = 2B_T^\star(\mathbf{e}) \cap \mathscr{M}(\mathbf{e}). \tag{3.13}$$

Remark 3.7. In the case of standard orthant cone ($\mathscr{X} = \mathbb{R}^n$, $\mathscr{C} = \mathbb{R}^n_+$ and $\mathbf{e} = \mathbf{1}$), Lemma 3.3 implies that for any two probability measures $\mu, \nu \in \mathscr{P}(\mathbf{1})$, the dual norm $\|\mu - \nu\|_H^\star$ is the total variation distance between μ and ν:

$$\|\mu - \nu\|_H^\star = \frac{1}{2}\|\mu - \nu\|_1 = \|\mu - \nu\|_{TV}$$

Before giving a representation of the extreme points of $B_H^\star(\mathbf{e})$, we define a *disjointness* relation $\perp$ on $\mathscr{P}(\mathbf{e})$.

Definition 3.1. For all $\nu,\pi\in\mathscr{P}(\mathbf{e})$, we say that ν and π are *disjoint*, denoted by $\nu\perp\pi$, if

$$\mu=\frac{\nu+\pi}{2}$$

for all $\mu\in\mathscr{P}(\mathbf{e})$ such that $\mu\succcurlyeq\frac{\nu}{2}$ and $\mu\succcurlyeq\frac{\pi}{2}$.

We have the following characterization of the disjointness property.

Lemma 3.4. *Let $\nu,\pi\in\mathscr{P}(\mathbf{e})$. The following assertions are equivalent:*

(a) $\nu\perp\pi$.

(b) The only elements $\rho,\sigma\in\mathscr{P}(\mathbf{e})$ satisfying

$$\nu-\pi=\rho-\sigma$$

are $\rho=\nu$ and $\sigma=\pi$.

Proof. (a)⇒ (b): Let any $\rho,\sigma\in\mathscr{P}(\mathbf{e})$ such that

$$\nu-\pi=\rho-\sigma.$$

Then it is immediate that

$$\nu+\sigma=\pi+\rho.$$

Let $\mu=\frac{\nu+\sigma}{2}=\frac{\pi+\rho}{2}$. Then $\mu\in\mathscr{P}(\mathbf{e})$, $\mu\succcurlyeq\frac{\nu}{2}$ and $\mu\succcurlyeq\frac{\pi}{2}$. Since $\nu\perp\pi$, we obtain that $\mu=\frac{\nu+\pi}{2}$. It follows that $\rho=\nu$ and $\sigma=\pi$.

(b)⇒ (a): Let any $\mu\in\mathscr{P}(\mathbf{e})$ such that $\mu\succcurlyeq\frac{\nu}{2}$ and $\mu\succcurlyeq\frac{\pi}{2}$. Then

$$\nu-\pi=(2\mu-\pi)-(2\mu-\nu).$$

From (b) we know that $2\mu-\pi=\nu$. □

We denote by $\mathrm{extr}(\cdot)$ the set of extreme points of a convex set.

Proposition 3.5. *The set of extreme points of $B_H^\star(\mathbf{e})$, denoted by $\mathrm{extr}\,B_H^\star(\mathbf{e})$, is characterized by:*

$$\mathrm{extr}\,B_H^\star(\mathbf{e})=\{\nu-\pi\mid\nu,\pi\in\mathrm{extr}\,\mathscr{P}(\mathbf{e}),\nu\perp\pi\}.$$

Proof. It follows from (3.7) that every point $\mu\in B_T^\star(\mathbf{e})$ can be written as

$$\mu=s\nu-t\pi$$

with $s+t=1,s,t\geqslant 0$, $\nu,\pi\in\mathscr{P}(\mathbf{e})$. Moreover, if $\mu\in\mathscr{M}(\mathbf{e})$, then

$$0=\langle\mu,\mathbf{e}\rangle=s\langle\nu,\mathbf{e}\rangle-t\langle\pi,\mathbf{e}\rangle=s-t,$$

thus $s=t=\frac{1}{2}$. Therefore every $\mu\in B_T^\star(\mathbf{e})\cap\mathscr{M}(\mathbf{e})$ can be written as

$$\mu=\frac{\nu-\pi}{2},\quad\nu,\pi\in\mathscr{P}(\mathbf{e}).$$

Therefore by (3.13) we proved that

$$B_H^\star(\mathbf{e}) = \{\nu - \pi : \nu, \pi \in \mathscr{P}(\mathbf{e})\}. \tag{3.14}$$

Now let $\nu, \pi \in \operatorname{extr} \mathscr{P}(\mathbf{e})$ and $\nu \perp \pi$. We are going to prove that $\nu - \pi \in \operatorname{extr} B_H^\star(\mathbf{e})$. Let $\nu_1, \pi_1, \nu_2, \pi_2 \in \mathscr{P}(\mathbf{e})$ such that

$$\nu - \pi = \frac{\nu_1 - \pi_1}{2} + \frac{\nu_2 - \pi_2}{2}.$$

Then

$$\nu - \pi = \frac{\nu_1 + \nu_2}{2} - \frac{\pi_1 + \pi_2}{2}.$$

By Lemma 3.4, the only possibility is $2\nu = \nu_1 + \nu_2$ and $2\pi = \pi_1 + \pi_2$. Since $\nu, \pi \in \operatorname{extr} \mathscr{P}(\mathbf{e})$ we obtain that $\nu_1 = \nu_2 = \nu$ and $\pi_1 = \pi_2 = \pi$. Therefore $\nu - \pi \in \operatorname{extr} B_H^\star(\mathbf{e})$.

Now let $\nu, \pi \in \mathscr{P}(\mathbf{e})$ such that $\nu - \pi \in \operatorname{extr} B_H^\star(\mathbf{e})$. Assume by contradiction that ν is not extreme in $\mathscr{P}(\mathbf{e})$ (the case in which π is not extreme can be dealt with similarly). Then, we can find $\nu_1, \nu_2 \in \mathscr{P}(\mathbf{e})$, $\nu_1 \neq \nu_2$, such that $\nu = \frac{\nu_1 + \nu_2}{2}$. It follows that

$$\mu = \frac{\nu_1 - \pi}{2} + \frac{\nu_2 - \pi}{2},$$

where $\nu_1 - \pi, \nu_2 - \pi$ are distinct elements of $B_H^\star(\mathbf{e})$, which is a contradiction. Next we show that $\nu \perp \pi$. To this end, let any $\rho, \sigma \in \mathscr{P}(\mathbf{e})$ such that

$$\nu - \pi = \rho - \sigma.$$

Then

$$\nu - \pi = \frac{\nu - \pi + \rho - \sigma}{2} = \frac{\nu - \sigma}{2} + \frac{\rho - \pi}{2}.$$

If $\sigma \neq \pi$, then $\nu - \sigma \neq \nu - \pi$ and this contradicts the fact that $\nu - \pi$ is extremal. Therefore $\sigma = \pi$ and $\rho = \nu$. From Lemma 3.4, we deduce that $\nu \perp \pi$.

□

Remark 3.8. In the case of standard orthant cone ($\mathscr{X} = \mathbb{R}^n$, $\mathscr{C} = \mathbb{R}^n_+$ and $\mathbf{e} = \mathbf{1}$), the set of extreme points of $\mathscr{P}(\mathbf{1})$ is the set of standard basis vectors $\{e_i\}_{i=1,\ldots,n}$. The extreme points are pairwise disjoint.

Remark 3.9. In the case of cone of semidefinite matrices ($\mathscr{X} = \mathrm{S}_n$, $\mathscr{C} = \mathrm{S}_n^+$ and $\mathbf{e} = I_n$), the set of extreme points of $\mathscr{P}(I_n)$ is

$$\operatorname{extr} \mathscr{P}(I_n) = \{xx^* \mid x \in \mathbb{C}^n, x^*x = 1\} \ ,$$

which are called *pure states* in quantum information terminology. Two extreme points xx^* and yy^* are disjoint if and only if $x^*y = 0$. To see this, note that if $x^*y = 0$ then any Hermitian matrix X such that $X \succcurlyeq xx^*$ and $X \succcurlyeq yy^*$ should satisfy $X \succcurlyeq xx^* + yy^*$. Hence by definition xx^* and yy^* are disjoint. Inversely, suppose that xx^* and yy^* are disjoint and consider the spectral decomposition of the matrix $xx^* - yy^*$, i.e., there is $\lambda \leqslant 1$ and two orthonormal vectors u, v such that $xx^* - yy^* = \lambda(uu^* - vv^*)$. It follows that $xx^* - yy^* = uu^* - ((1-\lambda)uu^* + \lambda vv^*)$. By Lemma 3.4, the only possibility is $yy^* = (1-\lambda)uu^* + \lambda vv^*$ and $xx^* = uu^*$ thus $\lambda = 1$, $u = x$ and $v = y$. Therefore $x^*y = 0$.

3.5 The operator norm induced by Hopf's oscillation seminorm

Consider two real Banach spaces $\mathscr{X}_1$ and $\mathscr{X}_2$. Let $\mathscr{C}_1 \subset \mathscr{X}_1$ and $\mathscr{C}_2 \subset \mathscr{X}_2$ be respectively two closed pointed convex normal cones with non empty interiors $\mathscr{C}_1^0$ and $\mathscr{C}_2^0$. Let $\mathbf{e}_1 \in \mathscr{C}_1^0$ and $\mathbf{e}_2 \in \mathscr{C}_2^0$. Then, we know from Section 3.4 that the two quotient spaces $(\mathscr{X}_1/\mathbb{R}\mathbf{e}_1, \|\cdot\|_H)$ and $(\mathscr{X}_2/\mathbb{R}\mathbf{e}_2, \|\cdot\|_H)$ are Banach spaces. The dual spaces of $(\mathscr{X}_1/\mathbb{R}\mathbf{e}_1, \|\cdot\|_H)$ and $(\mathscr{X}_2/\mathbb{R}\mathbf{e}_2, \|\cdot\|_H)$ are respectively $(\mathscr{M}(\mathbf{e}_1), \|\cdot\|_H^\star)$ and $(\mathscr{M}(\mathbf{e}_2), \|\cdot\|_H^\star)$ (see Lemma 3.3).

Let T denote a continuous linear map from $(\mathscr{X}_1/\mathbb{R}\mathbf{e}_1, \|\cdot\|_H)$ to $(\mathscr{X}_2/\mathbb{R}\mathbf{e}_2, \|\cdot\|_H)$. The operator norm of T, denoted by $\|T\|_H$, is given by:

$$\|T\|_H := \sup_{\|x\|_H=1} \|T(x)\|_H = \sup \frac{\omega(T(x)/\mathbf{e}_2)}{\omega(x/\mathbf{e}_1)} \quad .$$

By definition, the *adjoint operator* $T^\star : (\mathscr{M}(\mathbf{e}_2), \|\cdot\|_H^\star) \to (\mathscr{M}(\mathbf{e}_1), \|\cdot\|_H^\star)$ of T is:

$$\langle T^\star(\mu), x\rangle = \langle \mu, T(x)\rangle, \ \ \forall \mu \in \mathscr{M}(\mathbf{e}_2), x \in \mathscr{X}_1/\mathbb{R}\mathbf{e}_1.$$

The operator norm of $T^\star$, denoted by $\|T^\star\|_H^\star$, is then:

$$\|T^\star\|_H^\star := \sup_{\mu \in B_H^\star(\mathbf{e}_2)} \|T^\star(\mu)\|_H^\star.$$

A classical duality result (see [AB99, § 6.8]) shows that an operator and its adjoint have the same operator norm. In particular,

$$\|T\|_H = \|T^\star\|_H^\star.$$

Theorem 3.2. *Let $T : \mathscr{X}_1 \to \mathscr{X}_2$ be a bounded linear map such that $T(\mathbf{e}_1) \in \mathbb{R}\mathbf{e}_2$. Then,*

$$\|T\|_H = \frac{1}{2} \sup_{\nu,\pi \in \mathscr{P}(\mathbf{e}_2)} \|T^\star(\nu) - T^\star(\pi)\|_T^\star = \sup_{\nu,\pi \in \mathscr{P}(\mathbf{e}_2)} \sup_{x \in [0,\mathbf{e}_1]} \langle \nu - \pi, T(x)\rangle.$$

Moreover, the supremum can be restricted to the set of extreme points:

$$\|T\|_H = \frac{1}{2} \sup_{\substack{\nu,\pi \in \mathrm{extr}\,\mathscr{P}(\mathbf{e}_2) \\ \nu \perp \pi}} \|T^\star(\nu) - T^\star(\pi)\|_T^\star = \sup_{\substack{\nu,\pi \in \mathrm{extr}\,\mathscr{P}(\mathbf{e}_2) \\ \nu \perp \pi}} \sup_{x \in [0,\mathbf{e}_1]} \langle \nu - \pi, T(x)\rangle. \tag{3.15}$$

Proof. We already noted that $\|T\|_H = \|T^\star\|_H^\star$. Moreover,

$$\|T^\star\|_H^\star = \sup_{\mu \in B_H^\star(\mathbf{e}_2)} \|T^\star(\mu)\|_H^\star.$$

By the characterization of $B_H^\star(\mathbf{e}_2)$ in (3.14) and the characterization of the norm $\|\cdot\|_H^\star$ in Lemma 3.3, we get

$$\sup_{\mu \in B_H^\star(\mathbf{e}_2)} \|T^\star(\mu)\|_H^\star = \sup_{\nu,\pi \in \mathscr{P}(\mathbf{e}_2)} \|T^\star(\nu) - T^\star(\pi)\|_H^\star = \frac{1}{2} \sup_{\nu,\pi \in \mathscr{P}(\mathbf{e}_2)} \|T^\star(\nu) - T^\star(\pi)\|_T^\star$$

For the second equality, note that

$$\begin{aligned} \|T^\star(\nu) - T^\star(\pi)\|_T^\star &= \sup_{x \in [0,\mathbf{e}_1]} \langle T^\star(\nu) - T^\star(\pi), 2x - \mathbf{e}_1\rangle \\ &= 2 \sup_{x \in [0,\mathbf{e}_1]} \langle T^\star(\nu) - T^\star(\pi), x\rangle \quad . \end{aligned}$$

We next show that the supremum can be restricted to the set of extreme points. By the Banach-Alaoglu theorem, $B_H^\star(\mathbf{e}_2)$ is weak-star compact, and it is obviously convex. The dual space $\mathscr{M}(\mathbf{e}_2)$ endowed with the weak-star topology is a locally convex topological space. Thus by the Krein-Milman theorem, the unit ball $B_H^\star(\mathbf{e}_2)$, which is a compact convex set in $\mathscr{M}(\mathbf{e}_2)$ with respect to the weak-star topology, is the closed convex hull of its extreme points. So every element ρ of $B_H^\star(\mathbf{e}_2)$ is the limit of a net $(\rho_\alpha)_\alpha$ of elements in $\operatorname{conv}\big(\operatorname{extr} B_H^\star(\mathbf{e}_2)\big)$. Observe now that the function

$$\varphi : \mu \mapsto \|T^\star(\mu)\|_H^\star = \sup_{x\in B_H(\mathbf{e}_1)} \langle T^\star(\mu), x\rangle = \sup_{x\in B_H(\mathbf{e}_1)} \langle \mu, T(x)\rangle$$

which is a sup of weak-star continuous maps is convex and weak-star lower semi-continuous. This implies that

$$\begin{aligned}\varphi(\rho) &\leqslant \liminf_\alpha \varphi(\rho_\alpha)\\ &\leqslant \sup\{\varphi(\mu) : \mu \in \operatorname{conv}\big(\operatorname{extr} B_H^\star(\mathbf{e}_2)\big)\}\\ &= \sup\{\varphi(\mu) : \mu \in \operatorname{extr} B_H^\star(\mathbf{e}_2)\}\ .\end{aligned}$$

Using the characterization of the extreme points in Proposition 3.5, we get:

$$\sup_{\mu\in B_H^\star(\mathbf{e}_2)} \|T^\star(\mu)\|_H^\star = \sup_{\mu\in \operatorname{extr} B_H^\star(\mathbf{e}_2)} \|T^\star(\mu)\|_H^\star = \sup_{\substack{\nu,\pi\in\operatorname{extr}\mathscr{P}(\mathbf{e}_2)\\ \nu\perp\pi}} \|T^\star(\nu)-T^\star(\pi)\|_H^\star. \qquad \square$$

Remark 3.10. When $\mathscr{X}_1$ is of finite dimension, the set $[0,\mathbf{e}_1]$ is the convex hull of the set of its extreme points, hence, the supremum over the variable $x \in [0,\mathbf{e}_1]$ in (3.15) is attained at an extreme point. Similarly, if $\mathscr{X}_2$ is of finite dimension, the suprema over (ν,π) in the same equation are also attained, because the map φ in the proof of the previous theorem, which is a supremum of an equi-Lipschitz family of maps, is continuous (in fact, Lipschitz).

Remark 3.11. Theorem 3.2 should be compared with Proposition 12 of [RKW11] which can be stated as follows.

Proposition 3.6 (Proposition 12 in [RKW11]). *Let $\mathscr{V},\mathscr{V}'$ be two finite dimensional vector spaces and $L:\mathscr{V}\to\mathscr{V}'$ be a linear map and let $\mathscr{B}\subset\mathscr{V}$ and $\mathscr{B}'\subset\mathscr{V}'$ be bases. Then*

$$\sup_{v_1\neq v_2\in\mathscr{B}} \frac{\|L(v_1)-L(v_2)\|_{\mathscr{B}'}}{\|v_1-v_2\|_{\mathscr{B}}} = \frac{1}{2}\sup_{v_1,v_2\in\operatorname{extr}\mathscr{B}} \|L(v_1)-L(v_2)\|_{\mathscr{B}'} \tag{3.16}$$

The first term in (3.16) is called the *contraction ratio* of the linear map L, with respect to base norms (see Remark 3.6). One important applications of this proposition concerns the *base preserving* maps L such that $L(\mathscr{B})\subset\mathscr{B}'$. Let us translate this proposition in the present setting. Consider a linear map $T:\mathscr{X}_1/\mathbb{R}\mathbf{e}_1\to\mathscr{X}_2/\mathbb{R}\mathbf{e}_2$. Then $T^\star:\mathscr{X}_2^\star\to\mathscr{X}_1^\star$ is a base preserving linear map ($T^\star(\mathscr{P}(\mathbf{e}_2))\subset\mathscr{P}(\mathbf{e}_1)$) and so, Proposition 12 of [RKW11] shows that:

$$\sup_{\substack{\nu,\pi\in\mathscr{P}(\mathbf{e}_2)\\ \nu\neq\pi}} \frac{\|T^\star(\nu-\pi)\|_T^\star}{\|\nu-\pi\|_T^\star} = \frac{1}{2}\sup_{\nu,\pi\in\operatorname{extr}\mathscr{P}(\mathbf{e}_2)} \|T^\star(\nu)-T^\star(\pi)\|_T^\star \tag{3.17}$$

Hence, by comparison with [RKW11], the additional information here is the equality between the contraction ratio in Hilbert's seminorm of a unit preserving linear map, and the contraction ratio with respect to the base norms of the dual base preserving map. The latter is the primary object of interest

in quantum information theory whereas the former is of interest in the control/consensus literature. We also proved that the supremum in (3.17) can be restricted to pairs of *disjoint* extreme points ν, π. Finally, the expression of the contraction rate as the last supremum in Theorem 3.2 leads here to an abstract version of Dobrushin's ergodic coefficient, see Eqn (3.21) and Corollary 3.9 below.

Let us recall the definition of Hilbert's projective metric.

Definition 3.2 ([Bir57])**.** *Hilbert's projective metric* between two elements x and y of $\mathscr{C}_0$ is

$$d_H(x,y) = \log(M(x/y)/m(x/y)). \ . \tag{3.18}$$

Consider a linear operator $T : \mathscr{X}_1 \to \mathscr{X}_2$ such that $T(\mathscr{C}_1^0) \subset \mathscr{C}_2^0$. Following [Bir57, Bus73], the *projective diameter* of T is defined as below:

$$\operatorname{diam} T = \sup\{d_H(T(x),T(y)) : x,y \in \mathscr{C}_1^0\}.$$

Birkhoff's contraction formula [Bir57, Bus73] states that the oscillation ratio equals to the contraction ratio of T and they are related to its projective diameter.

Theorem 3.3 ([Bir57, Bus73])**.**

$$\sup_{x,y\in\mathscr{C}_1^0} \frac{\omega(T(x)/T(y))}{\omega(x/y)} = \sup_{x,y\in\mathscr{C}_1^0} \frac{d_H(T(x),T(y))}{d_H(x,y)} = \tanh(\frac{\operatorname{diam} T}{4}).$$

The projective diameter of $T^\star$ is defined by:

$$\operatorname{diam} T^\star = \sup\{d_H(T^\star(u),T^\star(v)) : u,v \in \mathscr{C}_2^\star \backslash 0\}.$$

Note that $\operatorname{diam} T = \operatorname{diam} T^\star$. This is because

$$\begin{aligned}\sup_{x,y\in\mathscr{C}_1^0} \frac{M(T(x)/T(y))}{m(T(x)/T(y))} &= \sup_{x,y\in\mathscr{C}_1^0} \sup_{u,v\in\mathscr{C}_2^\star\backslash 0} \frac{\langle u,T(x)\rangle\langle v,T(y)\rangle}{\langle u,T(y)\rangle\langle v,T(x)\rangle} \\ &= \sup_{u,v\in\mathscr{C}_2^\star\backslash 0} \frac{M(T^\star(u)/T^\star(v))}{m(T^\star(u)/T^\star(v))}\end{aligned}$$

Corollary 3.7 (Compare with [RKW11])**.** *Let $T : \mathscr{X}_1 \to \mathscr{X}_2$ be a bounded linear map such that $T(\mathbf{e}_1) \in \mathbb{R}\mathbf{e}_2$ and $T(\mathscr{C}_1^0) \subset \mathscr{C}_2^0$, then:*

$$\|T^\star\|_H^\star = \|T\|_H \leqslant \tanh(\frac{\operatorname{diam} T}{4}) = \tanh(\frac{\operatorname{diam} T^\star}{4})$$

Proof. It is sufficient to prove the inequality. For this, note that

$$\|T\|_H = \sup_{\substack{x\in\mathscr{X}_1 \\ \omega(x/\mathbf{e}_1)\neq 0}} \omega(T(x)/\mathbf{e}_2)/\omega(x/\mathbf{e}_1) = \sup_{\substack{x\in\mathscr{C}_1^0 \\ \omega(x/\mathbf{e}_1)\neq 0}} \omega(T(x)/\mathbf{e}_2)/\omega(x/\mathbf{e}_1).$$

Then we apply Birkhoff's contraction formula. □

Remark 3.12. Reeb et al. [RKW11] showed in a different way that

$$\|T^\star\|_H^\star \leqslant \tanh(\frac{\operatorname{diam} T^\star}{4}) \ ,$$

in a finite dimensional setting. The proof above shows that as soon as the duality formula $\|T^\star\|_H^\star = \|T\|_H$ has been obtained, the latter inequality follows from Birkhoff's contraction formula.

3.6 Application to discrete consensus operators on cones

A classical result, which goes back to Dœblin and Dobrushin, characterizes the Lipschitz constant of a Markov matrix acting on the space of measures (i.e., a row stochastic matrix acting on the left), with respect to the total variation norm (see the discussion in Section 3.7 below). The same constant characterizes the contraction ratio with respect to the "diameter" (Hilbert's seminorm) of the consensus system driven by this Markov matrix (i.e., a row stochastic matrix acting on the right). Consensus operators on cones extend Markov matrices. In this section, we extend to these abstract operators a number of known properties of Markov matrices.

A bounded linear map $T : \mathscr{X} \to \mathscr{X}$ is a *consensus operator* with respect to a unit vector $\mathbf{e}$ in the interior $\mathscr{C}^0$ of a closed convex pointed cone $\mathscr{C} \subset \mathscr{X}$ if it satisfies the two following properties:

(i) T is positive, i.e., $T(\mathscr{C}) \subset \mathscr{C}$.

(ii) T preserves the unit element $\mathbf{e}$, i.e., $T(\mathbf{e}) = \mathbf{e}$.

A time invariant discrete time consensus system can be described by

$$x_{k+1} = T(x_k), \quad k \in \mathbb{N}, \tag{3.19}$$

The main concern of consensus theory is the convergence of the orbit x_k to a *consensus state*, which is represented by a scalar multiple of the unit element. The case when $\|T\|_H < 1$ or equivalently $\|T^\star\|_H^\star < 1$ is of special interest; the following theorem shows that the iterates of T converge to a rank one projector with a rate bounded by $\|T\|_H$.

Theorem 3.4 (Geometric convergence to consensus)**.** *Let $T : \mathscr{X} \to \mathscr{X}$ be a consensus operator with respect to the unit element $\mathbf{e}$. If $\|T\|_H < 1$ or equivalently $\|T^\star\|_H^\star < 1$, then there is $\pi \in \mathscr{P}(\mathbf{e})$ such that for all $x \in \mathscr{X}$*

$$\|T^n(x) - \langle \pi, x\rangle \mathbf{e}\|_T \leqslant (\|T\|_H)^n \|x\|_H,$$

and for all $\mu \in \mathscr{P}(\mathbf{e})$

$$\|(T^\star)^n(\mu) - \pi\|_H^\star \leqslant (\|T\|_H)^n.$$

Proof. The intersection

$$\cap_n [m(T^n(x)/\mathbf{e}), M(T^n(x)/\mathbf{e})] \subset \mathbb{R}$$

is nonempty (as a non-increasing intersection of nonempty compact sets), and since $\|T\|_H < 1$ and

$$\omega(T^n(x)/\mathbf{e}) \leqslant (\|T\|_H)^n \omega(x/\mathbf{e}),$$

this intersection must be reduced to a real number $\{c(x)\} \subset \mathbb{R}$ depending on x, i.e.,

$$c(x) = \bigcap_n [m(T^n(x)/\mathbf{e}), M(T^n(x)/\mathbf{e})] \ .$$

Thus for all $n \in \mathbb{N}$,

$$-\omega(T^n(x)/\mathbf{e})\mathbf{e} \leqslant T^n(x) - c(x)\mathbf{e} \leqslant \omega(T^n(x)/\mathbf{e})\mathbf{e}.$$

Therefore by definition:

$$\|T^n(x) - c(x)\mathbf{e}\|_T \leqslant \omega(T^n(x)/\mathbf{e}). \leqslant (\|T\|_H)^n \|x\|_H.$$

It is immediate that:

$$c(x)\mathbf{e} = \lim_{n\to\infty} T^n(x)$$

from which we deduce that $c : \mathscr{X} \to \mathbb{R}$ is a continuous linear functional. Thus there is $\pi \in \mathscr{X}^\star$ such that $c(x) = \langle \pi, x \rangle$. Besides it is immediate that $\langle \pi, \mathbf{e} \rangle = 1$ and $\pi \in \mathscr{C}^\star$ because

$$x \in \mathscr{C} \Rightarrow c(x)\mathbf{e} \in \mathscr{C} \Rightarrow c(x) \geqslant 0 \Rightarrow \langle \pi, x \rangle \geqslant 0.$$

Therefore $\pi \in \mathscr{P}(\mathbf{e})$. Finally for all $\mu \in \mathscr{P}(\mathbf{e})$ and all $x \in \mathscr{X}$ we have

$$\begin{aligned} \langle (T^\star)^n(\mu) - \pi, x \rangle &= \langle \mu, T^n(x) - \langle \pi, x \rangle \mathbf{e} \rangle \\ &\leqslant \|\mu\|_T^\star \|T^n(x) - \langle \pi, x \rangle \mathbf{e}\|_T \\ &\leqslant (\|T\|_H)^n \|x\|_H. \end{aligned}$$

Hence

$$\|(T^\star)^n(\mu) - \pi\|_H^\star \leqslant (\|T\|_H)^n.$$

□

A time-dependent consensus system is described by

$$x_{k+1} = T_{k+1}(x_k), \quad k \in \mathbb{N} \tag{3.20}$$

where $\{T_k : k \geqslant 1\}$ is a sequence of consensus operators sharing a common unit element $\mathbf{e} \in \mathscr{C}^0$. Then if there is an integer $p > 0$ and a constant $\alpha < 1$ such that for all $i \in \mathbb{N}$

$$\|T_{i+p} \dots T_{i+1}\|_H \leqslant \alpha,$$

then the same lines of proof of Theorem 3.4 imply the existence of $\pi \in \mathscr{P}(\mathbf{e})$ such that for all $\{x_k\}$ satisfying (3.20),

$$\|x_k - \langle \pi, x_0 \rangle \mathbf{e}\|_T \leqslant \alpha^{\lfloor \frac{k}{p} \rfloor} \|x_0\|_H, \quad n \in \mathbb{N}.$$

Moreover, if $\{T_k : k \geqslant 1\}$ is a stationary ergodic random process, then the almost sure convergence of the orbits of (3.20) to a consensus state can be deduced by showing that

$$\mathbb{E}[\log \|T_{1+p} \dots T_1\|_H] < 0$$

for some $p > 0$, see Bougerol [Bou93]. Hence, in consensus applications, a central issue is to compute the operator norm $\|T\|_H$ of a consensus operator T.

A direct application of Theorem 3.2 leads to following characterization of the operator norm.

Corollary 3.8. *Let $T : \mathscr{X} \to \mathscr{X}$ be a consensus operator with respect to $\mathbf{e}$. Then,*

$$\|T\|_H = \|T^\star\|_H^\star = 1 - \inf_{\substack{\nu, \pi \in \operatorname{extr} \mathscr{P}(\mathbf{e}) \\ \nu \perp \pi}} \inf_{x \in [0, \mathbf{e}]} \langle \pi, T(x) \rangle + \langle \nu, T(\mathbf{e} - x) \rangle.$$

Proof. Since $T(\mathbf{e}) = \mathbf{e}$, we have:

$$\sup_{\substack{\nu, \pi \in \operatorname{extr} \mathscr{P}(\mathbf{e}) \\ \nu \perp \pi}} \sup_{x \in [0, \mathbf{e}]} \langle \nu - \pi, T(x) \rangle = \sup_{\substack{\nu, \pi \in \operatorname{extr} \mathscr{P}(\mathbf{e}) \\ \nu \perp \pi}} \sup_{x \in [0, \mathbf{e}]} 1 - \langle \pi, T(x) \rangle - \langle \nu, T(\mathbf{e} - x) \rangle.$$

□

3.7 Applications to classical linear consensus

In this section, we specialize the previous general results to the case of of the standard orthant cone ($\mathscr{X} = \mathbb{R}^n$, $\mathscr{C} = \mathbb{R}^n_+$ and $\mathbf{e} = \mathbf{1}$, Example 3.2). We recover the classical Dobrushin's ergodicity coefficient and some known convergence results of the consensus system.

A linear map $T : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$T(x) = Ax, \;\; x \in \mathbb{R}^n$$

is a consensus operator if and only if A is a row stochastic matrix. The operator norm corresponds to the contraction ratio of the matrix A with respect to the diameter Δ:

$$\|T\|_H = \tau(A) := \sup_{\Delta(x)\neq 0} \frac{\Delta(Ax)}{\Delta(x)},$$

and the dual operator norm corresponds to the Lipschitz constant of A' with respect to the total variation distance on the space of probability measures:

$$\|T\|^\star_H = \delta(A) := \sup_{\mu\neq\nu\in\mathscr{P}(\mathbf{1})} \frac{\|A'\mu - A'\nu\|_1}{\|\mu-\nu\|_1}$$

The value $\delta(A)$ is known as *Dobrushin's ergodicity coefficient* of the Markov chain with transition probability matrix A', see [LPW09]. Specializing Corollary 3.8 to this case, we get

$$\tau(A) = \delta(A) = 1 - \min_{i\neq j} \min_{I\subset\{1,\dots,n\}} \Big(\sum_{k\in I} A_{ik} + \sum_{k\notin I} A_{jk}\Big) \;\; .$$

The latter formula yields directly the following explicit form of Dobrushin's ergodicity coefficient [Dob56]:

$$\tau(A) = \delta(A) = 1 - \min_{i\neq j} \sum_{s=1}^{n} \min(A_{is}, A_{js}). \tag{3.21}$$

The above equality is a known result in the study of Markov chain. It is known that if $\tau(A) < 1$, then the Markov chain associated to A is ergodic [Sen91].

A (time-invariant) consensus system associated to the matrix A is described by:

$$x_{k+1} = Ax_k, \;\; k \in \mathbb{N} \;\; . \tag{3.22}$$

By Theorem 3.4, if $\tau(A) < 1$, then the consensus system (3.22) converges to a multiple of $\mathbf{1}$ with an exponential rate $\tau(A)$.

Remark 3.13. A simple classical situation in which $\tau(A) < 1$ is when there is a *Dœblin state*, i.e., an element $j \in \{1,\dots,n\}$ such that $A_{ij} > 0$ holds for all $i \in \{1,\dots,n\}$. Besides, a Dœblin state is represented by a node connected to all other nodes in the graph associated to a matrix A. Based on this observation, some graph connectivity conditions [VJAJ05, OT09, Mor05, AB09] characterizing the exponential convergence of consensus systems can be derived directly from Dobrushin's ergodicity coefficient (3.21). As far as we know, such connection between Dobrushin's ergodicity coefficient and the convergence of consensus system has been firstly observed in [MDA05].

Remark 3.14. For example, consider a time-variant linear consensus system:

$$x_{k+1} = A_k x_k, \;\; k \in \mathbb{N} \;\; , \tag{3.23}$$

where $\{A_k\}$ is a sequence of stochastic matrices. Moreau [Mor05] showed that if all the non-zero entries of the matrices $\{A_k\}$ are bounded from below by a positive constant $\alpha > 0$ and if there is $p \in \mathbb{N}$ such that for all $i \in \mathbb{N}$ there is a node connected to all other nodes in the graph associated to the matrix $A_{i+p} \ldots A_{i+1}$, then the system 3.23 is globally uniformly convergent. These two conditions imply exactly that there is a Dœblin state associated to the matrix $A_{i+p} \ldots A_{i+1}$. The uniform bound α is to have an upper bound on the contraction rate, more precisely,

$$\tau(A_{i+p} \ldots A_{i+1}) \leqslant 1 - \alpha, \;\; \forall i = 1, 2, \ldots$$

3.8 Applications to noncommutative consensus

In this section, we specialize the previous general results to a finite dimensional noncommutative space ($\mathscr{X} = \mathrm{S}_n$, $\mathscr{C} = \mathrm{S}_n^+$ and $\mathbf{e} = I_n$, Example 3.3).

A completely positive unital linear map $\Phi : \mathrm{S}_n \to \mathrm{S}_n$ is characterized by a set of matrices $\{V_1, \ldots, V_m\}$ satisfying

$$\sum_{i=1}^{m} V_i^* V_i = I_n \tag{3.24}$$

such that the map Φ is given by:

$$\Phi(X) = \sum_{i=1}^{m} V_i^* X V_i, \quad \forall X \in \mathrm{S}_n \;\; . \tag{3.25}$$

The matrices $\{V_i\}$ are called *Kraus operators*. The dual operator of Φ is given by:

$$\Psi(X) = \sum_{i=1}^{m} V_i X V_i^*, \;\; X \in \mathrm{S}_n \;\; .$$

It is a completely positive and trace-preserving map, called Kraus map. The map Φ and Ψ represent a purely quantum channel [SSR10, RKW11]. The map Φ acts between spaces of measures while the adjoint map Ψ is trace-preserving and acts between spaces of states (density matrices). The operator norm of $\Phi : \mathrm{S}_n / \mathbb{R} I_n \to \mathrm{S}_n / \mathbb{R} I_n$ is the contraction ratio with respect to the diameter of the spectrum:

$$\|\Phi\|_H = \sup_{X \in \mathrm{S}_n} \frac{\lambda_{\max}(\Phi(X)) - \lambda_{\min}(\Phi(X))}{\lambda_{\max}(X) - \lambda_{\min}(X)}.$$

The operator norm of the adjoint map $\Psi : \mathscr{P}(I_n) \to \mathscr{P}(I_n)$ is the contraction ratio with respect to the trace norm (the total variation distance):

$$\|\Psi\|_H^\star = \sup_{\rho_1, \rho_2 \in \mathscr{P}(I_n)} \frac{\|\Psi(\rho_1) - \Psi(\rho_2)\|_1}{\|\rho_1 - \rho_2\|_1}.$$

The values $\|\Phi\|_H$ and $\|\Psi\|_H^\star$ are respectively the noncommutative counterparts of $\tau(\cdot)$ and $\delta(\cdot)$.

Specializing Corollary 3.8 to the case of quantum operations, we obtain the noncommutative version of Dobrushin's ergodic coefficient.

Corollary 3.9. *Let Φ be a completely positive unital linear map defined in* (3.25)*. Then,*

$$\|\Phi\|_H = \|\Psi\|_H^\star = 1 - \min_{\substack{u,v:u^*v=0\\ u^*u=v^*v=1}} \min_{\substack{X=(x_1,\dots,x_n)\\ XX^*=I_n}} \sum_{i=1}^n \min\{u^*\Phi(x_ix_i^*)u, v^*\Phi(x_ix_i^*)v\} \tag{3.26}$$

Proof. It can be easily checked that

$$\operatorname{extr}[0,I_n] = \{P \in \mathrm{S}_n : P^2 = P\}.$$

Hence, Corollary 3.8 and Remark 3.9 yield:

$$\begin{aligned}\|\Phi\|_H = \|\Psi\|_H^\star &= 1 - \min_{\substack{u,v:u^*v=0\\ u^*u=v^*v=1}} \min_{P^2=P} u^*\Phi(I_n-P)u + v^*\Phi(P)v \\ &= 1 - \min_{\substack{u,v:u^*v=0\\ u^*u=v^*v=1}} \min_{\substack{X=(x_1,\dots,x_n)\\ XX^*=I_n}} \min_{J\subset\{1,\dots,n\}} \sum_{i\in J} u^*\Phi(x_ix_i^*)u + \sum_{i\notin J} v^*\Phi(x_ix_i^*)v\end{aligned}$$

from which (3.26) follows. □

Remark 3.15. For the noncommutative case, it is not evident whether more effective characterization of the contraction rate exists. Note that the dual operator norm was studied in quantum information theory, see [RKW11] and references therein. They provided a Birkhoff type upper bound (Corollary 9 in [RKW11]):

$$\|\Psi\|_H^* \leqslant \tanh(\operatorname{diam}\Psi/4) \ .$$

The value $\operatorname{diam}\Psi$ is not directly computable. This upper bound is equal to 1 if and only if $\operatorname{diam}\Psi = \infty$, which is satisfied if and only if there exist a pair of nonzero vectors $u, v \in \mathbb{C}^n$ such that:

$$\operatorname{span}\{V_iu : 1 \leqslant i \leqslant m\} \neq \operatorname{span}\{V_iv : 1 \leqslant i \leqslant m\}.$$

We next provide a much tighter, in fact necessary and sufficient, condition for the operator norm to be 1.

Corollary 3.10. *The following conditions are equivalent:*

1. $\|\Phi\|_H = \|\Psi\|_H^\star = 1$.
2. *There are nonzero vectors $u, v \in \mathbb{C}^n$ such that*

$$\langle V_iu, V_jv\rangle = 0, \ \ \forall i,j \in \{1,\dots,m\}.$$

3. *There is a rank one matrix $Y \subset \mathbb{C}^{n\times n}$ such that*

$$\operatorname{trace}(V_i^*V_jY) = 0, \ \ \forall i,j \in \{1,\cdots,m\}.$$

Proof. From Corollary 3.9 we know that $\|\Phi\|_H = 1$ if and only if there exist an orthonormal basis $\{x_1,\dots,x_n\}$ and two vectors $u, v \in \mathbb{C}^n$ of norm 1 such that

$$\sum_{i=1}^n \min\{\sum_{j=1}^m u^*V_j^*x_ix_i^*V_ju, \sum_{j=1}^m v^*V_j^*x_ix_i^*V_jv\} = 0 \ .$$

This is equivalent to that for each $i \in \{1,\dots,n\}$, either

$$x_i^* V_j u = 0, \ \ \forall j = 1,\dots,m$$

is true, or

$$x_i^* V_j v = 0, \ \ \forall j = 1,\dots,m$$

is true. This is equivalent to

$$\langle V_i u, V_j v\rangle = 0, \ \ \forall i,j \in \{1,\dots,m\} \ .$$

The equivalence between the second and the third condition is trivial by taking $Y = vu^*$. □

3.8.1 Convergence condition of noncommutative consensus system

We consider a time-invariant noncommutative consensus system:

$$X_{t+1} = \Phi(X_t), \ \ X_t \in \mathrm{S}_n, \ \ t = 1,2,\dots \tag{3.27}$$

where Φ is a completely positive unital map. To study the convergence of such system, Sepulchre, Sarlette and Rouchon [SSR10] proposed to study the contraction ratio

$$\alpha := \sup_{X \succ 0} d_H(\Phi(X), I_n)/d_H(X, I_n) \ .$$

They applied Birkhoff's contraction formula (Theorem 3.3) to give an upper bound on the contraction ratio α:

$$\alpha \leqslant \tanh(\operatorname{diam}\Phi/4) \ .$$

The following theorem is a direct corollary of Nussbaum [Nus94].

Theorem 3.5. *(Corollary of [Nus94, Thm2.3])*

$$\|\Phi\|_H = \lim_{\varepsilon \to 0^+} \big(\sup\{\frac{d_H(\Phi(X), I_n)}{d_H(X, I_n)} : 0 < d_H(X, I_n) \leqslant \varepsilon\}\big),$$

By this theorem, it is clear that the contraction ratio used in [SSR10] is an upper bound of the operator norm $\|\Phi\|_H$:

$$\|\Phi\|_H \leqslant \alpha \ .$$

We next provide an algebraic characterization of the global convergence of system (3.27). Let us consider a sequence of matrix subspaces defined by:

$$H_0 = \operatorname{span}\{I_n\}, \ \ H_{k+1} = \operatorname{span}\{V_i^* X V_j : X \in H_k, i,j = 1,\dots,m\} \ , k = 0,1,\dots, \tag{3.28}$$

Lemma 3.11. *There is $k_0 \leqslant n^2 - 1$ such that*

$$H_{k_0+s} = H_{k_0}, \ \ \forall s \in \mathbb{N}.$$

Proof. It follows from (3.24) that $H_{k+1} \supseteq H_k$ for all $k \in \mathbb{N}$. Besides, if for some $k_0 \in \mathbb{N}$ such that

$$H_{k_0+1} = H_{k_0} \ ,$$

then

$$H_{k_0+s} = H_{k_0}, \ \ \forall s \in \mathbb{N}.$$

This property also implies that if for some $k_0 \in \mathbb{N}$

$$H_{k_0+1} \neq H_{k_0} \ ,$$

then

$$H_{k_0-s+1} \neq H_{k_0-s} \ , \forall 1 \leqslant s \leqslant k_0 \ .$$

Since the dimension of H_k can not exceed n^2, the case

$$H_{k_0+1} \neq H_{k_0} \ ,$$

can not happen more than n^2 times. □

For all $k \in \mathbb{N}$, let G_k be the orthogonal complement of H_k. Then there is $k_0 \leqslant n^2 - 1$ such that

$$G_k \supseteq G_{k+1}, \ \forall k \in \mathbb{N}; \quad G_{k_0} = G_{k_0+s}, \ \forall s \in \mathbb{N} \tag{3.29}$$

Theorem 3.6. *The following conditions are equivalent:*

(1) There exists k such that $\|\Phi^k\|_H < 1$.

(2) Every orbit of the system (3.27) *converges to an equilibrium co-linear to* I_n.

(3) The subspace $\cap_k G_k$ *does not contain a rank one matrix.*

(4) There exists $k_0 \leqslant n^2 - 1$ *such that* $\|\Phi^{k_0}\|_H < 1$.

Proof. (1) $\Rightarrow$ (2): We apply Theorem 3.4 to the application Φ^k.

(2) $\Rightarrow$ (1): Hilbert's seminorm defines a norm in the orthogonal space to the identity matrix I_n. It follows from Gelfand's formula that

$$\lim_{k \to +\infty} \|\Phi^k\|_H^{1/k} = \max\{|\lambda| : \Phi(X + X^*) = \lambda X + \lambda^* X^*, X \perp I_n\}$$

Thus if (1) is not true, then there is $X \in \mathrm{S}_n$ and $\lambda \in \mathbb{C}$ such that $|\lambda| = 1$, $X \perp I$ and $\Phi(X) = \lambda X$. The system is therefore not globally convergent to an equilibrium co-linear to I_n.

(3) $\Leftrightarrow$ (1): Note that for all $k \in \mathbb{N}$,

$$\Phi^k(X) = \sum_{i_1,\ldots,i_k} V_{i_k}^* \ldots V_{i_1}^* X V_{i_1} \ldots V_{i_k}.$$

By Corollary 3.10, we know that $\|\Phi^k\|_H = 1$ if and only if the subspace G_k contains a a rank one matrix. Therefore , $\|\Phi^k\|_H = 1$ for all $k \in \mathbb{N}$ if and only if the subspace $\cap_k G_k$ contains a rank one matrix.

(3) $\Rightarrow$ (4): By (3.29), there is $k_0 \leqslant n^2 - 1$ such that $G_{k_0} = \cap_k G_k$. It follows that if (3) is true then there is $k_0 \leqslant n^2 - 1$ such that G_{k_0} does not contain a rank one matrix. Then by Corollary 3.10 we deduce that $\|\Phi^{k_0}\| < 1$ if (3) is true. □

Remark 3.16. A sufficient condition for the global convergence of the system (3.27) would be that there is $k_0 \leqslant n^2 - 1$ such that

$$H_{k_0} = \mathbb{C}^{n \times n}.$$

Such condition can be checked in polynomial time.

3.8.2 Irreducibility, primitivity and a complexity result

In this subsection, we first recall the definition of irreducibility and primitivity for the completely positive unital map Φ. Then we show that the global convergence of system (3.27) is equivalent to the primitivity of Φ if Φ is irreducible.

We denote by $S_k(\Phi)$ the linear space spanned by all the products of k Kraus operators $\{V_1,\dots,V_m\}$, and by $D_k(\Phi)$ the linear space spanned by all the products of at most k Kraus operators. We denote by $\mathscr{A}(\Phi)=\cup_{k\geqslant 1}D_k(\Phi)$ the algebra generated by the Kraus operators $\{V_1,\dots,V_m\}$.

Lemma 3.12. *There is $p\leqslant n^2$ such that $\mathscr{A}(\Phi)=D_p(\Phi)$.*

Proof. We know that

$$D_{k+1}\supset D_k\cup\{V_iX : X\in D_k, i=1,\dots,m\}\ \ ,k=1,2,\dots,$$

The remaining part of the proof is identical to that of Lemma 3.11. □

We next give some definitions analogous to the standard nonnegative matrix case. Usually, they are given for the Kraus map Ψ. It is equivalent to define them for the unital map Φ.

Definition 3.3 (Irreducibility [Far96])**.** The map Φ is irreducible if there is no face of S_n^+ invariant by Φ, where a face $\mathscr{F}$ of S_n^+ is a closed cone strictly contained in S_n^+ such that if $P\in\mathscr{F}$ then $[0,P]\in\mathscr{F}$.

Definition 3.4 (Strict positivity)**.** The map Φ is strictly positive if for all $X\succcurlyeq 0$, $\Phi(X)\succ 0$.

Definition 3.5 (Primitivity [SPGWC10])**.** The map Φ is primitive if there is $p>0$ such that Φ^p is strictly positive.

Proposition 3.13. *The map Φ is irreducible if and only if the algebra $\mathscr{A}(\Phi)$ is the whole $n\times n$ matrix algebra.*

Proof. It was shown [Far96, Theorem 2] that the reducibility is equivalent to the existence of a non-trivial (other than $\{0\}$ or $\mathbb{C}^n$) common invariant subspace of all $\{V_i\}$. By Burnside's theorem on matrix algebra (see [LR04]), the latter property holds if and only if the algebra $\mathscr{A}(\Phi)$ is not the whole matrix space. □

The following proposition is a noncommutative analogue of the property that a irreducible matrix is primitive if and only if it is aperiodic.

Proposition 3.14. *If Φ is irreducible, then the system* (3.27) *is globally convergent if and only if Φ is primitive.*

Proof. If Φ is primitive then there is k such that the unital map Φ^k is strictly positive. By Corollary 3.9, this implies that $\|\Phi^k\|_H<1$. Then we use Theorem 3.6 to obtain the global convergence of system (3.27).

Inversely, if there is $k>0$ such that $\alpha=\|\Phi^k\|_H<1$ then by Theorem 3.4, there is a unique invariant density matrix $\Pi\in\mathscr{P}(I_n)$ of Ψ^k such that for all $P\in\mathscr{P}(I_n)$,

$$\|\Psi^{nk}(P)-\Pi\|_H^\star\leqslant\alpha^n,\ \ \forall n\geqslant 0.$$

Since

$$\Psi^k(\Pi)=\Pi,$$

we know that

$$\Psi^k(\Psi(\Pi)) = \Psi(\Pi) \ .$$

Since Ψ^k has only one invariant matrix, it follows that Π is also the unique invariant density matrix of Ψ. We deduce that Π is of full rank by the irreducibility of Ψ. Again by Theorem 3.4, for all $x \in \mathbb{C}^n$ with norm equal to 1,

$$\|\Phi^{nk}(xx^*) - (x^*\Pi x)I_n\|_T \leqslant \alpha^n.$$

That is,

$$-\alpha^n I_n \leqslant \Phi^{nk}(xx^*) - (x^*\Pi x)I_n \leqslant \alpha^n I_n$$

Since Π is a density matrix of full rank, there is n_0 such that for all $x \in \mathbb{C}^n$,

$$\Phi^{n_0 k}(xx^*) \succcurlyeq \big(\lambda_{\min}(\Pi) - \alpha^{n_0}\big) I_n > 0.$$

Thus Φ is primitive. □

We shall use a characterization of primitivity given in [SPGWC10].

Theorem 3.7 ([SPGWC10])**.** *The unital completely positive map Φ is primitive if and only if there is $q \leqslant (n^2 - m + 1)n^2$ such that the space $S_q(\Phi)$ is of dimension n^2.*

Corollary 3.15. *Let Φ be a unital completely positive map determined by rational Kraus operators. Then checking whether Φ is irreducible can be done in polynomial time. If Φ is irreducible, then checking whether the system* (3.27) *is globally convergent can be done in polynomial time.*

Proof. To decide if Φ is irreducible, we shall compute the increasing sequence of matrix subspaces $D_s(\Phi)$, $s = 1, 2, \ldots$, and look for the first integer $k \leqslant n^2$ such that $D_k(\Phi) = D_{k+1}(\Phi)$. For a given s, we shall represent $D_s(\Phi)$ by a basis, i.e.,

$$D_s(\Phi) = \text{span}\{M_1, \cdots, M_l\}$$

where $M_i \in \mathbb{C}^{n\times n}$ are linearly independent matrices. Recall that extracting a basis from a family of rational vectors can be done in polynomial time in the bit model. It follows that the number of bits needed to code the basis elements remain polynomially bounded in the length of the input. Hence, a basis representation of the algebra $\mathscr{A}(\Phi)$ can be obtained in polynomial time.

Suppose now that Φ is irreducible. Then by Proposition 3.14, deciding if the system (3.27) is globally convergent reduces to checking whether Φ is primitive. By Theorem 3.7, Φ is primitive if and only if $S_q(\Phi)$ is of dimension n^2 for some $q \leqslant (n^2 - m + 1)n^2$. Arguing as above, a basis representation of $S_q(\Phi)$ can be computed in polynomial time. □

Remark 3.17. A natural method to decide whether Φ^k is a contraction for k large enough, would be to check whether 1 is the only eigenvalue of Φ on the unit circle and if it is algebraically simple. However, doing so in exact arithmetic appear to be not so tractable, whereas Corollary 3.15 leads to a polynomial algorithm in the bit model, when the map Φ is irreducible. When Φ is reducible, we do not know the algorithmic complexity of checking the conditions of Theorem 3.6. The application to noncommutative consensus over infinite dimensional Hilbert spaces also remains to be developed.

3.8.3 NP-hardness of deciding the strict positivity of a Kraus map

We have seen in Section 3.8.1 that deciding if $\|\Phi\|_H < 1$ and if the noncommutative consensus system 3.27 is globally convergent can both be reduced to finding a rank one matrix in certain matrix subspaces (Corollary 3.10 and Theorem 3.6). In this section, we study the complexity of deciding if a matrix subspace contains a rank one matrix. Our main result shows that deciding if there is a rank one matrix orthogonal to a given subspace of $\mathbb{C}^{n\times n}$ is NP-hard, even if this space is given as the linear span of the matrices arising in the representation of a Kraus map. Formally, we consider the following problem.

Problem 3.1 (Rank one matrix)**.** *Input:* integers n,m, and matrices $V_1,\dots,V_m \subset \mathbb{C}^{n\times n}$ with rational entries, satisfying

$$\sum_{i=1}^{m} V_i^* V_i = I_n \ .$$

Question: is there a rank one matrix in the orthogonal complement of the subspace of $\mathbb{C}^{n\times n}$ spanned by $\{V_1,\dots,V_m\}$?

Theorem 3.8. *The 3SAT problem is reducible in polynomial time to Problem 3.1.*

The proof is based on the following remark. An instance of 3SAT problem with N Boolean variables $X_1,\dots,X_N$ and M clauses can be coded by a system of polynomial equations in N complex variables $x_1,\dots,x_N$,

$$\begin{cases} (1+p_i x_{k_i^1})(1+q_i x_{k_i^2})(1+r_i x_{k_i^3}) = 0, \quad i = 1,\cdots,M \\ x_i^2 = 1, \quad i = 1,\cdots,N \end{cases} \tag{3.30}$$

where $k_i^1,k_i^2,k_i^3 \in \{1,\dots,N\}$, $p_i,q_i,r_i \in \{\pm 1\}$ and $k_i^1 \neq k_i^2 \neq k_i^3$ for all $1 \leqslant i \leqslant M$. The Boolean variable X_i is true if $x_i = 1$ and false if $x_i = -1$. For instance, the clause $X_1 \vee \neg X_2 \vee X_4$ corresponds to the polynomial $(1-x_1)(1+x_2)(1-x_4)$ and the clause $\neg X_6 \vee \neg X_1 \vee X_2$ corresponds to the polynomial $(1+x_6)(1+x_1)(1-x_2)$.

Therefore, to prove Theorem 3.8, it is sufficient to construct in polynomial time a set of Kraus operators $\{V_1,\dots,V_m\} \subset \mathbb{C}^n$ satisfying

$$\sum_{i=1}^{m} V_i^* V_i = I_n$$

such that there is a solution to (3.30) if and only if there are two nonzero vectors $x,y \in \mathbb{C}^n$ such that

$$x^* V_i y = 0, \ \ \forall i = 1,\dots,m \ .$$

We begin by the following basic lemma.

Lemma 3.16. *Let $a_k(\cdot,\cdot) : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$, $1 \leqslant k \leqslant M$ be a finite set of bilinear forms. There is a solution $x \in \mathbb{C}^n$ to the system*

$$a_k(x,x) = 0, \qquad 1 \leqslant k \leqslant M$$

if and only if there is a pair of non-zero vectors $x = (x_i)_{1\leqslant i\leqslant n}, y = (y_i)_{1\leqslant i\leqslant n} \in \mathbb{C}^n$ satisfying the system

$$\begin{cases} a_k(x,y) = 0, \qquad 1 \leqslant k \leqslant M \\ x_i y_j - x_j y_i = 0, \qquad 1 \leqslant i < j \leqslant n \ . \end{cases} \tag{3.31}$$

It is sufficient to note that the second equations require that x be proportional to y.

The next lemma shows that system (3.30) can be transformed into a set of homogeneous equations.

Lemma 3.17. *Let $N,M \in \mathbb{N}$. Let $(k_i^1)_i,(k_i^2)_i,(k_i^3)_i$ be three sequences of integers in $\{1,\cdots,N\}$. Let $(p_i)_i,(q_i)_i,(r_i)_i$ be three sequences of real numbers. Consider the following system of equations on the variables $(x_i)_{1\leqslant i\leqslant N}$:*

$$\begin{cases} (1+p_i x_{k_i^1})(1+q_i x_{k_i^2})(1+r_i x_{k_i^3}) = 0, \quad i=1,\cdots,M \\ x_i^2 = 1, \quad i=1,\cdots,N \end{cases} \tag{3.32}$$

The system (3.32) *has a solution $x \in \mathbb{C}^N$ if and only if there is a pair of nonzero vectors $x=(x_i)_{0\leqslant i\leqslant N+2M}, y=(y_i)_{0\leqslant i\leqslant N+2M} \in \mathbb{C}^{N+2M+1}$ satisfying the following system:*

$$\begin{cases} (x_0 + p_i x_{k_i^1} + q_i x_{k_i^2} + p_i q_i x_{N+i}) y_{N+M+i} = 0, \quad i=1,\cdots,M \\ x_{k_i^1} y_{k_i^2} - x_0 y_{N+i} = 0, \quad i=1,\cdots,M \\ (x_0 + r_i x_{k_i^3} - x_{N+M+i}) y_j = 0, \quad i=1,\cdots,M, \quad j=0,\ldots,N+2M \\ x_i y_i - x_0 y_0 = 0, \quad i=1,\cdots,N+M \\ x_i y_j - x_j y_i = 0, \qquad 0 \leqslant i < j \leqslant N+2M \end{cases} \tag{3.33}$$

Proof. A simple rewriting of the system (3.32) is:

$$\begin{cases} (1+p_i x_{k_i^1} + q_i x_{k_i^2} + p_i q_i x_{k_i^1} x_{k_i^2})(1+r_i x_{k_i^3}) = 0, \quad i=1,\cdots,M \\ x_i^2 = 1, \quad i=1,\cdots,N \end{cases} \tag{3.34}$$

By introducing $2M$ extra variables, denoted by $\{x_{N+i}\}_{1\leqslant i\leqslant 2M}$, to replace the variables $\{x_{k_i^1}x_{k_i^2}, 1+r_i x_{k_i^3}\}_{i\leqslant M}$, we rewrite the system (3.34) as:

$$\begin{cases} (1+p_i x_{k_i^1} + q_i x_{k_i^2} + p_i q_i x_{N+i}) x_{N+M+i} = 0, \quad i=1,\cdots,M \\ x_{k_i^1} x_{k_i^2} - x_{N+i} = 0, \quad i=1,\cdots,M \\ 1 + r_i x_{k_i^3} - x_{N+M+i} = 0, \quad i=1,\cdots,M \\ x_i^2 = 1, \quad i=1,\cdots,N+M \end{cases} \tag{3.35}$$

We next add an extra variable x_0 to replace the affine term 1 to construct a system of homogeneous polynomial equations of degree 2:

$$\begin{cases} (x_0 + p_i x_{k_i^1} + q_i x_{k_i^2} + p_i q_i x_{N+i}) x_{N+M+i} = 0, \quad i=1,\cdots,M \\ x_{k_i^1} x_{k_i^2} - x_0 x_{N+i} = 0, \quad i=1,\cdots,M \\ (x_0 + r_i x_{k_i^3} - x_{N+M+i}) x_j = 0, \quad i=1,\cdots,M, \quad j=0,\ldots,N+2M \\ x_i^2 - x_0^2 = 0, \quad i=1,\cdots,N+M \end{cases} \tag{3.36}$$

Then that there is a solution to (3.35) if and only if there is a solution $x=(x_i)_{0\leqslant i\leqslant N+2M}$ to (3.36) such that $x_0 \neq 0$. By Lemma 3.16, we know that the system (3.36) has a solution $x=(x_i)_{0\leqslant i\leqslant N+2M}$ with $x_0 \neq 0$ if and only if there is a pair of non-null vectors $x=(x_i)_{0\leqslant i\leqslant N+2M}$ and $y=(y_i)_{0\leqslant i\leqslant N+2M}$ with $x_0 y_0 \neq 0$ satisfying (3.33).

So far, we proved that there is a solution to (3.32) if and only if there is a pair of nonzero vectors $x,y \in \mathbb{C}^{N+2M+1}$ satisfying (3.33) such that $x_0 y_0 \neq 0$. We next prove by contradiction that all nonzero pair of solutions to (3.33) satisfy $x_0 y_0 \neq 0$.

Let $x=(x_i)_{0\leqslant i\leqslant N+2M}$ and $y=(y_i)_{0\leqslant i\leqslant N+2M}$ be a pair of nonzero solutions to (3.33) such that $x_0 y_0 = 0$. Since by the last constraint in (3.33), x and y are proportional to each other, we know that

$x_0 = y_0 = 0$. Suppose that there is $1 \leqslant i_0 \leqslant N+M$ such that $x_{i_0} \neq 0$, then by the fourth equation of (3.33) we know that:

$$x_{i_0} y_{i_0} = 0,$$

thus $y_{i_0} = 0$. This implies that y is a zero vector because x and y are proportional to each other. Hence $x_i = 0$ for all $i \leqslant N+M$. Now we apply this condition to the third equation in (3.33) to obtain:

$$x_{N+M+i} y_j = 0, \quad i = 1,\dots,M, \;\; j = 0,\dots,N+2M \;\; .$$

If x is a nonzero vector, necessarily there is i_0 such that $x_{N+M+i_0} \neq 0$, in that case y is a zero vector. Therefore we deduce that for all nonzero solution of (3.33), it is necessary that $x_0 y_0 \neq 0$.

□

Lemma 3.18. *Consider the system* (3.32) *in Lemma 3.17. We suppose in addition that $k_i^1 \neq k_i^2$ for all $1 \leqslant i \leqslant M$ and that $(p_i)_i, (q_i)_i, (r_i)_i$ are sequences of numbers in $\{\pm 1\}$. Let $n = N+2M+1$. There are a finite number of matrices $\{V_i\}_{1 \leqslant i \leqslant m} \subset \mathscr{C}^{n \times n}$ with entries in $\{0, \pm 1, \pm \frac{1}{3}\}$ such that the system* (3.32) *has a solution if and only if there is a rank one matrix in the matrix subspace:*

$$\mathscr{B} := \{X \in \mathbb{C}^{n \times n} : \operatorname{trace}(V_i X) = 0, \quad \forall i = 1,\dots,m\}.$$

Besides, the integer m can be bounded by a polynomial en N and M and the matrices $\{V_i\}_{1 \leqslant i \leqslant m}$ satisfy:

$$\sum_{i=1}^{m} V_i^* V_i = (2N+7M+4)^2 I_n$$

Proof. We denote by $\{e_i\}_{0 \leqslant i \leqslant N+2M}$ the standard basis vectors in $\mathbb{C}^{N+2M+1}$. We know from Lemma 3.17 that the system (3.32) admits a solution if and only if there is a pair of non-null vectors $x, y \in \mathbb{C}^n$ satisfying

$$\begin{cases} x^T (e_0 + p_i e_{k_i^1} + q_i e_{k_i^2} + p_i q_i e_{N+i}) e_{N+M+i}^T y = 0, \quad i = 1,\cdots,M \\ x^T (e_{k_i^1} e_{k_i^2}^T - e_0 e_{N+i}^T) y = 0, \quad i = 1,\cdots,M \\ x^T (e_0 + r_i e_{k_i^3} - e_{N+M+i}) e_j^T y = 0, \quad i = 1,\cdots,M, \quad j = 0,\dots,N+2M \\ x^T (e_i e_i^T - e_0 e_0^T) y = 0, \quad i = 1,\cdots,N+M \\ x^T (e_i e_j^T - e_j e_i^T) y = 0, \qquad 0 \leqslant i < j \leqslant N+2M \end{cases} \tag{3.37}$$

The system (3.37) has $3N+8M+1$ bilinear equations. Let $m_0 = 3N+8M+1$ and denote by $\{A_i\}_{1 \leqslant i \leqslant m_0}$ the matrices corresponding to the m_0 bilinear forms in (3.37). Recall that $(p_i)_i, (q_i)_i, (r_i)_i$ are sequences with numbers in $\{1,-1\}$. Therefore we transformed the system (3.32) to finding a rank one matrix in the matrix subspace given by:

$$\mathscr{B}_0 := \{X \in \mathbb{C}^{n \times n} : \operatorname{trace}(A_i X) = 0, \quad \forall i = 1,\dots,m_0\}.$$

where A_i have entries in $\{0,1,-1\}$. We check the five lines in (3.37) and obtain that

$$\begin{aligned} \sum_{i=1}^{m_0} A_i^* A_i = \;& \sum_{i=1}^{M} 4 e_{N+M+i} e_{N+M+i}^T + \sum_{i=1}^{M} (e_{k_i^2} e_{k_i^2}^T + e_{N+i} e_{N+i}^T) \\ & + \sum_{i=1}^{M} \sum_{j=0}^{N+2M} 3 e_j e_j^T + \sum_{i=1}^{N+M} (e_i e_i^T + e_0 e_0^T) \\ & + \sum_{i<j} (e_j e_j^T + e_i e_i^T) \end{aligned}$$

Therefore we have that

$$\sum_{i=1}^{m_0} A_i^* A_i = \begin{pmatrix} k_1 & & & \\ & k_2 & & \\ & & \ddots & \\ & & & k_n \end{pmatrix}$$

where $k_i \leqslant 2N+7M+4$ for all $1 \leqslant i \leqslant n$. Remark that due to the third line of equations in (3.37), for each $0 \leqslant j \leqslant N+2M$, there is an integer $1 \leqslant n_j \leqslant m_0$ such that

$$A_{n_j}^* A_{n_j} = 3 e_j e_j^T.$$

By letting $B_j = A_{n_j}/3$ we get that:

$$3 B_j^* B_j = e_j e_j^T.$$

For all $1 \leqslant j \leqslant n$ let $l_j = (2N+7M+4)^2 - n_j$. Let $m = m_0 + 3\sum_{j=1}^{n} l_j$ and $\{V_i\}_{1 \leqslant i \leqslant m}$ be the sequence of matrices containing $\{A_i\}_{1 \leqslant i \leqslant m_0}$ and $3l_j$ times the matrix B_j for all $1 \leqslant j \leqslant n$. Then we have

$$\sum_{i=1}^{m} V_i^* V_i = \sum_{i=1}^{m_0} A_i^* A_i + \sum_{j=1}^{n} 3 l_j B_j^* B_j = (2N+7M+4)^2 I_n.$$

Since for all j, B_j is co-linear to a matrix in $\{A_i\}_{i \leqslant m_0}$. The matrix subspace given by

$$\mathscr{B} := \{X \in \mathbb{C}^{n \times n} : \operatorname{trace}(V_i X) = 0, \quad \forall i = 1, \ldots, m\}$$

is equal to $\mathscr{B}_0$. Thus the system (3.32) admits a solution if and only if there is a rank one matrix in $\mathscr{B}$. □

We now give a proof for Theorem 3.8.

Proof. Let $k_i^1, k_i^2, k_i^3 \in \{1, \ldots, N\}$, $p_i, q_i, r_i \in \{\pm 1\}$ and $k_i^1 \neq k_i^2 \neq k_i^3$ for all $1 \leqslant i \leqslant M$. Let

$$\begin{cases} (1 + p_i x_{k_i^1})(1 + q_i x_{k_i^2})(1 + r_i x_{k_i^3}) = 0, & i = 1, \cdots, M \\ x_i^2 = 1, & i = 1, \cdots, N \end{cases} \tag{3.38}$$

be a system corresponding to an instance of 3SAT problem with N Boolean variables and M clauses. By Lemma 3.18, we can construct in polynomial time (with respect to N and M) a sequence of $n \times n$ matrices $\{V_i\}_{1 \leqslant i \leqslant m}$ with entries in $\{0, \pm\frac{1}{l}, \pm\frac{1}{3l}\}$ where $l = (2N+7M+4)$ such that there is a solution to (3.38) if and only if there is a rank one matrix in the subspace

$$\mathscr{B} := \{X \in \mathbb{C}^{n \times n} : \operatorname{trace}(V_i X) = 0, \quad i = 1, \ldots, m\}.$$

Besides, the matrices $\{V_i\}_{1 \leqslant i \leqslant m}$ satisfy

$$\sum_{i=1}^{m} V_i^* V_i = I_n.$$

□

Lemma 3.19. *A completely positive map $\Phi : \mathrm{S}_n^+ \to \mathrm{S}_n^+$ given by:*

$$\Phi(X) = \sum_{i=1}^{m} V_i^* Y V_i \tag{3.39}$$

is strictly positive if and only if there do not exist two nonzero vectors $x, y \in \mathbb{C}^n$ such that

$$x^* V_i y = 0, \ \ \forall i = 1, \ldots, m.$$

Proof. By definition, the map Φ is strictly positive if and only if for all nonzero vector $x \in \mathbb{C}^n$, the matrix

$$\Phi(xx^*) = \sum_{i=1}^{m} V_i^* xx^* V_i$$

is positive definite. This is equivalent to that for all nonzero vector $y \in \mathbb{C}^n$,

$$\sum_{i=1}^{m} y^* V_i^* xx^* V_i y = \sum_{i=1}^{n} (x^* V_i y)^2 > 0.$$

Therefore Φ is not strictly positive if and only if we can find nonzero vectors $x, y \in \mathbb{C}^n$ such that

$$x^* V_i y = 0, \ \ \forall i = 1, \ldots, m.$$

□

Then we get a corollary from Theorem 3.8 and Lemma 3.19.

Corollary 3.20. *Deciding whether a completely positive unital map Φ is strictly positive is NP-hard.*

Remark 3.18. Corollary 3.10 shows that $\|\Phi\|_H = 1$ is equivalent to the existence of two vectors $u, v \in \mathbb{C}^n$ of norm 1 such that

$$\langle V_i u, V_j v \rangle = 0, \ \ \forall i, j \in \{1, \ldots, m\} \ .$$

This condition is known to be equivalent to the existence of two pure *distinguishable* pure states which the quantum channel Φ takes into two orthogonal subspaces, see [MA05]. This distinguishability of two quantum states is a fundamental property of quantum systems [NC00]. Beigi and Shor [BS08] defined the quantum analogue of the classical clique problem in a graph. Namely, they denote by $\alpha(\Phi)$ the maximum number of distinguishable pure states in the quantum channel Φ. It is immediate that $\|\Phi\|_H = 1$ if and only if $\alpha(\Phi) \geqslant 2$. We refer to [BS08] for more information on complexity issues concerning the quantum clique problem.

3.8.4 Complexity of determining the global convergence of a noncommutative consensus system: an open question

We showed in Corollary 3.15 that deciding whether a noncommutative consensus system is globally convergent can be done in polynomial time if the quantum map is irreducible. However the complexity of determining the global convergence of a noncommutative consensus system is left unknown. We recall that by Theorem 3.6, this is equivalent to the following decision problem:

Problem 3.2. *Input:* integers n, m, and matrices $V_1, \ldots, V_m \subset \mathbb{C}^{n \times n}$ with rational entries, satisfying

$$\sum_{i=1}^{m} V_i^* V_i = I_n \ .$$

Question: Let H be the smallest subspace containing the identity matrix and satisfying:

$$I_n \in H, \ \ V_i^* X V_j \in H \ \ \forall X \in H \ .$$

Is there a rank one matrix in the orthogonal complement of the subspace H?

Indeed, the subspace H considered in the latter question corresponds to the union of the subspaces $\cup_k H_k$ defined in (3.28).

CHAPTER 4

The contraction rate in Hilbert's projective metric of flows on cones

In this chapter, we apply the formula of the contraction ratio of linear maps in Hilbert's seminorm, obtained in the previous chapter, to finite dimensional nonlinear flows. In other words, we deal with the continuous time analogue of the results of the previous chapter. We first deduce a characterization formula for the contraction rate in Hilbert's seminorm of nonlinear flows. Our characterization leads to an explicitly calculable formula in $\mathbb{R}^n$ equipped with the standard partial order. In particular, we obtain an explicit contraction rate bound for a class of nonlinear consensus protocols. Using Nussbaum's Finsler approach, we also derive from the formula obtained in the previous chapter a characterization of the contraction rate of nonlinear maps in Hilbert's projective metric. We apply the general formula to a nonlinear matrix differential equation and obtain an explicit contraction rate bound in Hilbert's projective metric.

This chapter is part of the preprint [GQ12b].

4.1 Introduction

In this chapter, we consider a *finite dimensional* vector space $\mathscr{X}$, a closed pointed convex cone $\mathscr{C} \subset \mathscr{X}$ with interior $\mathscr{C}_0$ and a distinguished element $\mathbf{e} \in \mathscr{C}_0$. Let $\mathscr{D} \subset \mathscr{X}$ be an open set and $\phi : \mathscr{D} \to \mathscr{X}$ be a continuously differentiable application. Since ϕ is locally Lipschitz, we know that for

all $x_0 \in \mathscr{D}$, there is a maximal interval $J(x_0)$ such that a unique solution $x(\cdot;x_0) : J(x_0) \to \mathscr{D}$ of

$$\dot{x}(t) = \phi(x(t)), \quad x(0) = x_0 \tag{4.1}$$

exists. Recall that the flow associated to (4.1) is an application $M_\cdot(\cdot) : \mathbb{R} \times \mathscr{D} \to \mathscr{D}$ defined by:

$$M_t(x_0) = x(t;x_0), \quad t \in J(x_0).$$

The flow $M_\cdot(\cdot)$ may not be everywhere defined on $\mathbb{R} \times \mathscr{D}$. Since ϕ is continuously differentiable, the flow is differentiable with respect to the state variable. We denote by $DM_t(x)$ the derivative of the flow M with respect to the state variable at point (t,x). Recall that

$$\dot{DM_t}(x)z = D\phi(M_t(x))(DM_t(x)z), \;\; t \in J(x), z \in \mathscr{X}.$$

Let $U \subset \mathscr{D}$ be a convex open set. For $x_0 \in U$ define:

$$t_U(x_0) := \sup\{t_0 \leqslant J(x_0) : x(t;x_0) \in U, \;\; \forall t \in [0,t_0)\}$$

the time when the solution of (4.1) leaves U.

Consider a continuously differentiable map $\phi : \mathscr{X} \to \mathscr{X}$ such that

$$\phi(x + \lambda \mathbf{e}) = \phi(x), \;\; \forall \lambda \in \mathbb{R}, x \in \mathscr{X} \;\; .$$

Then the flow is additively homogeneous with respect to $\mathbf{e}$, i.e.,

$$M_t(x + \lambda \mathbf{e}) = M_t(x) + \lambda \mathbf{e}, \;\; \lambda \in \mathbb{R}, x \in \mathscr{X}, t \in J(x) \;\; .$$

Our first main result concerns the optimal contraction rate of additively homogeneous flows in Hilbert's seminorm. More precisely, the latter contraction rate can be formulated as:

$$\alpha(U) := \sup\{\beta \in \mathbb{R} : \|M_t(x) - M_t(y)\|_H \leqslant e^{-\beta t}\|x - y\|_H, \forall x, y \in U, t \leqslant t_U(x) \wedge t_U(y)\} \;\; .$$

We apply the characterization of the contraction ratio of linear maps in Hilbert's seminorm in Theorem 3.2 to obtain the following characterization (Theorem 4.1):

$$\alpha(U) = \inf_{x \in U} \inf_{\nu, \pi \in \operatorname{extr} \mathscr{P}(\mathbf{e})} \inf_{\substack{z \in \operatorname{extr}([0,\mathbf{e}]) \\ \langle \nu, z \rangle + \langle \pi, \mathbf{e} - z \rangle = 0}} \langle \nu, D\phi(x)z \rangle + \langle \pi, D\phi(x)(\mathbf{e} - z) \rangle.$$

The notations are already introduced in Chapter 3. We apply the latter formula to nonlinear differential consensus system and deduce explicit contraction rates for some consensus systems studied in [SM03, Mor05], see Section 4.5.1 and 4.5.2.

Next suppose that ϕ is defined on the interior of the cone $\mathscr{C}_0$ and positively homogeneous:

$$\phi(\lambda x) = \lambda \phi(x) \;\; , \forall \lambda > 0, x \in \mathscr{C}_0 \;\; .$$

Such condition implies that the flow is also positively homogeneous, i.e.,

$$M_t(\lambda x) = \lambda M_t(x), \;\; t \in J(x) \;\; .$$

Let $U \subset \mathscr{C}_0$ be a convex open set. The second object of this chapter is to characterize the the optimal contraction rate on U in Hilbert's projective metric, that is,

$$\kappa(U) := \sup\{\alpha \in \mathbb{R} : d_H(M_t(x_1), M_t(x_2)) \leqslant e^{-\alpha t} d_H(x_1, x_2), \forall x_1, x_2 \in U, t \leqslant t_U(x_1) \wedge t_U(x_2)\}.$$

Our second main result shows that (Theorem 4.3):

$$\kappa(U) = \inf_{x\in U} \inf_{z\in[0,x]} \inf_{\substack{\nu,\pi\in\mathscr{P}(x)\\ \langle\pi,z\rangle+\langle\nu,x-z\rangle=0}} \langle\pi, D\phi(x)z\rangle + \langle\nu, D\phi(x)(x-z)\rangle \ . \tag{4.2}$$

To obtain formula (4.2), we apply Nussbaum's characterization of the contraction ratio of nonlinear maps in Hilbert's projective metric [Nus94, Coro 2.1] and our formula of the contraction ratio of linear maps in Hilbert's seminorm (Theorem 3.2).

For applications, we specialize the general formulas respectively to $\mathbb{R}^n$ equipped with the standard partial order (Section 4.5) and to the space of Hermitian matrices equipped with the Loewner order (Section 4.6). In particular, we obtain an explicit contraction rate bound for a class of nonlinear consensus protocols (Corollary 4.5).

4.2 Contraction rate of linear flows in Hilbert's seminorm

In this section, we consider the case when $\phi : \mathscr{X} \to \mathscr{X}$ is a linear application. The set of linear transformations on $\mathscr{X}$ is denoted by $\mathrm{End}(\mathscr{X})$ and $I : \mathscr{X} \to \mathscr{X}$ denotes the identity transformation. Let $L \in \mathrm{End}(\mathscr{X})$ such that $L(\mathbf{e}) = 0$. The next proposition characterizes the contraction rate of the flow associated to the linear differential equation

$$\dot{x} = L(x),$$

with respect to Hilbert's seminorm.

Proposition 4.1. *Let $L \in \mathrm{End}(\mathscr{X})$ be a linear transformation from $\mathscr{X}$ to $\mathscr{X}$ such that $L(\mathbf{e}) = 0$. The optimal constant α such that*

$$\|\exp(tL)x\|_H \leqslant e^{-\alpha t}\|x\|_H, \ \ \forall t \geqslant 0, x \in \mathscr{X}$$

can be characterized by:

$$h(L) := \inf_{\nu,\pi\in\mathrm{extr}\,\mathscr{P}(\mathbf{e})} \inf_{\substack{x\in\mathrm{extr}([0,\mathbf{e}])\\ \langle\nu,x\rangle+\langle\pi,\mathbf{e}-x\rangle=0}} \langle\nu, L(x)\rangle + \langle\pi, L(\mathbf{e}-x)\rangle. \tag{4.3}$$

Proof. We define a functional on $\mathrm{End}(\mathscr{X})$ by:

$$F(W) = \sup_{\nu,\pi\in\mathscr{P}(\mathbf{e})} \sup_{x\in[0,\mathbf{e}]} \langle\pi-\nu, W(x)\rangle, \ \ \forall W \in \mathrm{End}(\mathscr{X}) \ .$$

By Theorem 3.2, the optimal constant α is:

$$\begin{aligned}\alpha &= -\lim_{\varepsilon\to 0^+} \varepsilon^{-1}(\|\exp(\varepsilon L)\|_H - 1)\\ &= -\lim_{\varepsilon\to 0^+} \varepsilon^{-1}(F(\exp(\varepsilon L)) - F(I)) \ .\end{aligned} \tag{4.4}$$

Note that

$$F = \sup_{\nu,\pi\in\mathscr{P}(\mathbf{e})} \sup_{x\in[0,\mathbf{e}]} F_{\nu,\pi,x}$$

where $F_{\nu,\pi,x} : \mathrm{End}(\mathscr{X}) \to \mathbb{R}$ is defined by:

$$F_{\nu,\pi,x}(W) = \langle\pi-\nu, W(x)\rangle, \ \ \forall W \in \mathrm{End}(\mathscr{X}) \ .$$

Since the function $F_{\nu,\pi,x}$ is continuously differentiable on $\mathrm{End}(\mathscr{X})$ and the functions $F_{\nu,\pi,x}(W)$ and $DF_{\nu,\pi,x}(W)$ are jointly continuous on (ν,π,x,W), we know that $F:\mathrm{End}(\mathscr{X})\to\mathbb{R}$ defines a subsmooth function (see Appendix A). The limit in (4.4) coincides with the one-side directional derivative of F at point I in the direction L. Hence, by the formula of the one-side directional derivative of a subsmooth function (A.3), we get:

$$F(I;L)=\sup_{\nu,\pi,x\in T(I)}\langle\pi-\nu,L(x)\rangle$$

where

$$T(I)=\arg\max\{F_{\nu,\pi,x}(I):x\in[0,\mathbf{e}],\nu,\pi\in\mathscr{P}(\mathbf{e})\}.$$

Hence,

$$\begin{aligned}\alpha &= -F(I;L)\\ &= -\sup_{\nu,\pi\in\mathscr{P}(\mathbf{e})}\ \sup_{\substack{x\in[0,\mathbf{e}]\\ \langle\pi-\nu,x\rangle=1}}\langle\pi-\nu,L(x)\rangle\\ &= \inf_{\nu,\pi\in\mathscr{P}(\mathbf{e})}\ \inf_{\substack{x\in[0,\mathbf{e}]\\ \langle\nu,x\rangle+\langle\pi,\mathbf{e}-x\rangle=0}}\langle\nu,L(x)\rangle+\langle\pi,L(\mathbf{e}-x)\rangle.\end{aligned}$$

Since $\mathscr{X}$ is finite dimensional, the sets $\mathscr{P}(\mathbf{e})$ and $[0,\mathbf{e}]$ are both compact, and they are the convex hull of their extreme points. Henceforth, arguing as in Remark 3.10, we can replace $\mathscr{P}(\mathbf{e})$ and $[0,\mathbf{e}]$ by $\mathrm{extr}\,\mathscr{P}(\mathbf{e})$ and $\mathrm{extr}([0,\mathbf{e}])$, respectively. □

We now state the analogous result to Proposition 4.1, which applies to *time dependent* linear flows. Let $t_0>0$ and $L_\cdot(\cdot):[0,t_0)\times\mathscr{X}\to\mathscr{X}$ be a continuous application linear in the second variable such that $L_t(\mathbf{e})=0$ for all $t\in[0,t_0)$. We denote by $U(s,t)$ the evolution operator of the following linear time-varying differential equation:

$$\dot{x}(t)=L_t(x),\ \ t\in[0,t_0).$$

Then a slight modification of the proof of Proposition 4.1 leads to the following result.

Proposition 4.2. *The optimal constant α such that*

$$\|U(s,t)x\|_H\leqslant e^{-\alpha(t-s)}\|x\|_H,\ \ \forall s,t\in[0,t_0),x\in\mathscr{X}.$$

can be characterized by:

$$\inf_{t\in[0,t_0)}h(L_t)=\inf_{t\in[0,t_0)}\ \inf_{\nu,\pi\in\mathrm{extr}\,\mathscr{P}(\mathbf{e})}\ \inf_{\substack{x\in\mathrm{extr}([0,\mathbf{e}])\\ \langle\nu,x\rangle+\langle\pi,\mathbf{e}-x\rangle=0}}\langle\nu,L_t(x)\rangle+\langle\pi,L_t(\mathbf{e}-x)\rangle. \tag{4.5}$$

4.3 Contraction rate of nonlinear flows in Hilbert's seminorm

Suppose that the dynamics $\phi:\mathscr{X}\to\mathscr{X}$ satisfy

$$\phi(x+\lambda\mathbf{e})=\phi(x),\ \ \forall\lambda\in\mathbb{R},x\in\mathscr{X}\ \ . \tag{4.6}$$

It follows that:

$$D\phi(x)\mathbf{e}=0,\ \ \forall x\in\mathscr{X}\ \ . \tag{4.7}$$

We denote by M the flow associated to the differential equation (see Section 4.1 for notations):

$$\dot{x} = \phi(x). \tag{4.8}$$

By uniqueness of the solution, it is clear that for all $x_0 \in \mathscr{X}$ and $\lambda \in \mathbb{R}$,

$$M_t(x_0 + \lambda \mathbf{e}) = M_t(x_0) + \lambda \mathbf{e}, \quad t \in J(x_0). \tag{4.9}$$

Let U be a convex open set. In this section we characterize the contraction rate of the flow associated to ϕ on U with respect to Hilbert's seminorm:

$$\alpha(U) := \sup\{\beta \in \mathbb{R} : \|M_t(x) - M_t(y)\|_H \leqslant e^{-\beta t}\|x-y\|_H, \forall x,y \in U, t \leqslant t_U(x) \wedge t_U(y)\}. \tag{4.10}$$

Theorem 4.1. *Let ϕ satisfy* (4.6) *and $U \subset \mathscr{X}$ be a convex open set. We have*

$$\alpha(U) = \inf_{x \in U} h(D\phi(x))$$

where h is defined in (4.3).

Proof. Denote

$$\beta = \inf_{x \in U} h(D\phi(x)).$$

For any $x \in U$, define:

$$L_t = D\phi(M_t(x)), \;\; t \in [0, t_U(x)).$$

Then by (4.7) we see that

$$L_t(\mathbf{e}) = 0, \;\; t \in [0, t_U(x)).$$

Let any $z \in \mathscr{X}$. Then $DM_t(x)z : t \in [0, t_U(x))$ is the solution of the following linear time-varying differential equation:

$$\begin{cases} \dot{x} = L_t(x), \;\; t \in [0, t_U(x)), \\ x(0) = z. \end{cases}$$

By Proposition 4.2, it is immediate that:

$$\omega(DM_t(x)z/\mathbf{e}) \leqslant e^{-\beta t}\omega(z/\mathbf{e}), \;\; t \in [0, t_U(x)), z \in \mathscr{X} \;\; . \tag{4.11}$$

Let $x, y \in U$. Denote $\gamma(s) = sx + (1-s)y : s \in [0,1]$ and let any

$$0 < h < \min\{t_U(\gamma(s)) : s \in [0,1]\} \;\; .$$

Then by applying (4.11) to every $x = \gamma(s)$ we get:

$$\omega(M_h(x) - M_h(y)/\mathbf{e}) \leqslant \int_0^1 \omega(DM_h(\gamma(s))(x-y)/\mathbf{e})ds \leqslant e^{-\beta h}\omega(x-y/\mathbf{e}).$$

Therefore, for all $x, y \in U$,

$$\limsup_{h \to 0^+} \frac{\|M_h(x) - M_h(y)\|_H}{h} \leqslant -\beta\|x-y\|_H.$$

We deduce that for all $x, y \in U$ and $t < t_U(x) \wedge t_U(y)$,

$$\limsup_{h \to 0^+} \frac{\|M_{t+h}(x) - M_{t+h}(y)\|_H}{h} \leqslant -\beta\|M_t(x) - M_t(y)\|_H.$$

Therefore,

$$\|M_t(x)-M_t(y)\|_H \leqslant e^{-\beta t}\|x-y\|_H,\ \ t < t_U(x)\wedge t_U(y).$$

This implies that

$$\alpha(U) \geqslant \beta.$$

Inversely, for all $x\in U$, there is $t_0>0$ and $\varepsilon_0>0$ such that for all $h\leqslant t_0$, $0<\varepsilon\leqslant\varepsilon_0$ and $z\in\mathscr{X}$

$$\|M_h(x+\varepsilon z)-M_h(x)\|_H \leqslant e^{-\alpha(U)h}\|\varepsilon z\|_H.$$

Therefore,

$$\begin{aligned}\|DM_h(x)(z)\|_H &= \|\lim_{\varepsilon\to 0^+}\frac{M_h(x+\varepsilon z)-M_h(x)}{\varepsilon}\|_H \\ &= \lim_{\varepsilon\to 0^+}\frac{\|M_h(x+\varepsilon z)-M_h(x)\|_H}{\varepsilon} \leqslant e^{-\alpha(U)h}\|z\|_H.\end{aligned}$$

Recall from (4.9) that:

$$DM_h(x)\mathbf{e}=\mathbf{e}\ \ .$$

Hence by Theorem 3.2,

$$\sup_z \|DM_h(x)(z)\|_H/\|z\|_H = \|DM_h(x)\|_H = \sup_{\nu,\pi\in\mathscr{P}(\mathbf{e})}\sup_{z\in[0,\mathbf{e}]}\langle \nu-\pi, DM_h(x)z\rangle\ \ .$$

Hence for all $h\leqslant t_0$,

$$\sup_{\nu,\pi\in\mathscr{P}(\mathbf{e})}\sup_{z\in[0,\mathbf{e}]}\langle \nu-\pi, DM_h(x)z\rangle \leqslant e^{-\alpha(U)h}.$$

It is then immediate that for $h\leqslant t_0$,

$$\sup_{\nu,\pi\in\operatorname{extr}\mathscr{P}(\mathbf{e})}\ \sup_{\substack{z\in\operatorname{extr}([0,\mathbf{e}])\\ \langle\nu,z\rangle+\langle\pi,\mathbf{e}-z\rangle=0}} -\langle\nu, DM_h(x)(\mathbf{e}-z)\rangle - \langle\pi, DM_h(x)z\rangle \leqslant e^{-\alpha(U)h}-1.$$

Dividing the two sides by h and passing to the limit as $h\to 0$ we get:

$$-h(D\phi(x)) \leqslant -\alpha(U).$$

Therefore $\beta\geqslant\alpha(U)$. □

4.4 Contraction rate of nonlinear flows in Hilbert's projective metric

In this section, we apply Theorem 3.2 to determine the contraction rate of nonlinear flows in Hilbert's projective metric. Let $\phi:\mathscr{C}^0\to\mathscr{X}$ be a continuously differentiable function defined on the interior of the cone such that

$$\phi(\lambda x)=\lambda\phi(x)\ \ ,\forall\lambda>0, x\in\mathscr{C}^0\ \ .$$

We denote by M the flow associated to the differential equation (see Section 4.1 for notations):

$$\dot{x}=\phi(x). \tag{4.12}$$

By uniqueness of the solution, it is clear that for all $x \in \mathscr{C}^0$,

$$M_t(\lambda x) = \lambda M_t(x), \quad t \in J(x). \tag{4.13}$$

Let $U \subset \mathscr{C}^0$ be a convex open set. Define the optimal contraction rate of the flow in Hilbert's projective metric on U by:

$$\kappa(U) := \sup\{\alpha \in \mathbb{R} : d_H(M_t(x_1), M_t(x_2)) \leqslant e^{-\alpha t} d_H(x_1, x_2), \forall x_1, x_2 \in U, t \leqslant t_U(x_1) \wedge t_U(x_2)\}. \tag{4.14}$$

For $x \in \mathscr{C}^0$, define:

$$c(x) := \inf_{z \in [0,x]} \inf_{\substack{\nu, \pi \in \mathscr{P}(x) \\ \langle \pi, z \rangle + \langle \nu, x - z \rangle = 0}} \langle \pi, D\phi(x) z \rangle + \langle \nu, D\phi(x)(x - z) \rangle \tag{4.15}$$

Arguing as in the proof of Proposition 4.1, we get

$$c(x) = \inf_{z \in \mathrm{extr}[0,x]} \inf_{\substack{\nu, \pi \in \mathrm{extr}\, \mathscr{P}(x) \\ \langle \pi, z \rangle + \langle \nu, x - z \rangle = 0}} \langle \pi, D\phi(x) z \rangle + \langle \nu, D\phi(x)(x - z) \rangle \tag{4.16}$$

Proposition 4.3. *For all $x \in \mathscr{C}_0$,*

$$c(x) = -\lim_{t \to 0^+} t^{-1} (\sup_z \frac{\omega(DM_t(x) z / M_t(x))}{\omega(z/x)} - 1).$$

Proof. Let $x \in \mathscr{C}_0$. Define a function $F : \mathrm{End}(\mathscr{X}) \to \mathbb{R}$ by:

$$F(W) = \sup_{z \in [0,x]} \sup_{\pi, \nu \in \mathscr{P}(x)} \langle \frac{\nu}{\langle \nu, W(x) \rangle} - \frac{\pi}{\langle \nu, W(x) \rangle}, W(z) \rangle, \ \ \forall W \in \mathrm{End}(\mathscr{X}) \ \ .$$

Recall from (4.13) that

$$DM_t(x) x = M_t(x) \ \ , \ \ t \in J(x) \ \ .$$

Hence for all $t \in J(x)$, the function $DM_t(x) : \mathscr{X}/\mathbb{R}x \to \mathscr{X}/\mathbb{R}M_t(x)$ defines a linear map and by applying Theorem 3.2, we get:

$$\begin{aligned} \|DM_t(x)\|_H &= \sup_{z \in [0,x]} \sup_{\nu, \pi \in \mathscr{P}(M_t(x))} \langle \nu - \pi, DM_t(x) z \rangle \\ &= \sup_{z \in [0,x]} \sup_{\nu, \pi \in \mathscr{P}(x)} \langle \frac{\nu}{\langle \nu, DM_t(x) x \rangle} - \frac{\pi}{\langle \pi, DM_t(x) x \rangle}, DM_t(x) z \rangle \\ &= F(DM_t(x)). \end{aligned}$$

Therefore,

$$\begin{aligned} &\lim_{t \to 0^+} t^{-1} (\sup \frac{\omega(DM_t(x) z / M_t(x))}{\omega(z/x)} - 1) \\ &= \lim_{t \to 0^+} t^{-1} (\|DM_t(x)\|_H - 1) \\ &= \lim_{t \to 0^+} t^{-1} (F(DM_t(x)) - F(I)) \end{aligned} \tag{4.17}$$

Recall that the function $DM_t(x) : [0, J(x)) \to \mathrm{End}(\mathscr{X})$ satisfies:

$$\lim_{t \to 0^+} t^{-1} (DM_t(x) - I) = D\phi(x).$$

The following reasoning is similar to that in the proof of Proposition 4.1. First for $\nu,\pi \in \mathscr{P}(x)$ and $z \in [0,x]$ define the function $F_{\nu,\pi,z} : \mathrm{End}(\mathscr{X}) \to \mathscr{X}$ by:

$$F_{\nu,\pi,z}(W) = \langle \frac{\nu}{\langle \nu, W(x)\rangle} - \frac{\pi}{\langle \pi, W(x)\rangle}, W(z)\rangle, \ \ \forall W \in \mathrm{End}(\mathscr{X}) \ \ .$$

It is clear that $F_{\nu,\pi,z}(W)$ and the derivative $DF_{\nu,\pi,z}(W)$ are jointly continuous on (ν,π,z,W). Therefore the function F, which can be written as

$$F = \sup_{z\in[0,x]} \sup_{\pi,\nu\in\mathscr{P}(x)} F_{\nu,\pi,z} \ \ ,$$

is a subsmooth function. The limit in (4.17) equals to the one-side directional derivative of F at I in the direction $D\phi(x)$. By applying the formula of the one-side directional derivative of a subsmooth function (A.3) we get:

$$\begin{aligned} &\lim_{t\to 0^+} t^{-1}(F(DM_t(x)) - F(I)) \\ &= \sup_{\nu,\pi,z\in T(I)} DF_{\nu,\pi,z}(I)(D\phi(x)) \end{aligned}$$

where

$$T(I) = \{\nu,\pi \in \mathscr{P}(x), z \in [0,x] : \langle \nu - \pi, z\rangle = 1\}.$$

The derivative of $F_{\nu,\pi,z}$ at point I in the direction $D\phi(x)$ is:

$$\begin{aligned} &DF_{\nu,\pi,z}(I)(D\phi(x)) \\ &= \frac{\langle \nu, D\phi(x)z\rangle\langle \nu, x\rangle - \langle \nu, z\rangle\langle \nu, D\phi(x)x\rangle}{\langle \nu, x\rangle^2} - \frac{\langle \pi, D\phi(x)z\rangle\langle \pi, x\rangle - \langle \pi, z\rangle\langle \pi, D\phi(x)x\rangle}{\langle \pi, x\rangle^2} \\ &= \langle \tfrac{\nu}{\langle \nu,x\rangle}, D\phi(x)z\rangle - \langle \tfrac{\nu}{\langle \nu,x\rangle}, z\rangle\langle \tfrac{\nu}{\langle \nu,x\rangle}, D\phi(x)x\rangle - \langle \tfrac{\pi}{\langle \pi,x\rangle}, D\phi(x)z\rangle + \langle \tfrac{\pi}{\langle \pi,x\rangle}, z\rangle\langle \tfrac{\pi}{\langle \pi,x\rangle}, D\phi(x)x\rangle. \end{aligned}$$

For $(\nu,\pi,z) \in T(I)$, it is easy to check that

$$\langle \nu, z\rangle = 1, \ \ \langle \pi, z\rangle = 0 \ \ .$$

Therefore, for all $(\nu,\pi,z) \in T(I)$, we have

$$\begin{aligned} &DF_{\nu,\pi,z}(I)(D\phi(x)) \\ &= \langle \nu, D\phi(x)z\rangle - \langle \nu, D\phi(x)x\rangle - \langle \pi, D\phi(x)z\rangle. \end{aligned}$$

Hence,

$$\begin{aligned} &\lim_{t\to 0^+} t^{-1}(\sup \frac{\omega(DM_t(x)z/M_t(x))}{\omega(z/x)} - 1) \\ &\lim_{t\to 0^+} t^{-1}(F(DM_t(x)) - F(I)) \\ &= \sup_{\nu,\pi,z\in T(I)} DF_{\nu,\pi,z}(W)(D\phi(x)) \\ &= \sup_{z\in[0,x]} \sup_{\substack{\nu,\pi\in\mathscr{P}(x) \\ \langle \nu-\pi,z\rangle = 1}} \langle \nu, D\phi(x)z\rangle - \langle \nu, D\phi(x)x\rangle - \langle \pi, D\phi(x)z\rangle \\ &= -c(x). \end{aligned}$$

□

We shall need to use Nussbaum's characterization of the contraction ratio of nonlinear maps in Hilbert's projective metric, in terms of the oscillation ratio of the derivative.

Theorem 4.2 (Coro 2.1, [Nus94]). *Let $U \subset \mathscr{C}^0$ be a convex open set such that $tU \subset U$ for all $t > 0$. Let $f : U \to \mathscr{C}^0$ be a continuously differentiable map such that $\omega(f(x)/f(y)) = 0$ whenever $x, y \in U$ and $\omega(x/y) = 0$. For each $x \in U$ define $\lambda(x)$, λ_0 and k_0 by:*

$$\lambda(x) := \inf\{c > 0 : \omega(Df(x)z/f(x)) \leqslant c\omega(z/x) \text{ for all } z \in \mathscr{X}\},$$

$$\lambda_0 := \sup\{\lambda(x) : x \in U\},$$

$$k_0 := \inf\{c > 0 : d_H(f(x), f(y)) \leqslant c d_H(x, y) \text{ for all } x, y \in U\}.$$

Then it follows that $\lambda_0 = k_0$.

Below is the main theorem of this section, which characterizes the contraction rate $\kappa(U)$.

Theorem 4.3. *Let $U \subset \mathscr{C}^0$ denote a convex open set such that $\lambda U = U$ for all $\lambda > 0$. Then*

$$\kappa(U) = \inf_{x \in U} c(x). \tag{4.18}$$

Proof. Fix $x_0 \in U$. By the Cauchy-Lipschitz theorem, there is $r > 0$ and $t_0 > 0$ such that the flow is well-defined on $[0, t_0] \times B(x_0; r)$ where $B(x_0; r)$ is the open ball of radius r centered at x_0. We assume that $B(x_0; r) \subset U$ and

$$t_U(x) \geqslant t_0, \ \ \forall x \in B(x_0; r) \ .$$

Denote the set

$$G := \underset{\lambda > 0}{\cup} \lambda B(x_0; r) \ .$$

By (4.14), for every $t \leqslant t_0$, the application M_t is well defined on G and

$$d_H(M_t(x), M_t(y)) \leqslant e^{-\kappa(U)t} d_H(x, y), \ \ \forall x, y \in G.$$

By Theorem 4.2, the latter formula implies that:

$$\omega(DM_t(x)z/M_t(x)) \leqslant e^{-\kappa(U)t}\omega(z/x) \ \ \forall x \in G, z \in \mathscr{X}.$$

Therefore by Proposition 4.3,

$$-c(x_0) = \limsup_{t \to 0^+} \frac{1}{t}(\sup_z \frac{\omega(DM_t(x_0)z/M_t(x_0))}{\omega(z/x_0)} - 1) \leqslant -\kappa(U).$$

It follows that

$$\kappa(U) \leqslant \inf_{x \in U} c(x).$$

Next we show the inverse inequality. Denote

$$c = \inf_{x \in U} c(x).$$

Then for all $x \in U$, $z \in \mathscr{X}$ and $t \in t_U(x)$,

$$\begin{aligned}
&\limsup_{h \to 0^+} \frac{\omega(DM_{t+h}(x)z/M_{t+h}(x)) - \omega(DM_t(x)z/M_t(x))}{h} \\
&= \limsup_{h \to 0^+} \frac{\omega(DM_h(M_t(x))(DM_t(x)z))/M_h(M_t(x))) - \omega(DM_t(x)z/M_t(x))}{h} \\
&= \limsup_{h \to 0^+} \frac{\omega(DM_t(x)z/M_t(x))}{h}(\frac{\omega(DM_h(M_t(x))(DM_t(x)z))/M_h(M_t(x))}{\omega(DM_t(x)z/M_t(x))} - 1) \\
&\leqslant -c(M_t(x))\omega(DM_t(x)z/M_t(x)) \\
&\leqslant -c\omega(DM_t(x)z/M_t(x)).
\end{aligned}$$

Note that the first inequality in the last formula is due to Proposition 4.3. Therefore, for all $x \in U$, $z \in \mathscr{X}$ and $t \in t_U(x)$ we have that,

$$\omega(DM_t(x)z/M_t(x)) \leqslant e^{-ct}\omega(z/x).$$

Let $x, y \in U$ and define $\gamma(s) = (1-s)x+sy$, $0 \leqslant s \leqslant 1$. By the compactness of the set $\{\gamma(s) : s \in [0,1]\}$, we know that

$$t_0 := \inf\{t_U(\gamma(s)) : s \in [0,1]\} > 0.$$

Therefore, using the Finsler structure of Hilbert's projective metric ([Nus94, Thm 2.1]), we get that for every $t \leqslant t_0$,

$$\begin{aligned} &d_H(M_t(x),M_t(y)) \leqslant \textstyle\int_0^1 \omega(DM_t(\gamma(s))(y-x)/M_t(\gamma(s)))ds \\ &\leqslant \textstyle\int_0^1 e^{-ct}\omega(y-x/\gamma(s))ds \\ &= e^{-ct}d_H(x,y). \end{aligned}$$

Consequently we proved that for all $x, y \in U$

$$\limsup_{h\to 0^+} \frac{d_H(M_h(x),M_h(y)) - d_H(x,y)}{h} \leqslant -cd_H(x,y).$$

This implies that for all $x, y \in U$ and $t < t_U(x) \wedge t_U(y)$:

$$\begin{aligned} &\limsup_{h\to 0^+} \frac{d_H(M_{t+h}(x),M_{t+h}(y)) - d_H(M_t(x),M_t(y))}{h} \\ &= \limsup_{h\to 0^+} \frac{d_H(M_h(M_t(x)),M_h(M_t(y))) - d_H(M_t(x),M_t(y))}{h} \\ &\leqslant -cd_H(M_t(x),M_t(y)). \end{aligned}$$

It follows that

$$d_H(M_t(x),M_t(y)) \leqslant e^{-ct}d_H(x,y), \quad \forall x,y \in U, t < t_U(x) \wedge t_U(y).$$

Therefore

$$\kappa(U) \geqslant c.$$

□

4.5 Applications to standard positive cone

In this section, we apply the previous results to the case $\mathscr{X} = \mathbb{R}^n$, $\mathscr{C} = \mathbb{R}^n_+$ and $\mathbf{e} = \mathbf{1}$. For $x \in \mathbb{R}^n$ we denote by $\delta(x)$ the diagonal matrix with entries x.

4.5.1 Contraction rate of linear flows in Hilbert's seminorm

In this subsection, we specialize Proposition 4.1 to $\mathbb{R}^n$ equipped with the standard partial order.

Corollary 4.4. *Let $A \in \mathbb{R}^{n\times n}$ be a square matrix such that $A\mathbf{1} = 0$. Then the best constant α such that*

$$\Delta(e^{At}x) \leqslant e^{-\alpha t}\Delta(x), \quad \forall x \in \mathbb{R}^n, t \geqslant 0 \ ,$$

can be characterized by:

$$h(A) = \min_{i\neq j} \big(A_{ji} + A_{ij} + \sum_{k\notin\{i,j\}} \min(A_{ik}, A_{jk})\big). \tag{4.19}$$

Proof. Recall that

$$\begin{aligned}\operatorname{extr}(\mathscr{P}(\mathbf{1})) &= \{e_1,\dots,e_n\},\\ \operatorname{extr}[0,\mathbf{1}] &= \{\sum_{i\in I} e_i : I\subset\{1,\dots,n\}\}.\end{aligned}$$

Therefore we have:

$$\begin{aligned}h(A) &= \min_{i\neq j}\min_{\substack{I\subset\{1,\dots,n\}\\ i\notin I, j\in I}} \sum_{k\in I} A_{ik} + \sum_{k\notin I} A_{jk}\\ &= \min_{i\neq j} A_{ij} + A_{ji} + \min_{\substack{I\subset\{1,\dots,n\}\\ i\notin I, j\in I}} \sum_{k\in I\setminus\{j\}} A_{ik} + \sum_{k\notin I\cup\{i\}} A_{jk}\\ &= \min_{i\neq j} A_{ij} + A_{ji} + \sum_{k\notin\{i,j\}} \min(A_{ik}, A_{jk}).\end{aligned}$$

□

Consider the order-preserving case, i.e. $A_{ij}\geqslant 0$ for $i\neq j$. Such situation was studied extensively in the context of consensus dynamics. In particular, let $G=(V,E)$ be a graph and equip each arc $(i,j)\in E$ a weight $C_{ij}>0$ (the node j is connected to i). One of the consensus systems that Moreau [Mor05] studied is:

$$\dot{x}_i = \sum_{(i,j)\in E} C_{ij}(x_j - x_i),\ \ i=1,\dots,n\ .$$

This can be written as $\dot{x}=Ax$, where $A_{ij}=C_{ij}$ for $i\neq j$ and $A_{ii}=\sum_j C_{ij}$ is a *discrete Laplacian*. A general result of Moreau implies that if there is a node connected by path to all other nodes in the graph G, then the system is globally convergent. Our results show that if $h(C)>0$ then the system converges exponentially to consensus with rate $h(C)$.

Remark 4.1. The condition $h(C)=0$ means that there are two nodes disconnected with each other ($C_{ij}+C_{ji}=0$) and all other nodes are connected by arc to at most one of them ($\sum_{k\notin\{i,j\}}\min(C_{ik},C_{jk})=0$). The condition $h(C)>0$, though more strict than Moreau's connectivity condition, gives an explicit contraction rate.

In addition, our result applies to not necessarily order-preserving flows. For example, consider the matrix

$$A=\begin{pmatrix} -3 & 1 & 2\\ 1 & 0 & -1\\ 1 & 1 & -2\end{pmatrix}.$$

A basic calculus shows that $h(A)=1$. Therefore, every orbit of the linear system $\dot{x}=Ax$ converges exponentially with rate 1 to a multiple of the unit vector.

Remark 4.2. We point out that as a contraction constant, $h(A)$ makes sense only when $A\mathbf{1}=0$. However, as a functional h is well defined on the space of square matrices. Moreover, since the diagonal elements do not account in the formula (4.19), it is clear that for any square matrix $B\in\mathbb{R}^{n\times n}$ and $x\in\mathbb{R}^n$

$$h(B)=h(B-\delta(x)).$$

4.5.2 Applications to nonlinear differential consensus systems

Let $G=(V,E)$ denote a directed graph. We equip every arc $(i,j)\in E$ with a weight $C_{ij}>0$. For $(i,j)\notin E$, we set $C_{i,j}=0$. Consider the following nonlinear consensus protocol [SM03]:

$$\dot{x}_k=\sum_{(i,k)\in E} C_{ik}f_{ik}(x_i-x_k),\ \ k=1,\dots,n, \tag{4.20}$$

where we suppose that every map $f_{ik}:\mathbb{R}\to\mathbb{R}$ is differentiable. When every f_{ik} is the identity map, the operator at the right hand-side of (4.20) is the discrete Laplacian of the digraph G, in which C_{ik} is the conductivity of arc (i,k).

Corollary 4.5. *Let $w>0$. Suppose that*

$$\alpha:=\inf\{f'_{ik}(t):t\in[-w,w],(i,k)\in E\}\geqslant 0. \tag{4.21}$$

Consider the convex open set

$$U(w)=\{x\in\mathbb{R}^n:\Delta(x)<w\}.$$

For all $x(\cdot):[0,T]\to\mathbb{R}^n$ satisfying $x(0)\in U(w)$ and (4.20)*, we have:*

$$\Delta(x(t))\leqslant e^{-h(C)\alpha t}\Delta(x(0)),\ \ \forall 0\leqslant t\leqslant T.$$

Proof. For all $x\in U(w)$,

$$h(D\phi(x))=\min_{i\neq j}\frac{\partial\phi_i(x)}{\partial x_j}+\frac{\partial\phi_j(x)}{\partial x_i}+\sum_{k\neq i,j}\min(\frac{\partial\phi_i(x)}{\partial x_k},\frac{\partial\phi_j(x)}{\partial x_k}),$$

where

$$\frac{\partial\phi_i(x)}{\partial x_j}=C_{ij}f'_{ij}(x_j-x_i),\ \ i\neq j.$$

Hence for all $x\in U(w)$,

$$h(D\phi(x))\geqslant\min_{i\neq j}C_{ij}\alpha+C_{ji}\alpha+\sum_{k\neq i,j}\min(C_{ik}\alpha,C_{jk}\alpha)=\alpha h(C).$$

We apply Theorem 4.1 and consider $y=\mathbf{1}$ in (4.10) to get:

$$\Delta(x(t))\leqslant e^{-\alpha h(C)t}\Delta(x(0)),\ \ t\leqslant t_{U(w)}(x(0))\ \ .$$

Since $\alpha\geqslant 0$ and $h(C)\leqslant 0$, we deduce that

$$\Delta(x(t))\leqslant\Delta(x(0))\leqslant w,\ \ t\leqslant t_{U(w)}(x(0))\ \ .$$

Hence the set $U(w)$ is invariant. In other words, for all $x(\cdot):[0,T]\to\mathbb{R}^n$ satisfying $x(0)\in U(w)$ and (4.20), we know that

$$\{x(t):t\in[0,T]\}\subset U(w)\ \ .$$

□

Example 4.3. The Kuramoto equation [Str00] is a special case of the protocol (4.20).

$$\dot{\theta}_i = \sum_{j:(i,j)\in E} C_{ij}\sin(\theta_j - \theta_i), i = 1,\dots,n. \tag{4.22}$$

Let $w < \pi/2$. Then

$$\inf\{\cos(t) : t \in [-w,w]\} \geqslant \cos w > 0.$$

We apply Corollary 4.5 and obtain that for all $\theta(\cdot) : [0,T] \to \mathbb{R}^n$ satisfying equation (4.22) and $\Delta(\theta(0)) < w$, we have:

$$\Delta(\theta(t)) \leqslant e^{-h(C)\cos(w)t}\Delta(\theta(0)), \ \ \forall t \geqslant 0.$$

In particular, for all $\theta(0) \in (-\pi/4, \pi/4)^n$, the solution of equation (4.22) satisfies:

$$\Delta(\theta(t)) \leqslant e^{-h(C)\cos(\Delta(\theta(0)))t}\Delta(\theta(0)), \ \ \forall t \geqslant 0.$$

Remark 4.4. Moreau [Mor05] showed that if there is a node connected by path to all other nodes in the graph (V,E), then the Kuramoto system (4.22) is globally convergent on the set $(-\pi/2, \pi/2)^n$. Compared to his results (see Remark 4.1), our condition for convergence is more strict but we obtain an explicit exponential contraction rate.

Example 4.5. Another class of maps satisfying (4.21) is when $f_{ik}(t) = \arctan(t)$ for all $i,k \in \{1,\dots,n\}$. Consider the following system

$$\dot{x}_i = \sum_{j:(i,j)\in E} C_{ij}\arctan(x_j - x_i), i = 1,\dots,n. \tag{4.23}$$

Then we obtain in the same way as in Example 4.3 that for all $x(0) \in \mathbb{R}^n$, the solution of (4.23) satisfies:

$$\Delta(x(t)) \leqslant e^{-\frac{h(C)t}{1+\Delta(x(0))^2}}\Delta(x(0)), \ \ \forall t \geqslant 0.$$

Example 4.6. (Discrete p-Laplacian) We now analyze the degenerate case of the p-Laplacian consensus dynamics for $p \in (1,2)\cup(2,+\infty)$. Then latter can be described by the following dynamical system in $\mathbb{R}^n$:

$$\dot{v}_i = \sum_{j:(i,j)\in E} C_{ij}(v_j - v_i)\left|C_{ij}(v_i - v_j)\right|^{p-2}, \ \ i = 1,\dots,n. \tag{4.24}$$

Let $\alpha > 0$ and consider the convex open set:

$$U(\alpha) := \{v : \max_{i\neq j}|v_i - v_j| < \alpha\}.$$

Let $0 < \beta < \alpha$ and consider a convex open set $V(\beta)$ contained in $\{v : \min_{i\neq j}|v_i - v_j| > \beta\}$ so that the vector field of (4.24) is C^1 in $V(\beta)$. A basic calculus shows that for $v \in V(\beta)$,

$$\frac{\partial\phi_i(v)}{\partial v_j} = \begin{cases} 0, & (i,j) \notin E \\ (p-1)|v_i - v_j|^{p-2}C_{ij}^{p-1}, & (i,j) \in E \end{cases}$$

Let C^{p-1} denote the matrix with entries C_{ij}^{p-1}. Recall that $h(C^{p-1}) \geqslant 0$. Then we have:

$$h(D\phi(x)) \geqslant \begin{cases} (p-1)h(C^{p-1})\beta^{p-2}, & p > 2, x \in V(\beta) \ , \\ (p-1)h(C^{p-1})\alpha^{p-2}, & 1 < p < 2 \ , x \in V(\beta)\cap U(\alpha) \ . \end{cases}$$

We remark that the contraction rate on $V(\beta)$ tends to $+\infty$ when p tends to $+\infty$.

4.5.3 Contraction rate of nonlinear flows in Hilbert's projective metric

Suppose that $\phi : \operatorname{int}\mathbb{R}^n_+ \to \mathbb{R}^n$ is a continuously differentiable function such that

$$\phi(\lambda x) = \lambda\phi(x), \ \ \forall\lambda > 0, x \in \operatorname{int}\mathbb{R}^n_+ \ .$$

We specialize Theorem 4.3 to obtain a contraction rate characterization in Hilbert's projective metric of the flow associated to the equation:

$$\dot{x} = \phi(x) \ . \tag{4.25}$$

Corollary 4.6. *Let $U \subset \operatorname{int}\mathbb{R}^n_+$ be a convex open set. When $\mathscr{X} = \mathbb{R}^n$ and $\mathscr{C} = \mathbb{R}^+_n$, the contraction rate on U in Hilbert's projective metric of the flow associated to* (4.25) *can be characterized as below:*

$$\kappa(U) = \inf_{x\in U} c(x) = \inf_{x\in U} h(A(x)),$$

where

$$A(x) = \delta(x)^{-1} D\phi(x)\delta(x)$$

and h is defined in (4.19)*.*

Proof. It is sufficient to remark that in this special case:

$$\operatorname{extr}\mathscr{P}(x) = \delta(x)^{-1}\operatorname{extr}\mathscr{P}(\mathbf{1}) \ ,$$

and

$$\operatorname{extr}[0,x] = \delta(x)\operatorname{extr}([0,\mathbf{1}]) \ .$$

Therefore,

$$\begin{aligned} c(x) &= \inf_{z\in\operatorname{extr}[0,x]} \ \inf_{\substack{\pi,\nu\in\operatorname{extr}\mathscr{P}(x)\\ \langle\nu,z\rangle+\langle\pi,x-z\rangle=0}} \langle\nu, D\phi(x)z\rangle + \langle\pi, D\phi(x)(x-z)\rangle \\ &= \inf_{z\in\operatorname{extr}[0,\mathbf{1}]} \ \inf_{\substack{\pi,\nu\in\operatorname{extr}\mathscr{P}(\mathbf{1})\\ \langle\nu,z\rangle+\langle\pi,x-z\rangle=0}} \langle\delta(x)^{-1}\nu, D\phi(x)\delta(x)z\rangle + \langle\delta(x)^{-1}\pi, D\phi(x)\delta(x)(\mathbf{1}-z)\rangle \\ &= h(A(x)). \end{aligned}$$

□

Remark 4.7. Consider the linear flow in $\mathbb{R}^n$ of the following equation:

$$\dot{x} = Ax,$$

where $A_{ij} \geqslant 0$, for all $i \neq j$, so that the flow is order-preserving. Let x be in the interior of $\mathbb{R}^n_+$. Then we have

$$\delta(x)^{-1}A\delta(x)_{ij} = A_{ij}\frac{x_j}{x_i}, \ \ i,j = 1,\ldots,n.$$

Therefore,

$$h(\delta(x)^{-1}A\delta(x)) = \min_{i\neq j} A_{ji}\frac{x_i}{x_j} + A_{ij}\frac{x_j}{x_i} + \sum_{k\notin\{i,j\}} \min(A_{ik}\frac{x_k}{x_i}, A_{jk}\frac{x_k}{x_j}).$$

The global contraction rate (restricted to $\mathscr{C}^0$) is then

$$\inf_{x\in\mathscr{C}^0} h(\delta(x)^{-1}A\delta(x)) = \min_{i\neq j} 2\sqrt{A_{ij}A_{ji}}.$$

(This formula may be alternatively obtained by differentiating with respect to t, at point 0, the contraction ratio of $I+tA$, using Birkhoff's theorem.)

It follows that a positive global contraction rate exists if and only if $A_{ij}>0$ for all $i\neq j$. However, a strict *local* contraction may occur even if there is $A_{ij}=0$ for some $i\neq j$. Let $K>1$ and consider the convex open set

$$U(K)=\{x\in\mathbb{R}^n : \frac{1}{K}\leqslant\frac{x_i}{x_j}\leqslant K\}.$$

Then the local contraction rate with respect to $U(K)$ is

$$\inf_{x\in U(K)} h(\delta(x)^{-1}A\delta(x))\geqslant\frac{h(A)}{K}.$$

Therefore, $h(A)>0$ is sufficient to have a strict local contraction. Moreover, the above bound on the contraction rate increases (faster convergence) as the orbit approaches to consensus, i.e., a multiple of $\mathbf{1}$.

4.6 Applications to the space of Hermitian matrices

We now specialize our general results to the case $\mathscr{X}=\mathrm{S}_n$, $\mathscr{C}=\mathrm{S}_n^+$ and $\mathbf{e}=I_n$.

4.6.1 Contraction rate of a linear flow in Hilbert's seminorm

In this subsection, we specialize Formula (4.3) to a linear flow in S_n associated to

$$\dot{X}=\phi(X) \tag{4.26}$$

where $\phi:\mathrm{S}_n\to\mathrm{S}_n$ is linear and $\phi(I_n)=0$.

Corollary 4.7. *The contraction rate of the linear flow associated to* (4.26) *can be characterized by:*

$$h(\phi)=\inf_{\substack{X=(x_1,\dots,x_n)\\ XX^*=I_n}}\big(x_1^*\phi(x_2x_2^*)x_1+x_2^*\phi(x_1x_1^*)x_2+\sum_{k=3}^{n}\min(x_1^*\phi(x_kx_k^*)x_1,x_2^*\phi(x_kx_k^*)x_2)\big). \tag{4.27}$$

where x_i is the i-th column vector of each unitary matrix X. □

Proof. Recall that

$$\mathrm{extr}(\mathscr{P}(I_n))=\{xx^* : x\in\mathbb{C}^n, x^*x=1\},\ \ \mathrm{extr}[0,I_n]=\{P\in\mathrm{S}_n : P^2=P\}.$$

Then,

$$\begin{aligned}
h(\phi)&=\inf_{x_1^*x_1=x_2^*x_2=1}\ \inf_{\substack{P^2=P\\ Px_1=0,Px_2=x_2}} x_1^*\phi(P)x_1+x_2^*\phi(I_n-P)x_2\\
&=\inf_{x_1^*x_1=x_2^*x_2=1}\ \inf_{\substack{P=x_2x_2^*+\dots x_kx_k^*\\ P^2=P,Px_1=0}}\sum_{i=2}^{k}x_1^*\phi(x_ix_i^*)x_1+x_2^*\phi(I_n-P)x_2\\
&=\big(\inf_{x_1^*x_1=x_2^*x_2=1} x_1^*\phi(x_2x_2^*)x_1+x_2^*\phi(x_1x_1^*)x_2\\
&\qquad+\inf_{\substack{X=(x_1,x_2,\dots,x_n)\\ XX^*=I_n}}\sum_{i=3}^{k}x_1^*\phi(x_ix_i^*)x_1+\sum_{i=k+1}^{n}x_2^*\phi(x_ix_i^*)x_2\big).
\end{aligned}$$

□

As pointed out in Remark 4.2, h is a functional well defined for all linear applications from S_n to S_n. It is interesting to remark that for any linear application $\Psi : \mathrm{S}_n \to \mathrm{S}_n$ and any square matrix $Z \in \mathbb{C}^{n\times n}$,

$$h(\Psi) = h(\Phi)$$

where Φ is defined by

$$\Phi(X) = \Psi(X) - ZX - XZ^*, \ \ \forall X \in \mathrm{S}_n \ \ .$$

4.6.2 Contraction rate of nonlinear flows in Hilbert's projective metric

Suppose that $\phi : \hat{\mathrm{S}}_n^+ \to \mathrm{S}_n$ is a continuously differentiable function such that

$$\phi(\lambda X) = \lambda \phi(X), \ \ \forall \lambda > 0, X \in \hat{\mathrm{S}}_n^+ \ \ .$$

We specialize the contraction formula (4.18) to obtain the contraction rate in Hilbert's projective metric of the flow associated to the following equation on $\hat{\mathrm{S}}_n^+$:

$$\dot{X} = \phi(X) \ \ . \tag{4.28}$$

Corollary 4.8. *Let $U \subset \hat{\mathrm{S}}_n^+$ be a convex open set, the contraction rate on U of the flow associated to* (4.28) *in Hilbert's projective metric can be characterized by:*

$$\kappa(U) = \inf_{P\in U} c(P) = \inf_{P\in U} h(\Phi(P))$$

where $\Phi(P) : \mathrm{S}_n^+ \to \mathrm{S}_n^+$ is a linear application given by:

$$\Phi(P)(Z) = P^{-\frac{1}{2}} D\phi(P)(P^{\frac{1}{2}} Z P^{\frac{1}{2}}) P^{-\frac{1}{2}} \tag{4.29}$$

and $h : \mathrm{S}_n \to \mathbb{R}$ is defined in (4.27)*.*

Proof. Remark that in this special case,

$$\mathrm{extr}[0, P] = P^{\frac{1}{2}}(\mathrm{extr}[0, I_n])P^{\frac{1}{2}},$$

and

$$\mathrm{extr}(\mathscr{P}(P)) = P^{-\frac{1}{2}}(\mathrm{extr}\,\mathscr{P}(I_n))P^{-\frac{1}{2}}.$$

The desired formula is obtained the same way as in the proof of Corollary 4.6. □

Example 4.8. As an example, let us show a calculus of contraction rate using Corollary 4.8 for the following differential equation in S_n:

$$\dot{P} = \phi(P) := \frac{-PBP}{\mathrm{trace}(CP)} + AP + PA^* \tag{4.30}$$

where $B, C \in \hat{\mathrm{S}}_n^+$. Let $\hat{P} \in \mathrm{S}_n^+$. Then the linear application $\Phi(P) : \mathrm{S}_n \to \mathrm{S}_n$ defined in (4.29) is given by:

$$\begin{aligned}\Phi(P)(Z) &= P^{-\frac{1}{2}} D\phi(P)(P^{\frac{1}{2}} Z P^{\frac{1}{2}}) P^{-\frac{1}{2}} \\ &= (-ZP^{\frac{1}{2}}BP^{\frac{1}{2}} - P^{\frac{1}{2}}BP^{\frac{1}{2}}Z)\,\mathrm{trace}(CP)^{-1} \\ &\quad + P^{\frac{1}{2}}BP^{\frac{1}{2}}\,\mathrm{trace}(CP)^{-2}\,\mathrm{trace}(CP^{\frac{1}{2}}ZP^{\frac{1}{2}}) + P^{-\frac{1}{2}}AP^{\frac{1}{2}}Z + ZP^{\frac{1}{2}}A^*P^{-\frac{1}{2}}, \ \ \forall Z \in \mathrm{S}_n \ \ .\end{aligned}$$

Therefore let $x, y \in \mathbb{C}^n$ such that $x^*y = 0$ then

$$y^*\Phi(P)(xx^*)y = (y^*P^{\frac{1}{2}}BP^{\frac{1}{2}}y)(x^*P^{\frac{1}{2}}CP^{\frac{1}{2}}x)\operatorname{trace}(CP)^{-2}.$$

Let $\{x_1, \dots, x_n\}$ be an orthonormal basis. Denote

$$\begin{aligned}
\alpha_1 &= x_1^*P^{\frac{1}{2}}BP^{\frac{1}{2}}x_1 \ , \\
\alpha_2 &= x_2^*P^{\frac{1}{2}}BP^{\frac{1}{2}}x_2 \ , \\
\beta_1 &= x_1^*P^{\frac{1}{2}}CP^{\frac{1}{2}}x_1 \ , \\
\beta_2 &= x_2^*P^{\frac{1}{2}}CP^{\frac{1}{2}}x_2 \ .
\end{aligned}$$

Without loss of generality, we assume that $\alpha_1 \leqslant \alpha_2$. Then

$$\begin{aligned}
&x_1^*\Phi(P)(x_2x_2^*)x_1 + x_2^*\Phi(P)(x_1x_1^*)x_2 + \sum_{k=3}^{n} \min(x_1^*\Phi(P)(x_kx_k^*)x_1, x_2^*\Phi(P)(x_kx_k^*)x_2) \\
&= \big(\alpha_1\beta_2 + \alpha_2\beta_1 + \alpha_1 \sum_{k=3}^{n} x_k^*P^{\frac{1}{2}}CP^{\frac{1}{2}}x_k\big)\operatorname{trace}(CP)^{-2} \\
&= \big(\alpha_1\beta_2 + \alpha_2\beta_1 + \alpha_1(\operatorname{trace}(CP) - \beta_1 - \beta_2)\big)\operatorname{trace}(CP)^{-2} \\
&= \big(\alpha_1 \operatorname{trace}(CP) + \beta_1(\alpha_2 - \alpha_1)\big)\operatorname{trace}(CP)^{-2} \\
&\geqslant \lambda_{\min}(BP)\operatorname{trace}(CP)^{-1}.
\end{aligned}$$

Therefore by the definition in (4.27),

$$h(\Phi(P)) \geqslant \lambda_{\min}(BP)\operatorname{trace}(CP)^{-1}.$$

Let us consider the convex open set

$$U = \{P \in \hat{\mathrm{S}}_n^+ : d_H(P, I_n) < K\}.$$

Then,

$$\begin{aligned}
\inf_{P\in U} h(\Phi(P)) &\geqslant \inf_{P\in U} \lambda_{\min}(BP)\operatorname{trace}(CP)^{-1} \\
&\geqslant \frac{\lambda_{\min}(BP)}{n\lambda_{\max}(CP)} \geqslant \frac{\lambda_{\min}(B)\lambda_{\min}(P)}{n\lambda_{\max}(C)\lambda_{\max}(P)} \\
&\geqslant \frac{\lambda_{\min}(B)}{n\lambda_{\max}(C)e^K}
\end{aligned}$$

Let $\alpha = \dfrac{\lambda_{\min}(B)}{n\lambda_{\max}(C)e^K}$. Then by Corollary 4.8, for all $P_1, P_2 \in U$ we have:

$$d_H(M_t(P_1), M_t(P_2)) \leqslant e^{-\alpha t} d_H(P_1, P_2), \ \ 0 \leqslant t < t_U(P_1) \wedge t_U(P_2).$$

If A, B, C are matrices such that for some $\lambda_0 \in \mathbb{R}$,

$$\phi(I_n) = -B\operatorname{trace}(C)^{-1} + A + A' = \lambda_0 I_n,$$

then we know that

$$M_t(I_n) = e^{\lambda_0 t} I_n.$$

In that case, for $P \in U$ we have:

$$d_H(M_t(P), e^{\lambda_0 t} I_n) \leqslant e^{-\alpha t} d_H(P, I_n), \ \ 0 \leqslant t < t_U(P).$$

It follows that $t_U(P) = +\infty$ and therefore every solution of equation (4.30) converges exponentially to a scalar multiplication of I_n.

Part II

Max-plus based numerical methods for optimal control problems

CHAPTER 5

Max-plus basis methods: general principle and asymptotic approximation error estimates

In this chapter, we first review the general principle of max-plus basis methods. Then, we establish a negative result, showing that some form of curse dimensionality is unavoidable for these methods, but also for more classical approximate dynamic programming methods like stochastic dual dynamic programming, in which a convex value function is approximated by a supremum of affine functions. Indeed, we show that asymptotically, the minimal approximation error in the L_1 or L_∞ norm, for a smooth convex function, using at most n affine minorants, is equivalent to $1/n^{2/d}$, as the number of basis functions n goes to infinity. We derive the latter result as an analogue of Gruber's best asymptotic error estimates of approximating a convex body using circumscribed polytopes. We also give explicit asymptotic constants, respectively for the L_1 or L_∞ norm. Both constants rely on the determinant of the Hessian matrix of the convex function to approximate. We deduce that an attenuation of the curse of dimensionality occurs (fewer basis functions are needed) when the convex function to be approximated is "flat" in some direction, i.e., when its Hessian matrix has some eigenvalues close to zero.

This chapter extends the theoretical part of the conference article [GMQ11].

5.1 Introduction

In this chapter, we consider the following finite horizon optimal control problem

Problem 5.1.

$$v(x,T) := \sup_{\mathbf{u}\in\mathscr{U}_T} \int_0^T \ell(\mathbf{x}(s),\mathbf{u}(s))ds + \phi(\mathbf{x}(T)) \ ;$$

$$\dot{\mathbf{x}}(s) = f(\mathbf{x}(s),\mathbf{u}(s)), \quad \mathbf{x}(0) = x, \quad \mathbf{x}(s)\in X, \mathbf{u}(s)\in U \ . \tag{5.1}$$

Here, $X \subset \mathbb{R}^d$ is the set of states, $U \subset \mathbb{R}^m$ is the set of actions, T denotes the horizon and $\mathscr{U}_T$ denotes the locally integrable control functions with values in U:

$$\mathscr{U}_T := L^1([0,T];U) \ .$$

The Lagrangian $\ell : X \times U \to \mathbb{R}$, the terminal reward $\phi : X \to \mathbb{R}$, and the dynamics $f : X \times U \to \mathbb{R}^d$ are given functions. The supremum is taken over all the control functions $\mathbf{u}$ and system trajectories $\mathbf{x}$ satisfying (5.1), and v is the *value function*, depending on the initial condition $x \in X$ and the final horizon $T > 0$. We will assume here for simplicity that the set X is invariant by the dynamics (5.1) for all choices of the control function $\mathbf{u} \in \mathscr{U}_T$. Under certain regularity assumptions, it is known that $v(x,t)$ is a viscosity solution of the Hamilton-Jacobi partial differential equation (HJ PDE)[CL83, LS85]:

$$\begin{cases} \frac{\partial v}{\partial t} - H(x, \frac{\partial v}{\partial x}) = 0, & \forall (x,t) \in X \times (0,T] \ , \\ v(x,0) = \phi(x), & \forall x \in X \ . \end{cases} \tag{5.2}$$

where

$$H(x,p) = \sup_{u\in U} p' f(x,u) + \ell(x,u), \ \ x \in X, p \in \mathbb{R}^d$$

denotes the Hamiltonian of the optimal control problem. Several techniques have been proposed in the literature to solve the latter HJ PDE. We mention, for example, the finite difference schemes [CL84], the discrete dynamic programming method by Capuzzo Dolcetta [CD83] or the semi-Lagrangian method developed by Falcone, Ferretti and Carlini [Fal87, FF94, CFF04], the high order ENO schemes introduced by Osher, Sethian and Shu [OS88, OS91], the discontinuous Galerkin method by Hu and Shu [HS99], the ordered upwind methods for convex static Hamilton-Jacobi equations by Sethian and Vladimirsky [SV03] which is an extension of the fast marching method for the Eikonal equations [Set99], and the antidiffusive schemes for advection of Bokanowski and Zidani [BZ07]. However, these methods generally require the generation of a grid on the state space. Thus they suffer from the so-called curse of dimensionality, meaning that the execution time grows exponentially with the dimension of the state space. The question of the attenuation of the curse of dimensionality has received much attention by the numerical optimal control community. We mention the domain decomposition algorithm [CFLS94, FLS94] and the patchy domain decomposition technique [NK07, CCFP12]. In the discrete dynamic programming community, specially in the study of Markov decision processes, various techniques have also been proposed to reduce the curse of dimensionality, including the approximate policy iteration [Ber11], the classification-based policy iteration [LGM10] and the point based value iteration [CLZ97].

Recently a new class of methods has been developed after the work of Fleming and McEneaney [FM00], see also the works of McEneaney [McE07], of Akian, Gaubert and Lakhoua [AGL08], of McEneaney, Deshpande and Gaubert [MDG08], of Sridharan *et al.* [SGJM10], and of Dower and McEneaney [DM11].. These methods are referred to as *max-plus basis methods* since they all rely on

max-plus algebra. Their common idea is to approximate the value function by a supremum of finitely many "basis functions" and to propagate the supremum forward in time by exploiting the max-plus linearity of the Lax-Oleinik semigroup.

To compare this new class of methods with the classical ones, the first question to understand is why, and to what extent, max-plus techniques can attenuate the curse of dimensionality. To this end, based on the Gruber's best error estimates of convex body covering by circumscribed polytopes [Gru93a, Gru93b], we establish asymptotic minimal error estimates of semiconvex function approximation by finitely quadratic basis functions (Theorem 5.3 and 5.4). Our results imply that the curse of dimensionality is unavoidable for all class of numerical methods which approximate a smooth convex function by a finite number of affine functions, including the stochastic dual dynamic programming method [Sha11]. However, the asymptotic constants show that an attenuation of the curse of dimensionality is possible for value functions with negligible determinant of Hessian matrix.

The main object of this chapter is to recall the general principle of max-plus basis methods and to show the inherent curse of dimensionality to the family of max-plus basis methods based on c-semiconvex transforms. In Section 5.2, we review the general principle of the methods. In Section 5.3, we state the asymptotic best error estimates of semiconvex based approximation. The proof is given is Section 5.4.

5.2 Max-plus numerical methods to solve optimal control problems

5.2.1 The Lax-Oleinik semigroup

Let $(S_T)_{T\geqslant 0}$ be the *Lax-Oleinik* semigroup, i.e., the *evolution semigroup* of the Hamilton-Jacobi equation (5.2). Then for every horizon $T>0$, S_T is a map which associates to the terminal reward ϕ the value function $v(x,T)$ on horizon T:

$$S_T[\phi](x) = v(x,T) = \sup \int_0^T \ell(\mathbf{x}(s),\mathbf{u}(s))ds + \phi(\mathbf{x}(T)) \ ; \tag{5.3}$$

$$\dot{\mathbf{x}}(s) = f(\mathbf{x}(s),\mathbf{u}(s)), \quad \mathbf{x}(0)=x, \quad \mathbf{x}(s)\in X, \mathbf{u}(s)\in U \ . \tag{5.4}$$

By *semigroup*, we mean that

$$S_{t+s} = S_t \circ S_s, \ \ \forall t,s \geqslant 0 \ .$$

Recall that the *max-plus semiring*, $\mathbb{R}_{max}$, is the set $\mathbb{R}\cup\{-\infty\}$, equipped with the addition $(a,b)\mapsto \max(a,b)$ and the multiplication $(a,b)\mapsto a+b$. For all functions f,g from X to $\mathbb{R}_{max}$ and $\lambda\in\mathbb{R}_{max}$, we denote by $f\vee g$ the pointwise maximum of f and g, namely,

$$(f\vee g)(x) = \max(f(x),g(x)), \ \ x\in X \ ,$$

and by $\lambda + f$ the function f modified by the constant λ:

$$(\lambda + f)(x) = \lambda + f(x), \ \ x\in X \ .$$

It is known that the semigroup S_t is *max-plus linear* [Mas87, KM97, AQV98], i.e.,

$$S_t[f\vee g] = S_t[f]\vee S_t[g], \qquad S_t[\lambda + g] = \lambda + S_t[g] \ . \tag{5.5}$$

We shall see that the max-plus basis methods exploit these properties to solve the optimal control problem (5.3).

5.2.2 Max-plus linear spaces

A set $\mathscr{W}$ of functions from $\mathbb{R}^d$ to $\mathbb{R}_{\max}$ is a max-plus linear space if for all $\phi_1, \phi_2 \in \mathscr{W}$ and $\lambda \in \mathbb{R}$, the functions $\phi_1 \vee \phi_2$ and $\lambda + \phi_1$ belong to $\mathscr{W}$. A max-plus linear space $\mathscr{W}$ is (conditionally) *complete* if the pointwise supremum of any family of functions of $\mathscr{W}$ that is bounded from above by an element of $\mathscr{W}$ is finite.

Let $\mathscr{B}$ be a set of functions from $\mathbb{R}^d$ to $\mathbb{R}$ (*max-plus basis functions*). The complete max-plus (linear) space $\overline{\operatorname{span}}\mathscr{B}$ of functions generated by $\mathscr{B}$ is defined to be the set of arbitrary linear combinations of elements of $\mathscr{B}$, in the max-plus sense, so that an element ϕ of $\overline{\operatorname{span}}\mathscr{B}$ reads

$$\sup_{\omega \in \mathscr{B}} (a(\omega) + \omega)$$

for some family $(a(\omega))_{w \in \mathscr{B}}$ of elements in $\mathbb{R}_{\max}$. The (non complete) space $\operatorname{span}\mathscr{B}$ is defined in a similar way, but the linear combination must now involve a finite family, meaning that $a(\omega)$ should equal to $-\infty$ for all but finitely many values of $\omega \in \mathscr{B}$. We refer the reader to [LMS01, CGQ04, McE06] for more background on max-plus linear spaces.

If $\mathscr{W}$ is a complete max-plus linear space of functions $\mathbb{R}^d \to \mathbb{R}_{\max}$, and if ψ is any function $\mathbb{R}^d \to \mathbb{R}_{\max}$, the *max-plus projection* of ψ onto $\mathscr{W}$ is defined to be

$$P_{\mathscr{W}}(\psi) := \max\{\phi \in \mathscr{W} \mid \phi \leqslant \psi\} \tag{5.6}$$

(by writing max, we mean that the supremum element of the set under consideration belongs to this set, which follows from the completeness of $\mathscr{W}$).

All the previous definitions can be dualized, replacing max by min, and $-\infty$ by $+\infty$. In particular, a complete min-plus linear space is a set $\mathscr{Z}$ of functions $\mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that $-\mathscr{Z} := \{-w \mid w \in \mathscr{Z}\}$ is a complete max-plus linear space. Then, we define the dual projector $P^{\mathscr{Z}}$ by

$$P^{\mathscr{Z}}(\psi) := \min\{\phi \in \mathscr{Z} \mid \phi \geqslant \psi\} \ ,$$

for all functions $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$.

We list below some typical choices of basis functions, appearing in the literature (see [FM00, AGL08, McE07]).

Example 5.1 (linear basis functions [FM00]). Consider linear basis functions:

$$\mathscr{B}_S := \{p'x : p \in S\} \ ,$$

where S is a possibly infinite subset of $\mathbb{R}^d$. Then the complete max-plus linear space $\mathscr{W}_S$ spanned by $\mathscr{B}_S$ is:

$$\mathscr{W}_S := \{\sup_{p \in S} p'x + a(p) : a \in \mathbb{R}_{\max}^S\}.$$

If ψ is a convex function, then

$$P_{\mathscr{W}_S}(\psi) = \sup_{p \in S} p'x - \psi^*(p) \ ,$$

where ψ^* denotes the convex conjugate of function ψ. Figure 5.1 is an illustration of max-plus projection of one dimensional convex function where the convex function is

$$\psi(x) = \frac{x^2}{2} \ ,$$

and the set S is

$$S = \{-1, 1\}.$$

If $S = \mathbb{R}^d$, then it follows from Fenchel-Moreau theorem that the space $\overline{\text{span}}\mathscr{B}_S$ coincides with the space of convex (lower semicontinuous) functions. We then say that linear basis functions are *adapted* for approximating convex functions.

Example 5.2 (quadratic basis functions [FM00, AGL08]). Quadratic basis functions, with the same Hessian $c > 0$, refer to the basis set:

$$\mathscr{B}_{c,S} := \{-\frac{c}{2}|x|^2 + p'x : p \in S\} \ , \tag{5.7}$$

where S is a possibly infinite subset of $\mathbb{R}^d$. The complete max-plus linear space $\mathscr{W}_{c,S}$ spanned by $\mathscr{B}_{c,S}$ is then:

$$\mathscr{W}_{c,S} := \{\sup_{p\in S} -\frac{c}{2}|x|^2 + p'x + a(p) : a \in \mathbb{R}^S_{\max}\}.$$

Denote by I_d the identity matrix of dimension d. We abuse the notation to denote the quadratic function

$$I_d(x) := x'x, \ \ x \in \mathbb{R}^d.$$

Recall that a function ψ is c-semiconvex if the function $\psi + \frac{c}{2}I_d$ is convex. It is direct that

$$P_{\mathscr{W}_{c,S}}(\psi) = P_{\mathscr{W}_S}(\psi + \frac{c}{2}I_d) - \frac{c}{2}I_d \ ,$$

and

$$P_{\mathscr{W}_{c,S}}(\psi) = \sup_{p\in S} -\frac{c}{2}|x|^2 + p'x - (\psi + \frac{c}{2}I_d)^*(p).$$

Figure 5.2 is an illustration of max-plus projection of one dimensional semiconvex function where $c = 1$, the c-semiconvex function is

$$\psi(x) = \sin(x), \ \ x \in \mathbb{R} \ ,$$

and the set S is

$$S = \{-4, 0, 1.8, 3.5\}.$$

The same, if $S = \mathbb{R}^d$, then it follows from Fenchel-Moreau theorem that the space $\mathscr{W}_{c,S}$ coincides with the space of *c-semiconvex* (lower semicontinuous) functions. In [AGL08], such basis functions are called P_2 finite elements.

Example 5.3 (P_1 finite element [AGL08]). The P_1 finite elements or Lipschitz finite elements, with constant $b > 0$, refer to basis functions of the following form:

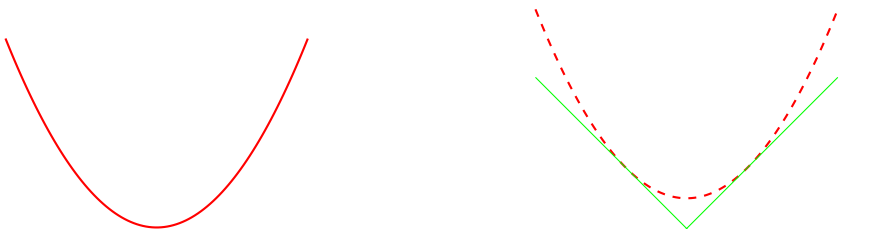
$$\mathscr{T}_{b,S} := \{-b|x - x_0|_1 : x_0 \in S\}$$

where S is a possibly infinite subset of $\mathbb{R}^d$. The min-plus linear space $\mathscr{Z}_{b,S}$ spanned by $-\mathscr{T}_{b,S}$ is then:

$$\mathscr{Z}_{b,S} := \{\inf_{x_0\in S} b|x - x_0|_1 + a(x_0) : a \in \mathbb{R}^S_{\min}\}.$$

If ψ is a Lipschitz function with constant b, then

$$P^{\mathscr{Z}_{b,S}}(\psi) = \inf_{x_0\in S} b|x - x_0|_1 + \psi(x_0).$$



(a) convex function $\psi(x) = x^2/4$

(b) projection $P_{\mathscr{W}}(\psi) = x-1 \vee -x-1$

Figure 5.1: An example of max-plus projection of convex function

Figure 5.3 is an illustration of min-plus projection of Lipschitz function using P_1 finite elements where $b = 1$, the b-Lipschitz function is

$$\psi(x) = \cos(x), \quad \forall x \in \mathbb{R} \ ,$$

and the set S is

$$S = \{1, \pm\pi, \pm\frac{\pi}{2}\} \ .$$

5.2.3 Max-plus basis methods-general principle

The general approach of max-plus basis methods for solving the optimal control problem (Problem 5.1) works as follows. First we discretize the time interval $[0,T]$ by small step $\tau > 0$ such that $T = N\tau$ for some integer $N > 0$. Choose a set of basis functions $\mathscr{B}$. For $k \in \{0,\ldots,N\}$, the value function $v(\cdot, k\tau)$ at time $k\tau$ will be approximated by a finite max-plus linear combination v_h^k of basis functions, i.e.,

$$v(x,k\tau) \simeq v_h^k(x) = \sup_{i\in I_k}(\lambda_i^k + w_i^k(x)), \quad k = 0,1,\ldots,N-1 \ ,$$

where $\{w_i^k : i \in I_k\} \subseteq \mathscr{B}$ for all k. Then, the coefficients $\{\lambda_i^k\}_{i\in I_k}$ and the functions $\{w_i^k\}_{i\in I_{k+1}}$ need to be inductively determined. Using the semigroup property, we know that

$$v(\cdot,(k+1)\tau) = S_\tau[v(\cdot,k\tau)], \qquad k = 0,1,\ldots,N-1 \ . \tag{5.8}$$

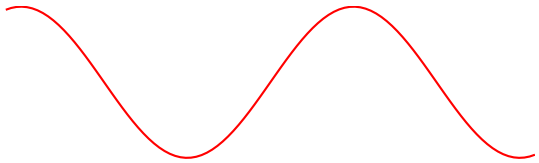
We require the max-plus basis approximation v_h^k of $v(\cdot,k\tau)$ to satisfy the analogous relation, at least approximately:

$$v_h^{k+1} \simeq S_\tau[v_h^k] = \sup_{i\in I_k}(\lambda_i^k + S_\tau[w_i^k]), \quad k = 0,\ldots,N-1 \ . \tag{5.9}$$

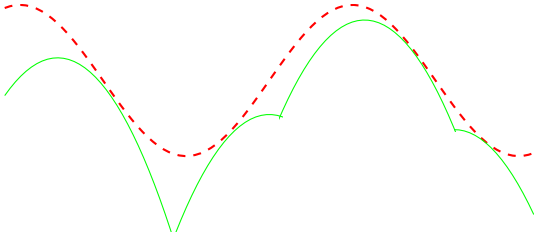
Next we decompose the problem into two subproblems:

1. Semigroup approximation:
$$S_\tau[w_i^k] \simeq \tilde{S}_\tau(w_i^k), \quad i \in I_k \ ;$$
2. Max-plus projection:
$$\sup_{i\in I_k}(\lambda_i^k + \tilde{S}_\tau[w_i^k]) \simeq \sup_{i\in I_{k+1}} \lambda_i^{k+1} + w_i^{k+1} \ .$$

Depending on the problem structure, the way that we address the two subproblems can be different. In the next section, we present some examples of max-plus basis methods.

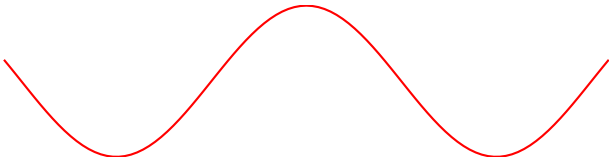


(a) semiconvex function $\psi(x) = \sin x$

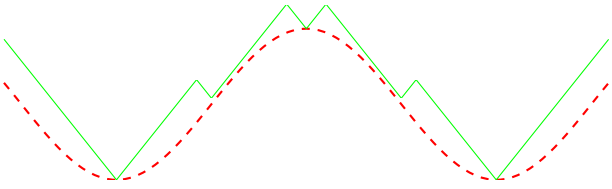


(b) projection $P_{\mathscr{W}}(\psi) = -\frac{x^2}{2} + 1.55 \vee -\frac{(x-1.8)^2}{2} + 2.8 \vee -\frac{(x+4)^2}{2} + 2.3 \vee -\frac{(x-3.5)^2}{2} + 1.35$

Figure 5.2: An example of max-plus projection of semi-convex function



(a) semiconvex function $\psi(x) = \cos x$



(b) projection $P^{\mathscr{Z}}(\psi) = |x| + 1 \wedge |x - \frac{\pi}{2}| \wedge |x + \frac{\pi}{2}| \wedge |x + \pi| + 1 \wedge |x - \pi| - 1$

Figure 5.3: An example of min-plus projection of Lipschitz function

5.2.4 Max-plus basis methods-examples

The first subproblem is the simplest one: computing $S_\tau[w_i^k]$ is equivalent to solving an optimal control problem, but the horizon τ is small, and the terminal reward w_i^k (typically a quadratic function) has a regularizing and a "concavifying" effect, which implies that the global optimum can be accurately approached (by reduction to a convex programming problem), leading to various approximations with a consistency error of $O(\tau^r)$, with $r = 3/2, 2$, or sometimes better, depending on the scheme, see [McE06, AGL08, LAK07].

We next present two different approaches to solve the second subproblem.

Example 5.4 (Method of Flemming and McEneaney). In the original paper [FM00], the authors choose a finite set of basis functions

$$\mathscr{B} = \{\omega_i : i = 1, \dots, n\}.$$

Denote by $\mathscr{W}$ the max-plus linear space spanned by $\mathscr{B}$. The approximation functions $\{v_h^k\}_k$ are required to satisfy:

$$v_h^{k+1} = \sup_{i=1}^{n} \lambda_i^k + P_{\mathscr{W}}(\tilde{S}[\omega_i]), \;\; k = 0, \dots, N-1 \;\; .$$

Let B be a n by n matrix such that:

$$P_{\mathscr{W}}(\tilde{S}_\tau[\omega_i]) = \sup_{j=1}^{n} B_{ji} + \omega_j, \;\; \forall i = 1, \dots, n.$$

Then the recursive equations of the coefficients $\{\lambda^k\}_k$ are:

$$\lambda_i^{k+1} = \sup_{j=1}^{n} B_{ij} + \lambda_j^k, \quad i = 1, \dots, n, \;\; k = 0, \dots, N-1 \;\; .$$

Example 5.5 (Max-plus finite element). In the max-plus finite element method of Akian, Gaubert and Lakhoua [AGL08], the authors choose a finite set of basis functions:

$$\mathscr{B} = \{\omega_i : i = 1, \dots, n\},$$

and a finite set of test functions:

$$\mathscr{T} = \{z_i : i = 1, \dots, m\}.$$

Denote by $\mathscr{W}$ the max-plus linear space spanned by $\mathscr{B}$ and $\mathscr{Z}$ the min-plus linear space spanned by $\mathscr{T}$. The approximation functions $\{v_h^k\}_k$ are required to satisfy:

$$v_h^{k+1} = P^{\mathscr{Z}} P_{\mathscr{W}}((\sup_{i=1}^{n} \lambda_i^k + \tilde{S}_\tau[\omega_i])), \;\; k = 0, \dots, N-1 \;\; .$$

Let K and M be two n by m matrices such that:

$$P^{\mathscr{Z}}(\omega_i) = \inf_{j=1}^{m} M_{ji} + z_j, \quad P^{\mathscr{Z}}(\tilde{S}_\tau[\omega_i]) = \inf_{j=1}^{m} K_{ji} + z_j, \;\; i = 1, \dots, n.$$

Then the recursive equations of the coefficients $\{\lambda^k\}_k$ are

$$\lambda_i^{k+1} = \min_{j=1,\dots,n} (-M_{ji} + \max_{l=1,\dots,m} (K_{jl} + \lambda_l^k)) \quad i = 1, \dots, n, \;\; k = 0, \dots, N-1 \;\; .$$

5.2.5 Max-plus basis methods-complexity and error bound

According to the general principle of max-plus basis methods, there is no direct discretization of the state space (only a discretization of the time interval). That is the special feature of max-plus basis methods, compared to the classical numerical methods. The accuracy of the method is limited by the semigroup approximation and the max-plus projection (see Section 5.2.3). We measure the approximation error in L_p norm:

$$\|v(\cdot,T)-v_h^N\|_p := \begin{cases} (\int_X |v(x,T)-v_h^N(x)|^p dx)^{1/p}, & p>0 \\ \sup_{x\in X} |v(x,T)-v_h^N(x)|, & p=+\infty \end{cases}$$

Most often we consider the L_∞ norm because the semigroup S_t is nonexpansive with respect to the L_∞ norm.

5.2.5.a Method of Flemming and McEneaney The arithmetic complexity of the method of Flemming and McEneaney, presented in Example 5.4, is polynomial with respect to the number of basis functions n and to the number of steps N.

The following lemma regarding the approximation error is immediate:

Lemma 5.1. *If the semigroup approximation is a subapproximation, namely,*

$$\tilde{S}_\tau[\omega_i] \leqslant S_\tau[\omega_i], \ \ \forall i=1,\dots,n \ \ ,$$

then for all p,

$$\|v(\cdot,T)-v_h^N\|_p \geqslant \|v(\cdot,T)-P_{\mathscr{W}}(v(\cdot,T))\|_p \ \ .$$

5.2.5.b Max-plus finite element method The arithmetic complexity of the max-plus finite element methods, presented in Example 5.5, is polynomial with respect to the number of basis functions n, to the number of test functions m and to the number of steps N.

In [AGL08], the following error bound is proved by using the nonexpansive property of the semigroup S_t in the sup norm.

Lemma 5.2 ([AGL08])**.** *The approximation error of max-plus finite element method, presented in Example 5.5, is bounded by the sum of the semigroup approximation error and the projection error:*

$$\begin{aligned}\|v(\cdot,T)-v_h^N\|_\infty \leqslant (1+N)\big(&\max_{i=1,\dots,n}\|S_\tau[\omega_i]-\tilde{S}_\tau[\omega_i]\|_\infty \\ &+\max_{k=0,\dots,N}\|v(\cdot,k\tau)-P_{\mathscr{W}}(v(\cdot,k\tau))\|_\infty \\ &+\max_{k=0,\dots,N}\|v(\cdot,k\tau)-P^{\mathscr{Z}}(v(\cdot,k\tau))\|_\infty\big)\end{aligned}$$

Actually the above error bound also holds for the method of Flemming and McEneaney:

Lemma 5.3. *The approximation error of the method of Flemming and McEneaney, presented in Example 5.4, is bounded by the sum of the semigroup approximation error and the projection error:*

$$\begin{aligned}\|v(\cdot,T)-v_h^N\|_\infty \leqslant (1+N)\big(&\max_{i=1,\dots,n}\|S_\tau[\omega_i]-\tilde{S}^\tau[\omega_i]\|_\infty \\ &+\max_{i=1,\dots,n}\|P_{\mathscr{W}}(\tilde{S}_\tau[\omega_i])-\tilde{S}_\tau[\omega_i]\|_\infty\big)\end{aligned}$$

The proof follows the same idea as that of Lemma 5.2.

As we discussed in Section 5.2.4, the semigroup approximation error is often of order $O(\tau^2)$ or $O(\tau)$, depending on the approximation scheme. If we want to have a better idea about the complexity of the method, we need to estimate the error order with respect to the number of basis functions n. For this, let us focus at the max-plus projection error estimates (the min plus projection error estimates can be done in the same way). The following projection error bound is given in [LAK07], Proposition 64.

Proposition 5.4 ([LAK07])**.** *Let X be a bounded convex subset of $\mathbb{R}^d$ and $S \subset \mathbb{R}^d$ be a finite set. Let $0 < a < c$ and let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a $(c-a)$-semiconvex function. Assume that for all $x \in \mathrm{ri}X$, $\partial(\psi + \frac{c}{2} I_d) \cap \mathrm{conv}\, S \neq \emptyset$.Let $\hat{S} = (\cup_{x \in \mathrm{ri}X} \partial(\psi + \frac{c}{2} I_d)(x)) \cap \mathrm{conv}\, S$, then*

$$\|\psi - P_{\mathscr{W}_{c,S}}(\psi)\|_\infty \leqslant \frac{1}{2a} \big(\sup_{x \in \hat{S}} \inf_{y \in S} |x-y|_2 \big)^2.$$

The term $\sup_{x \in \hat{S}} \inf_{y \in S} |x-y|_2$ corresponds to the maximal diameter of the Voronoi tessellation of the domain $\hat{S}$ by the discrete set S (see [OBSC00]). This proposition reveals the hidden space discretization nature in max-plus basis methods. It is known [Hla49, Rog64] that the minimal number $n(\varepsilon)$ of discrete points to get a Voronoi tessellation of diameter ε of a compact in $\mathbb{R}^d$ is equivalent to ε^{-d} as ε goes to infinity:

$$n(\varepsilon) \sim \frac{1}{\varepsilon^d}, \text{ as } \varepsilon \to 0.$$

Thus we know that the minimal projection error is bounded by $O(\frac{1}{n^{\frac{2}{d}}})$.

$$\min_{|S|=n} \|\psi - P_{\mathscr{W}_{c,S}}(\psi)\|_\infty = O(\frac{1}{n^{\frac{2}{d}}}), \text{ as } n \to +\infty.$$

However, an upper bound on the minimal projection error is not enough to understand to what extent the max-plus basis methods can attenuate the curse of dimensionality. The object of the next section is to give an asymptotic estimation of the minimal max-plus projection error of semiconvex functions, as the number of basis functions n tends to infinity.

5.3 Curse of dimensionality for semiconvex based approximations

In this section, we give an asymptotic estimate of the minimal max-plus projection error as the number of basis functions tends to infinity, in the special case in which the basis functions take the form:

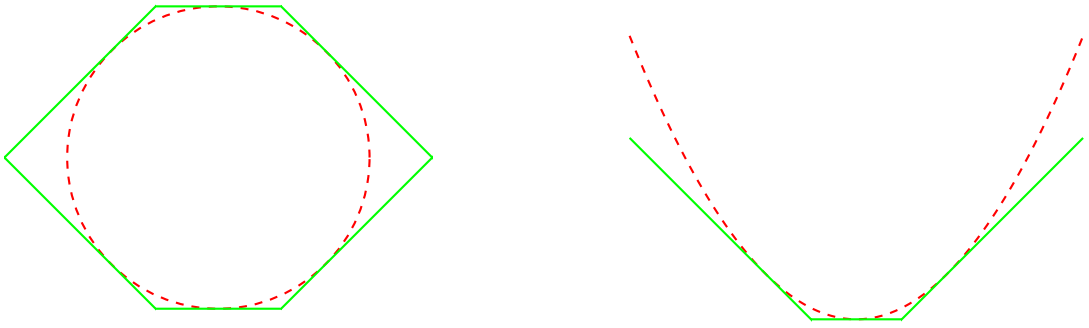
$$\mathscr{B}_{c,S} = \{-\frac{c}{2}|x|^2 + p'x : p \in S\}.$$

as in the max-plus basis method [FM00], or in the P_2 type finite element method of [AGL08, LAK07].

Let $X \subseteq \mathbb{R}^d$ be a full dimensional compact convex subset. For two functions f and g from $\mathbb{R}^d$ to $\mathbb{R}$, denote by

$$\delta_X^1(f,g) := \int_X |f(x) - f(x)| dx, \qquad \delta_X^\infty(f,g) := \sup_{x \in X} |f(x) - g(x)|$$

respectively the L_1 and the L_∞ metric between two functions f and g, measured on the compact X. Let $c \in \mathbb{R}$, $\varepsilon > 0$ and $\psi : \mathbb{R}^d \to \mathbb{R}$ be a $(c-\varepsilon)$-semiconvex function. The minimal L_1 and L_∞ max-plus



(a) approximate a convex body by circumscribed polytope

(b) approximate a convex function by supremum of its finitely affine minorants

Figure 5.4: Similarity between the two approximation problems

approximation error on X of ψ using at most n basis functions in $\mathscr{B}_{c,\mathbb{R}^d}$ is defined by:

$$\delta^1_{X,n}(\psi,c) = \inf\{\delta^1_X\big(\psi, P_{\mathscr{W}_{c,S}}(\psi)\big) : S \subseteq \mathbb{R}^d, |S| = n\} \tag{5.10}$$

$$\delta^\infty_{X,n}(\psi,c) = \inf\{\delta^\infty_X\big(\psi, P_{\mathscr{W}_{c,S}}(\psi)\big) : S \subseteq \mathbb{R}^d, |S| = n\} \tag{5.11}$$

When $c = 0$, we denote simply $\delta^1_{X,n}(\psi)$ and $\delta^\infty_{X,n}(\psi)$.

Remark 5.6. As noted in Example 5.2, if ψ is c-semiconvex, then

$$P_{\mathscr{W}_{c,S}}(\psi) = P_{\mathscr{W}_S}(\psi + \frac{c}{2}I_d) - \frac{c}{2}I_d \ .$$

Hence, for a $(c-\varepsilon)$-semiconvex function ψ, we have

$$\delta^\infty_{X,n}(\psi,c) = \delta^\infty_{X,n}(\psi + \frac{c}{2}I_d), \quad \delta^1_{X,n}(\psi,c) = \delta^1_{X,n}(\psi + \frac{c}{2}I_d).$$

In other words, the error estimation of max-plus projection of semiconvex function using quadratic basis functions is equivalent to that of max-plus projection of convex function using linear basis functions. By considering the epigraph, the latter approximation is closely related to the approximation of a convex body by a circumscribed polytope, see an illustration in Figure 5.4.

The error estimates of approximating a convex body using polytopes have been studied by many authors, see [Hla49, Sch87, Lud99, Bör00, Gru93a, Gru93b]. Following [Gru93a, Gru93b], let $C \subset \mathbb{R}^d$ be a convex body of non empty interior and $\mathscr{P}^C_{(n)}$ be the set of polytopes having at most n facets and circumscribed to C. For $P \in \mathscr{P}^C_{(n)}$, the approximation error of C by P related to the Hausdorff metric δ^H and to the volume difference δ^V are defined respectively by:

$$\delta^H(C,P) := \max_{x\in C}\min_{y\in P}\|x-y\|, \quad \delta^V(C,P) := \mathrm{vol}(C\backslash P).$$

Gruber [Gru93a, Gru07] considered the minimal approximation errors with respect to the Hausdorff metric and the volume difference:

$$\delta^H_n(C) := \inf\{\delta^H(C,P) : P \in \mathscr{P}^C_{(n)}\}$$

$$\delta^V_n(C) := \inf\{\delta^V(C,P) : P \in \mathscr{P}^C_{(n)}\}$$

He established the following asymptotic formulas:

Theorem 5.1. *[Gru93a] Let $C \subset \mathbb{R}^d$ be a convex body with non empty interior. Suppose that the boundary of C is of class $\mathscr{C}^2$ and the Gaussian curvature κ_C is strictly positive. Then,*

$$\delta_n^H(C) \sim \frac{1}{2}\Big(\frac{\vartheta_{d-1}}{\mathscr{K}_{d-1}}\int_{\partial C}(\kappa_C(x))^{\frac{1}{2}}d\sigma(x)\Big)^{\frac{2}{d-1}} n^{\frac{2}{1-d}}, \quad \text{as } n\to\infty$$

where $\mathscr{K}_d$ and ϑ_d denote respectively the volume of the unit ball in $\mathbb{R}^d$ and the minimum density of covering of $\mathbb{R}^d$ with Euclidean balls of unit radius (see Theorem 5.6 below).

Theorem 5.2. *[Gru93b] Let $C \subset \mathbb{R}^d$ be a convex body as in Theorem 5.1. Then, there is a constant δ_d depending only on d, such that:*

$$\delta_n^V(C) \sim \frac{\delta_d}{2}\Big(\int_{\partial C}\kappa_C(x)^{\frac{1}{d+1}}d\sigma(x)\Big)^{\frac{d+1}{d-1}} n^{\frac{2}{1-d}}, \quad \text{as } n\to\infty.$$

Here, σ is the ordinary surface area measure on the boundary of C.

In both of the proofs of the last two theorems, the boundary ∂C is partitioned into finitely many pieces, each of which is associated to a supporting hyperplane of C. For each supporting hyperplane H, a Cartesian coordinate system in H with origin at the intersection point p and the interior unit normal vector of ∂C at p form a Cartesian coordinate system in $\mathbb{R}^d$. Let $P \in \mathscr{P}^C_{(n)}$. The lower part of ∂C and ∂P with respect to the last coordinate can be represented by the graph of a strongly convex function f and the graph of $\sup_{i\in I} g_i$ where $\{g_i : i \in I\}$ is a finite set of affine minorants of f. Then, $\delta^H(C,P)$ and $\delta^V(C,P)$ are estimated through the L_1 or the L_∞ metric between the strongly convex function f and the supremum of its affine minorants $\sup_{i\in I} g_i$, on the Cartesian coordinate system associated to each supporting hyperplane.

It is not difficult to see that the technique used by Gruber can be adapted directly to obtain similar asymptotic formulas for the max-plus approximation error of a strongly convex function by the supremum of its affine minorants, as the number of minorants goes to infinity. In fact our problem is simpler since now we have a universal hyperplane ($H = \mathbb{R}^{d-1}$) thus a universal Cartesian coordinate system. Below are the analogous results for the asymptotic semiconvex based max-plus approximation error estimates:

Theorem 5.3 (L_∞ approximation error)**.** *Let $c \in \mathbb{R}$, $\varepsilon > 0$ and let $X \subset \mathbb{R}^d$ denote any full dimensional compact convex subset. If $\psi(x) : \mathbb{R}^d \to \mathbb{R}$ is $(c-\varepsilon)$-semiconvex of class $\mathscr{C}^2$, then we have:*

$$\delta^\infty_{X,n}(\psi,c) \sim \frac{1}{2}\Big(\frac{\vartheta_d}{\mathscr{K}_d}\int_X(\det(\psi''_x + cI_d))^{\frac{1}{2}}dx\Big)^{\frac{2}{d}} n^{-\frac{2}{d}} \ , \text{ as } n\to\infty.$$

Theorem 5.4 (L_1 approximation error)**.** *Let c, ε, X and $\psi(x)$ be as in Theorem 5.3. Then we have:*

$$\delta^1_{X,n}(\psi,c) \sim \frac{\delta_d}{2}\Big(\int_X(\det(\psi''_x + cI_d))^{\frac{1}{d+2}}dx\Big)^{\frac{d+2}{d}} n^{-\frac{2}{d}} \ , \text{ as } n\to\infty.$$

Here det means determinant.

The proof of these theorems is reported to Section 5.4, it builds on analogous methods and results of Theorem 5.1 and 5.2 in [Gru93a, Gru93b].

The following theorem is a direct corollary:

Theorem 5.5. *Assume that the value function of Problem 5.1 is $\mathscr{C}^2$ and c-semiconvex. Then, for any max-plus basis method providing an approximation from below of of the value function by a supremum of n quadratic functions, the L^1 and L^∞ approximation error are both $\Omega\big(\frac{1}{n^{2/d}}\big)$ as $n\to\infty$.*

Remark 5.7. Theorem 5.5 states that the curse of dimensionality is unavoidable in max-plus basis methods for certain class of optimal control problems. However, Theorems 5.3 and 5.4 also show that if the determinant of the Hessian matrix of $\psi + \frac{c}{2} I_d$ is close to zero, then the asymptotic constants will be close to zero. In that case, an attenuation of the curse of dimensionality should be observed.

5.4 Proof of asymptotic estimates

In this section, we present the proof of Theorems 5.3 and 5.4. The proof follows in essence the same lines as that of Theorem 5.1 and 5.2 by Gruber [Gru93a, Gru93b]. However, we choose to include a full proof in order to make the thesis self-contained. Throughout the section, X is a convex compact in $\mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}$ a strictly convex function of class $\mathscr{C}^2$. We prove the following asymptotic L_∞ and L_1 error estimates:

$$\delta^{\infty}_{X,n}(f) \sim \frac{1}{2}\Big(\frac{\vartheta_d}{\mathscr{H}_d}\int_X (\det(f''_x))^{\frac{1}{2}} dx\Big)^{\frac{2}{d}} n^{-\frac{2}{d}} \ , \text{ as } n \to \infty. \tag{5.12}$$

$$\delta^{1}_{X,n}(f) \sim \frac{\delta_d}{2}\Big(\int_X (\det(f''_x))^{\frac{1}{d+2}}\, dx\Big)^{\frac{d+2}{d}} n^{-\frac{2}{d}} \ , \text{ as } n \to \infty. \tag{5.13}$$

As noted in Remark 5.6, once the last two formulas are proved, Theorems 5.3 and 5.4 can be deduced immediately.

5.4.1 Preparations of the proof

We gather in this subsection the notions and theorems needed for the proof. We shall need the following notion of Bregman distance:

Definition 5.1 ([Brè67])**.** For any two points x and y of X, the *Bregman distance* from x to y, associated to the function f, is defined by

$$\mathrm{dist}_B(x;y) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle. \tag{5.14}$$

One interpretation of the Bregman distance $\mathrm{dist}_B(x;y)$ is the error at point x by approximating the convex function f by its affine minorant which is exact at point y. We call y the *contact point* of the affine minorant $f(y) + \langle \nabla f(y), x - y\rangle$. Note that the Bregman distance is positive definite ($\mathrm{dist}_B(x,y) \geqslant 0$ and the equality holds if and only if $x = y$), but it may not be symmetric. Besides, in general the Bregman distance does not satisfy the triangular inequality.

The hessian matrix of f at point x defines a positive definite quadratic form, denoted by f''_x. Then the family of positive definite forms $\{f''_x : x \in \mathbb{R}^d\}$ defines a Riemannian metric and a Riemannian measure on $\mathbb{R}^d$.

Definition 5.2 (Riemannian metric)**.** Let $x, y \in \mathbb{R}^d$. The Riemannian metric between x and y with respect to the family $\{f''_x : x \in \mathbb{R}^d\}$, denoted by $\gamma(x,y)$, is equal to:

$$\gamma(x,y) = \inf\{\int_0^1 (\dot{u}'_t f''_{u_t} \dot{u}_t)^{\frac{1}{2}}\, dt \,|\, u_t \in \mathscr{C}^1([0,1];\mathbb{R}^d), u_0 = x, u_1 = y\}.$$

Definition 5.3 (Riemannian measure)**.** The Riemannian measure (area) of a compact $J \subset \mathbb{R}^d$ with respect to the family $\{f''_x : x \in \mathbb{R}^d\}$, denoted by $A_\gamma(J)$, is defined by:

$$A_\gamma(J) := \int_J (\det f''_x)^{\frac{1}{2}}\, dx \qquad (dx = dx^1 \cdots dx^d).$$

The distance between two sets $U, V \subset \mathbb{R}^d$ with respect to γ is defined as:

$$\mathrm{dist}_\gamma(U,V) = \inf\{\gamma(x,y) : x \in U, y \in V\}.$$

For $x \in \mathbb{R}^d$ and $\rho > 0$, a Riemannian disc $B_\gamma(x;\rho)$ centered at x with radius $\rho > 0$ is the set $\{y \in \mathbb{R}^d : \gamma(x,y) \leqslant \rho\}$, distinguished with an Euclidean disc $B(x;\rho)$ centered at x with radius $\rho > 0$ which is the set $\{y \in \mathbb{R}^d : \|x-y\| \leqslant \rho\}$. A Bregman disc $B_b(x;\rho)$ centered at x with radius $\rho > 0$ is the set $\{y \in \mathbb{R}^d : \mathrm{dist}_B(y;x) \leqslant \rho\}$. For a compact $J \subseteq \mathbb{R}^d$ and $\rho > 0$, we denote by $n(J,\rho)$, $\hat{n}(J,\rho)$ and $k(J,\rho)$ respectively the minimal number of Euclidean, Riemannian and Bregman discs of radius ρ to cover J.

The following asymptotic estimate on the minimal covering number of manifold with Riemannian discs is essential in the proof of (5.12). It is a direct consequence of Lemma 1 in [Gru93a].

Lemma 5.5 (Corollary of [Gru93a, Lemma 1])**.** *Let $J \subset \mathbb{R}^d$ be a compact.*

$$\hat{n}(J,\rho) \sim \frac{\vartheta_d}{\mathscr{K}_d}\Big(\int_J (\det f''_x)^{\frac{1}{2}}\, dx\Big)\rho^{-d}, \quad \text{as} \quad \rho \to 0\ . \tag{5.15}$$

This lemma uses essentially the continuity between the Euclidean metric and Riemannian metric, together with the following result on minimum covering with Euclidean discs proved by Hlawka [Hla49].

Theorem 5.6. *[Hla49, Satz 29] Let J be a convex body in $\mathbb{R}^d$ of measure $v(J)$, then there is a constant ϑ_d independent of J such that:*

$$\vartheta_d = \lim_{\tau \to 0} n(J,\tau)\mathscr{K}_d \tau^d / v(J).$$

We shall need the following theorem on optimum quantization in the proof of (5.13):

Theorem 5.7. *[Gru07, Thm 33.2] Let $J \subseteq \mathbb{R}^d$ a subset of measure $v(J) > 0$. Let q be a positive definite quadratic form on $\mathbb{R}^d$, then:*

$$\inf_{S \subseteq \mathbb{R}^d, |S| = m} \int_J \min_{t \in S}\{q(s-t)\}\, ds \sim \delta_d v(J)^{\frac{d+2}{d}} (detq)^{\frac{1}{d}} \frac{1}{m^{\frac{2}{d}}}, \quad \text{as} \quad \mathrm{m} \to \infty.$$

Finally, we state some useful assertions that can be checked directly.

Let $0 < \rho_0 < 1$. By the strict convexity of f and the compactness of X, there is a compact $\tilde{X} \supset X$ such that

$$\mathrm{dist}_B(y;x) > \rho_0, \qquad \forall y \in X, x \notin \tilde{X}. \tag{5.16}$$

In other words, if there is a set of affine minorants $\{f_i\}_{i=1,\ldots,n}$ such that

$$\delta_X^\infty(f, \sup_i f_i) \leqslant \rho_0,$$

then the contact points $\{x_i\}_{i=1,\dots,n}$ are all in $\tilde{X}$. We may also assume that

$$\gamma(x,y) > \rho_0, \qquad \forall y \in X, x \notin \tilde{X}. \tag{5.17}$$

In the following we fix δ_0 and $\tilde{X}$ satisfying (5.16) and (5.17).

Let $\lambda > 1$. By the continuity of f'', for any $p \in \mathbb{R}^d$, there is an open convex $U \ni p$ such that:

$$\frac{1}{\lambda^2} f''_x \preccurlyeq f''_p \preccurlyeq \lambda^2 f''_x, \qquad \forall \quad x \in U. \tag{5.18}$$

Now let $0 < \delta < \frac{1}{2}\mathrm{dist}_\gamma(p, \partial U)$, then there is an open set V $\subset$ U containing p such that:

$$\mathrm{dist}_\gamma(V, \partial U) > \delta. \tag{5.19}$$

5.4.2 Proof of sup norm error asymptotic estimate

We give the proof of L_∞ error asymptotic L_∞ error estimate (5.12).

Proof. Let $\lambda > 1$. By the compactness of $\tilde{X}$, there are points $p_l \in \tilde{X}$, $l = 1, \dots, m$ with corresponding neighborhoods U_l, V_l and $\rho_l > 0$ such that $\{U_1, \dots, U_m\}$ are all convex and the followings hold:

$$\tilde{X} \subset V_1 \cup \dots V_m, \quad V_l \subset U_l, \quad l = 1, \dots, m \tag{5.20}$$

$$\mathrm{dist}_\gamma(V_l, \partial U_l) > \rho_l, \quad l = 1, \dots, m \tag{5.21}$$

$$\frac{1}{\lambda^2} f''_x \preccurlyeq f''_l \preccurlyeq \lambda^2 f''_x, \qquad \forall x \in U_l, l = 1, \dots, m, \tag{5.22}$$

where $f''_l = f''_{p_l}$. We proceed by proving several claims.

Claim 1:

$$x, y \in U_l, \gamma(x,y) < \mathrm{dist}_\gamma(x, \partial U_l) \Rightarrow \lambda^{-2} \leqslant \gamma^2(x,y)/(x-y)' f''_l (x-y) \leqslant \lambda^2 \ .$$

Let $x, y \in U_l$. By the convexity of U_l, the straight line $u_t = x + t(y-x)$, $t \in [0,1]$ is included in U_l. Hence,

$$\begin{aligned}
\gamma(x,y) &\leqslant \int_0^1 ((x-y)' f''_{u_t} (x-y))^{\frac{1}{2}} \, dt \\
&\leqslant \lambda \int_0^1 ((x-y)' f''_l (x-y))^{\frac{1}{2}} \, dt \qquad \text{by (5.22)} \\
&= \lambda ((x-y)' f''_l (x-y))^{\frac{1}{2}} \ .
\end{aligned}$$

If $\gamma(x,y) < \mathrm{dist}_\gamma(x, \partial U_l)$, then all geodesic lines (with respect to the Riemannian metric) between x and y are included in U_l. Consider a geodesic line $u_t : [0,1] \to U_l$ between x and y, we have:

$$\begin{aligned}
((x-y)' f''_l (x-y))^{\frac{1}{2}} &\leqslant \int_0^1 (\dot{u}'_t f''_l \dot{u}_t)^{\frac{1}{2}} \, dt \\
&\leqslant \lambda \int_0^1 (\dot{u}'_t f''_{u_t} \dot{u}_t)^{\frac{1}{2}} \, dt \qquad \text{by (5.22)} \\
&= \lambda \gamma(x,y) \ .
\end{aligned}$$

We then proved **Claim 1**.

Claim 2:

$$x,y \in U_l, \gamma(x,y) < \mathrm{dist}_\gamma(x,\partial U_l) \Rightarrow \tfrac{1}{2\lambda^4} \leqslant \mathrm{dist}_B(x,y)/\gamma(x,y)^2 \leqslant \tfrac{\lambda^4}{2} \ .$$

Let $x,y \in U_l$. By Taylor's formula, there is $\xi \in [0,1]$ such that:

$$\mathrm{dist}_B(x,y) = \frac{1}{2}(x-y)' f''_{x+\xi(y-x)}(x-y) \tag{5.23}$$

By (5.22) and **Claim 1**:

$$\begin{aligned}
\mathrm{dist}_B(x,y) &= \frac{1}{2}(x-y)' f''_{x+\xi(y-x)}(x-y) \\
&\leqslant \frac{1}{2}\lambda^2 (x-y)' f_l''(x-y) && \text{by (5.22)} \\
&\leqslant \frac{1}{2}\lambda^4 \gamma(x,y)^2 && \text{by } \textbf{Claim 1}.
\end{aligned}$$

The same,

$$\begin{aligned}
\mathrm{dist}_B(x,y) &= \frac{1}{2}(x-y)' f''_{x+\xi(y-x)}(x-y) \\
&\geqslant \frac{1}{2\lambda^2}(x-y)' f_l''(x-y) \\
&\geqslant \frac{1}{2\lambda^4}\gamma(x,y)^2
\end{aligned}$$

We then proved **Claim 2**.

Claim 3:

There is $M > 1$ such that

$$\frac{\gamma(x,y)^2}{\mathrm{dist}_B(x,y)} \leqslant M, \ \ \forall x,y \in \tilde{X}. \tag{5.24}$$

Claim 3 follows directly from **Claim 2** and the compactness of $\tilde{X}$.

Claim 4:

Let $\rho < \min\{\rho_0,\rho_1,\ldots,\rho_m\}^2/M$. If $\{f_i\}_{i=1}^n$ is a set of affine minorants with contact points $\{x_1,\ldots,x_n\}$ such that

$$\delta_X^\infty(f, \sup_{i=1}^n f_i) \leqslant \rho, \tag{5.25}$$

then X is covered by n Riemannian discs with centers $\{x_1,\ldots,x_n\} \subseteq \tilde{X}$ and radius $\lambda^2\sqrt{2\rho}$.

By the definition of Bregman distance, the inequality (5.25) implies that

$$\max_{y\in X} \min_{i=1,\ldots,n} \mathrm{dist}_B(y;x_i) \leqslant \rho.$$

Thus by (5.16) we may assume that $\{x_1,\ldots,x_n\} \subset \tilde{X}$. Moreover, for all $y \in X$, there is $i \in \{1,\ldots,n\}$ such that $\mathrm{dist}_B(y;x_i) \leqslant \rho$. By **Claim 3**, we have:

$$\gamma(x_i,y) \leqslant \sqrt{M\rho} \leqslant \min\{\rho_1,\ldots,\rho_m\}.$$

By (5.20), there is $l \in \{1,\dots,m\}$ such that $x_i \in V_l$. Now using (5.21) we obtain that

$$\gamma(x_i,y) < \rho_l < \mathrm{dist}_\gamma(V_l,\partial U_l) \leqslant \mathrm{dist}_\gamma(x_i,\partial U_l).$$

Thus by **Claim 2**, we obtain that

$$\gamma(x_i,y) \leqslant \lambda^2\sqrt{2\rho}.$$

We then proved **Claim 4**.

Claim 5:

Let $\rho < \min\{\rho_0,\dots,\rho_m\}$. If X is covered by n Riemannian discs of radius ρ with centers $\{x_1,\dots,x_n\}$, then

$$\delta_X^\infty(f,\sup_{i=1}^n f_i) \leqslant \frac{\rho^2\lambda^4}{2}$$

where $f_i(x) = f(x_i) + \langle \nabla f(x_i), x - x_i\rangle$, for $i = 1,\cdots,n$.

For all $y \in X$, there is $i \in \{1,\dots,n\}$ such that $\gamma(x_i,y) \leqslant \rho$. By (5.17) we may assume that $\{x_1,\dots,x_n\} \subset \tilde{X}$. Let $l \in \{1,\dots,m\}$ such that $x_i \in V_l$. Then by (5.21) we obtain that:

$$\gamma(x_i,y) < \rho_l < \mathrm{dist}_\gamma(V_l,\partial U_l) \leqslant \mathrm{dist}_\gamma(x_i,\partial U_l).$$

Now using **Claim 2** we deduce that:

$$\mathrm{dist}_B(x_i,y) \leqslant \frac{\lambda^4}{2}\gamma(x_i,y)^2 \leqslant \frac{\rho^2\lambda^4}{2}.$$

We then proved **Claim 5**.

By **Claim 4** et **Claim 5**, for $\rho > 0$ sufficiently small we have:

$$\hat{n}(X,\sqrt{2\rho}\lambda^2) \leqslant k(X,\rho), \quad k(X,\frac{\rho^2\lambda^4}{2}) \leqslant \hat{n}(X,\rho)$$

i.e. for τ sufficiently small,

$$\hat{n}(X,\sqrt{2\tau}\lambda^2) \leqslant k(X,\tau) \leqslant \hat{n}(X,\frac{\sqrt{2\tau}}{\lambda^2}).$$

By Lemma 5.5, we obtain that:

$$\hat{n}(X,\sqrt{2\tau}\lambda^2) \sim \vartheta_d A_\gamma(X)/\mathscr{K}_d(\sqrt{2\tau}\lambda^2)^d, \quad \text{as } \tau \to 0$$

$$\hat{n}(X,\frac{\sqrt{2\tau}}{\lambda^2}) \sim \vartheta_d A_\gamma(X)/\mathscr{K}_d(\frac{\sqrt{2\tau}}{\lambda^2})^d, \quad \text{as } \tau \to 0$$

Since $\lambda > 1$ is arbitrary, it follows that:

$$k(X,\tau) \sim \vartheta_d A_\gamma(X)/\mathscr{K}_d(2\tau)^{\frac{d}{2}}, \quad \text{as } \tau \to 0.$$

Denote $\beta = (\vartheta_d A_\gamma(X)/\mathscr{K}_d)^{\frac{d}{2}}/2$. We then have

$$\tau k(X,\tau)^{\frac{2}{d}} \sim \beta, \qquad \text{as} \quad \tau \to 0. \tag{5.26}$$

Consider the decreasing sequence $\tau_i = \frac{1}{i}$ with $i \in \mathbb{N}$. Then there is an increasing sequence of integers n_i with $i \in \mathbb{N}$ such that $n_i = k(X, \tau_i)$. Let $n \in [n_i, n_{i+1})$, then

$$\tau_{i+1} < \delta^{\infty}_{X,n}(f) \leqslant \tau_i.$$

It follows that

$$\frac{\tau_{i+1} n_i^{\frac{2}{d}}}{n^{\frac{2}{d}}} < \delta^{\infty}_{X,n}(f) \leqslant \frac{\tau_i n_{i+1}^{\frac{2}{d}}}{n^{\frac{2}{d}}}.$$

Note that (5.26) implies that $\tau_i n_{i+1}^{\frac{2}{d}} \sim \beta$ and $\tau_{i+1} n_i^{\frac{2}{d}} \sim \beta$ as i goes to infinity. In consequence,

$$\delta^{\infty}_{X,n}(f) \sim \frac{\beta}{n^{\frac{2}{d}}}$$

where

$$\beta = \frac{1}{2} \Big(\frac{\vartheta_d}{\mathscr{H}_d} \int_X (\det f'')^{\frac{1}{2}} \, dx \Big)^{\frac{2}{d}} \quad .$$

□

5.4.3 Proof of average error asymptotic estimate

We give the proof of asymptotic L_1 error estimate (5.13).

Proof. Let $\lambda > 1$. By the compactness of X, there are points $p_l \in X$, $l = 1, \ldots, m$ with corresponding neighborhoods U_l, V_l and $\rho_l > 0$ such that $\{U_1, \ldots, U_m\}$ are convex and the followings hold:

$$X \subset V_1 \cup \ldots V_m, \quad V_l \subset U_l, \quad l = 1, \ldots, m \tag{5.27}$$

$$\mathrm{dist}_{\gamma}(V_l, \partial U_l) > \rho_l, \quad l = 1, \ldots, m \tag{5.28}$$

$$\frac{1}{\lambda} f''_x \preccurlyeq f''_l \preccurlyeq \lambda f''_x, \;\; \frac{1}{\lambda} \det f''_x \leqslant \det f''_l \leqslant \lambda \det f''_x \qquad \forall x \in U_l, \;\; l = 1, \ldots, m. \tag{5.29}$$

where $f''_l = f''_{p_l}$. We may assume

$$X = \bigcup_l V_l \;\; ; \quad V_i \cap V_j = \emptyset, \quad i \neq j. \tag{5.30}$$

Let

$$\varepsilon = \min\{\rho_1, \ldots, \rho_m\}.$$

Now for each $l \in \{1, \ldots, m\}$, there is a compact $J_l \subset V_l$ such that

$$\mathrm{dist}(J_l, \mathrm{bd}\, U_l) > \varepsilon \;\; , \tag{5.31}$$

$$\int_{J_l} (\det f_x)^{\frac{1}{d+2}} \, dx \geqslant \frac{1}{\lambda} \int_{V_l} (\det f''_x)^{\frac{1}{d+2}} \, dx \;\; . \tag{5.32}$$

It follows that

$$X \supset \cup_{l=1}^m J_l \;\; ; \quad J_i \cap J_j = \emptyset, \;\; i \neq j, \tag{5.33}$$

and

$$\sum_{l=1}^m \int_{J_l} (\det f_x)^{\frac{1}{d+2}} \, dx \geqslant \frac{1}{\lambda} \int_X (\det f''_x)^{\frac{1}{d+2}} \, dx \;\; . \tag{5.34}$$

Claim 1:

For sufficiently large n,

$$\delta^1_{X,n}(f) \geqslant \frac{\delta}{2\lambda^{\frac{3d+3}{d}}}\Big(\int_X (\det f'')^{\frac{1}{d+2}}\Big)^{\frac{d+2}{d}} \frac{1}{n^{\frac{2}{d}}}. \tag{5.35}$$

Let $\{f_i\}_{i=1,\ldots,n}$ be a set of affine minorants of f given by the contact points $\{s_i : i = 1,\ldots,n\}$ such that

$$\delta^1_{X,n}(f) = \int_X \big(f(x) - \sup_{i=1}^n f_i(x)\big)\, dx.$$

For $i \in \{1,2,\cdots,n\}$, denote:

$$F_i := \{x \in X : \sup_{j=1}^n f_j(x) = f_i(x)\}.$$

For each $l \in \{1,2,\cdots,m\}$, denote:

$$K_{n,l} := \{i : F_i \cap J_l \neq \emptyset\}, \quad k_{n,l} = |K_{n,l}|.$$

By (5.31), we know that for sufficiently large n:

$$s_i \in U_l, \ \ \forall i \in K_{n,l}, \ \ l \in \{1,\ldots,m\}. \tag{5.36}$$

By the strict convexity of f, we know that as n goes to infinity:

$$k_{n,l} \to \infty, \quad \forall l \in \{1,2,\ldots,m\}. \tag{5.37}$$

By (5.33), we know that as n goes to infinity:

$$k_{n,1} + k_{n,2} + \cdots + k_{n,m} \leqslant n. \tag{5.38}$$

The L_1 approximation error on the compact J_l is given by:

$$\begin{aligned}&\int_{J_l} (f(x) - \sup_{i=1}^n f_i(x))\, dx \\ &= \sum_{i \in K_{n,l}} \Big(\int_{F_i \cap J_l} f(x) - f(s_i) - \langle \nabla f(s_i), x - s_i\rangle\Big) dx.\end{aligned}$$

Applying Taylor's formula and using (5.36), the convexity of U_l and (5.29), we know that for all $l = 1,\ldots,m$:

$$\begin{aligned}&\int_{J_l} (f(x) - \sup_{i=1}^n f_i(x))\, dx \\ &\geqslant \sum_{i \in K_{n,l}} \int_{F_i \cap J_l} \frac{1}{2\lambda}(x - s_i)' f_l''(x - s_i)\, dx\end{aligned}$$

Hence:

$$\begin{aligned}
\delta^1_{X,n}(f) &\geqslant \delta^1_X(f, \sup_{i=1}^n f_i) \\
&\geqslant \sum_{l=1}^m \int_{J_l} (f(x) - \sup_{i=1}^n f_i(x))\, dx && \text{by (5.33)} \\
&\geqslant \frac{1}{2\lambda} \sum_{l=1}^m \sum_{i \in K_{n,l}} \int_{F_i \cap J_l} (x-s_i)' f_l''(x-s_i)\, dx \\
&\geqslant \frac{1}{2\lambda} \sum_{l=1}^m \int_{J_l} \min_{i \in K_{n,l}} (x-s_i)' f_l''(x-s_i)\, dx \\
&\geqslant \frac{1}{2\lambda} \sum_{l=1}^m \inf_{\substack{S \subset \mathbb{E}^d \\ |S| = k_{n,l}}} \int_{J_l} \min_{y \in S} (x-y)' f_l''(x-y)\, dx
\end{aligned}$$

Now we apply the asymptotic formula given in Theorem 5.7. For sufficiently large n, as $k_{n,l} \to +\infty$ we have:

$$\begin{aligned}
\delta^1_{X,n}(f) &\geqslant \frac{\delta_d}{2\lambda^2} \sum_l v(J_l)^{\frac{d+2}{d}} (\det f_l'')^{\frac{1}{d}} k_{n,l}^{-\frac{2}{d}} \\
&= \frac{\delta_d}{2\lambda^2} \Big(\frac{k_{n,1} + \cdots + k_{n,m}}{k_{n,1} + \cdots + k_{n,m}}\Big) \sum_{l=1}^m k_{n,l} \big(v(J_l) (\det f_l'')^{\frac{1}{d+2}} k_{n,l}^{-1}\big)^{\frac{d+2}{d}} \\
&= \frac{\delta_d}{2\lambda^2} (k_{n,1} + \cdots + k_{n,m}) \sum_{l=1}^m \frac{k_{n,l} \big(v(J_l) (\det f_l'')^{\frac{1}{d+2}} k_{n,l}^{-1}\big)^{\frac{d+2}{d}}}{k_{n,1} + \cdots + k_{n,m}}
\end{aligned}$$

Recall that for a convex function g from $\mathbb{R}$ to $\mathbb{R}$, for all $y_1, \ldots, y_m \in \mathbb{R}$ and $k_1, \ldots, k_m > 0$ we have:

$$\sum_{l=1}^m \frac{k_l g(y_l)}{k_l + \cdots + k_m} \geqslant g\Big(\sum_{l=1}^m \frac{k_l y_l}{k_1 + \cdots + k_m}\Big).$$

Thus,

$$\begin{aligned}
\delta^1_{X,n}(f) &\geqslant \frac{\delta_d}{2\lambda^2} (k_{n,1} + \cdots + k_{n,m}) \Big(\sum_{l=1}^m \frac{v(J_l) (\det f_l'')^{\frac{1}{d+2}}}{k_{n,1} + \cdots + k_{n,m}}\Big)^{\frac{d+2}{d}} \\
&= \frac{\delta_d}{2\lambda^2} (k_{n,1} + \cdots + k_{n,m})^{-\frac{2}{d}} \Big(\sum_{l=1}^m v(J_l) (\det f_l'')^{\frac{1}{d+2}}\Big)^{\frac{d+2}{d}} \\
&\geqslant \frac{\delta_d n^{-\frac{2}{d}}}{2\lambda^2} \Big(\sum_{l=1}^m \int_{J_l} (\det f_l'')^{\frac{1}{d+2}}\, dx\Big)^{\frac{d+2}{d}} && \text{by (5.38)} \\
&\geqslant \frac{\delta_d n^{-\frac{2}{d}}}{2\lambda^{2+\frac{1}{d}}} \Big(\sum_{l=1}^m \int_{J_l} (\det f_x'')^{\frac{1}{d+2}}\, dx\Big)^{\frac{d+2}{d}} && \text{by (5.29)} \\
&\geqslant \frac{\delta_d n^{-\frac{2}{d}}}{2\lambda^{2+\frac{1}{d}+\frac{d+2}{d}}} \Big(\int_X (\det f_x'')^{\frac{1}{d+2}}\, dx\Big)^{\frac{d+2}{d}} && \text{by (5.34)}
\end{aligned}$$

We then proved **Claim 1**.

Claim 2:

For sufficiently large n,

$$\delta^1_{X,n}(f) \leqslant \frac{\lambda^{\frac{2d+3}{d}}\delta_d n^{-\frac{2}{d}}}{2}\Big(\int_X (\det f''_x)^{\frac{1}{d+2}}\,dx\Big)^{\frac{d+2}{d}} \tag{5.39}$$

For $l = 1,\dots,m$, let

$$\tau_l = \frac{\int_{V_l}(\det f''_x)^{\frac{1}{d+2}}\,dx}{\int_X (\det f''_x)^{\frac{1}{d+2}}\,dx} \tag{5.40}$$

and $k_{n,l} = \lfloor \tau_l n \rfloor$. We have:

$$k_{n,1} + k_{n,2} + \cdots + k_{n,m} \leqslant n \tag{5.41}$$

$$k_{n,l} \geqslant \frac{\tau_l n}{\lambda},\ \ l = 1,\dots,m, \quad \text{as } n \to +\infty \tag{5.42}$$

$$k_{n,l} \to +\infty,\ \ l = 1,\dots,m, \quad \text{as } n \to +\infty \tag{5.43}$$

For each $l = 1,\dots,m$, choose $k_{n,l}$ points $\{s_{n,l,i} | i = 1,2,\cdots,k_{n,l}\} \subseteq \mathbb{R}^d$ such that

$$\int_{V_l} \min_{i\leqslant k_{n,l}} (x - s_{n,l,i})' f''_l (x - s_{n,l,i})\,dx = \inf_{|S|=k_{n,l}} \int_{V_l} \min_{s\in S} (s-x)' f''_l (s-x)\,dx. \tag{5.44}$$

For each point $s_{n,l,i}$, we take the affine function:

$$f_{n,l,i}(x) = f(s_{n,l,i}) + \langle \nabla f(s_{n,l,i}), x - s_{n,l,i}\rangle.$$

Therefore,

$$\begin{aligned}
\delta^1_{X,n}(f) &\leqslant \int_X f(x) - \sup_{l,i} f_{n,l,i}(x)\,dx \\
&= \sum_{l=1}^m \int_{V_l} f(x) - \sup_{l,i} f_{n,l,i}(x)\,dx \qquad \text{by (5.30)} \\
&\leqslant \sum_{l=1}^m \int_{V_l} f(x) - \sup_{i\leqslant k_{n,l}} f_{n,l,i}(x)\,dx \\
&= \sum_{l=1}^m \int_{V_l} \min_{i\leqslant k_{n,l}} f(x) - f(s_{n,l,i}) - \langle \nabla f(s_{n,l,i}), x - s_{n,l,i}\rangle\,dx.
\end{aligned}$$

Again by Taylor's formula, the last term equals to

$$\sum_{l=1}^m \int_{V_l} \min_{i\leqslant k_{n,l}} \frac{1}{2}(x - s_{n,l,i})' f''_{s_{n,l,i}+\xi(x)(x-s_{n,l,i})}(x - s_{n,l,i})\,dx$$

where $\xi : \mathbb{R}^d \to [0,1]$. By (5.28), we know that for sufficiently large n, $s_{n,l,i} \in U_l$ for all $i = 1,\dots,k_{n,l}$. Therefore we deduce from the convexity of U_l that $s_{n,l,i} + \xi(x)(x - s_{n,l,i}) \in U_l$ for all $x \in V_l$, $l \in \{1,\dots,m\}$ and $i \in \{1,\dots,k_{n,l}\}$. Now using (5.29) we get:

$$\begin{aligned}
\delta^1_{X,n}(f) &\leqslant \sum_{l=1}^m \int_{V_l} \min_{i\leqslant k_{n,l}} \frac{1}{2}(x - s_{n,l,i})' f''_{s_{n,l,i}+\xi(x)(x-s_{n,l,i})}(x - s_{n,l,i})\,dx \\
&\leqslant \tfrac{\lambda}{2} \sum_{l=1}^m \int_{V_l} \min_{i=1\leqslant k_{n,l}} (x - s_{n,l,i})' f''_l (x - s_{n,l,i})\,dx
\end{aligned}$$

Now by the asymptotic formula in Theorem 5.7, we get:

$$\delta^1_{X,n}(f) \leqslant \frac{\lambda^2\delta_d}{2}\sum_{l=1}^{m} v(V_l)^{\frac{d+2}{d}}(\det f_l'')^{\frac{1}{d}}k_{n,l}^{-\frac{2}{d}}$$

Therefore,

$$\begin{aligned}
\delta^1_{X,n}(f) &\leqslant \frac{\lambda^{2+\frac{2}{d}}\delta_d}{2}\sum_l v(V_l)^{\frac{d+2}{d}}(\det f_l'')^{\frac{1}{d}}(\tau_l n)^{-\frac{2}{d}} \\
&= \frac{\lambda^{\frac{2d+2}{d}}\delta_d}{2}\sum_l \big(v(V_l)(\det f_l'')^{\frac{1}{d+2}}\big)^{\frac{d+2}{d}}(\tau_l n)^{-\frac{2}{d}} \\
&\leqslant \frac{\lambda^{\frac{2d+2}{d}+\frac{1}{d}}\delta_d}{2}\sum_l \Big(\int_{V_l}(\det f_x'')^{\frac{1}{d+2}}\,dx\Big)^{\frac{d+2}{d}}(\tau_l n)^{-\frac{2}{d}} && \text{by (5.29)} \\
&= \frac{\lambda^{\frac{2d+3}{d}}\delta_d n^{-\frac{2}{d}}}{2}\sum_l \Big(\int_{V_l}(\det f_x'')^{\frac{1}{d+2}}\Big)\Big(\int_X (f_x'')^{\frac{1}{d+2}}\,dx\Big)^{\frac{2}{d}} && \text{by (5.40)} \\
&= \frac{\lambda^{\frac{2d+3}{d}}\delta_d n^{-\frac{2}{d}}}{2}\Big(\int_X(\det f_x'')^{\frac{1}{d+2}}\,dx\Big)^{\frac{d+2}{d}}
\end{aligned}$$

Then we proved **Claim 2**.

Since (5.35) and (5.39) are shown for arbitrary $\lambda > 1$, we deduce the asymptotic formula (5.13). □

CHAPTER 6

A refinement of McEneaney's curse of dimensionality free method

The curse of dimensionality free method, introduced by McEneaney for infinite horizon switched optimal control problems, is a special interesting class of max-plus basis methods by its cubic growth rate with respect to the dimension of the state space. In this chapter we focus on the algorithmic aspects of McEneaney's curse of dimensionality free method. We show that the optimal pruning problem, which is a critical step in the implementation of the method, can be formulated as a continuous version of the facility location or k-center combinatorial optimization problems, in which the connection costs arise from a Bregman distance. We derive from our approach a refinement of the curse of dimensionality free method introduced previously by McEneaney, with a higher accuracy for a comparable computational cost.

This chapter extends the algorithmic part of the conference article [GMQ11].

6.1 Introduction

In the previous chapter, we established a negative result showing that the curse of dimensionality is inherent to the family of max-plus basis methods based on c-semiconvex transforms. However, this theoretical negative result is contrasted by the experimental efficiency of McEneaney's curse of dimensionality free method, firstly developed in [McE07] (see also [MK10, MDG08, SGJM10]), which often give approximations of an acceptable accuracy for a modest amount of basis functions.

In its original form [McE07], the method applies to an infinite-horizon optimal switching problem involving M linear quadratic models such that the corresponding HJ PDE is written as:

$$0 = H(x, \nabla V) = \max_{m \in \mathscr{M}} \{H^m(x, \nabla V)\} \tag{6.1}$$

where $\mathscr{M} = \{1, 2, \cdots, M\}$ and each H^m is a linear/quadratic form, originating from a linear quadratic optimal control problem:

$$H^m(x, p) = (A^m x)' p + \frac{1}{2} x' D^m x + \frac{1}{2} p' \Sigma^m p,$$

where (A^m, D^m, Σ^m) are matrices meeting certain conditions.

The solution of (6.1) is approximated by iterating a finite-horizon semigroup until a large enough propagation horizon is reached. This finite-horizon semigroup itself is approximated by a semigroup for a system where the switch is only allowed to happen at the integer multiples of a time step τ. The value function V is then approximated by a supremum of quadratic forms which are obtained by solving Riccati equations. If we keep all the other parameters fixed, the growth of the execution time is only cubic as the dimension grows, related to the solution of Riccati equations. However, the number of quadratic forms is multiplied by the number of systems M at each iteration thus the complexity still grows exponentially as the required accuracy tends to zero. Hence the curse of dimensionality is replaced by a *curse of complexity*.

In [MDG08], the method has been extended directly to a larger problem class where the Hamiltonian is now given or well-approximated by pointwise maximum of linear quadratic functions (possibly with linear terms). More specifically, the corresponding HJ PDE is now

$$0 = H(x, \nabla V) = \max_{m \in \mathscr{M}} \{H^m(x, \nabla V)\} \ , \tag{6.2}$$

where $\mathscr{M} = \{1, 2, \cdots, M\}$ and each H^m has the form:

$$H^m(x, p) = (A^m x)' p + \frac{1}{2} x' D^m x + \frac{1}{2} p' \Sigma^m p + (l_1^m)' x + (l_2^m)' p + \alpha^m \ ,$$

where $(A^m, D^m, \Sigma^m, l_1^m, l_2^m, \alpha^m)$ are parameters of proper dimension meeting certain conditions. The motivation for this problem class is that pointwise maximum of quadratic functions possibly with linear terms can approximate, arbitrarily closely, any semiconvex function. In [MDG08], the authors developed an SDP based pruning method in order to attenuate the curse of complexity. In this way, high dimensional instances (with state dimensions from 6 to 15) inaccessible by other methods could be solved [MDG08, SGJM10].

In this chapter, we focus our attention on the algorithmic aspects of McEneaney's curse of dimensionality free method. In Section 6.2, we restate the problem and the assumptions in order to make the thesis self-contained. In Section 6.3 we review the principle of the method and the SDP based pruning algorithm proposed in [MDG08]. Then, we show in Section 6.4.2 that the optimal pruning problem can be formulated as a continuous version of the k-median or k-center problem, depending on the choice of the norm. The discrete versions of these problems are NP-hard. Hence, we propose several heuristics (combining facility location heuristics and Shor SDP relaxation scheme). Experimental results are given in Section 6.5, showing that by combining the primal version of the method with improved pruning algorithms, a higher accuracy is reached for a similar running time, by comparison with [McE07, MDG08].

6.2 Problem class

We consider the optimal control problem for switched linear quadratic system studied in [MDG08] (see also [McE07, MK10]). Let d be the dimension of the state space and k be the dimension of the control space. Let $\mathscr{M} = \{1,2,\dots,M\}$. For each $m \in \mathscr{M}$, there is a linear quadratic optimal control problem with matrix parameters given by $A^m, D^m \in \mathbb{R}^{d\times d}$, $\sigma^m \in \mathbb{R}^{d\times k}$, $l_1^m, l_2^m \in \mathbb{R}^d$, and $\gamma, \alpha^m \in \mathbb{R}$. The infinite horizon switched optimal control problem is:

Problem 6.1.

$$V(x) = \sup_{\mathbf{u}\in W}\sup_{\mu\in\mathscr{D}_\infty}\sup_{T<\infty} J(x,T;\mathbf{u},\mu)$$

where

$$\begin{aligned}
&J(x,T;\mathbf{u},\mu) = \int_0^T L^{\mu(s)}(\mathbf{x}(s)) - \frac{\gamma^2}{2}|\mathbf{u}(s)|^2\,ds \ , \\
&\mathscr{D}_\infty \doteq \{\mu : [0,\infty) \to \mathscr{M} : \text{measurable}\} \ , \\
&W \doteq L_2^{\text{loc}}([0,\infty);\mathbb{R}^k) \ , \\
&L^m(x) = \frac{1}{2}x'D^m x + (l_1^m)'x + \alpha^m, \ \ m \in \mathscr{M} \ .
\end{aligned}$$

and the state dynamics are given by

$$\dot{\mathbf{x}}(s) = A^{\mu(s)}\mathbf{x}(s) + \sigma^{\mu(s)}\mathbf{u}(s) + l_2^{\mu(s)}; \ \ \mathbf{x}(0) = x \in \mathbb{R}^d \ . \tag{6.3}$$

The problem has its origin in H_∞ control of nonlinear systems, see [Sor96, McE98, Ali11]. The control $(\mathbf{u}(\cdot),\mu(\cdot))$ here should be identified with the disturbance term in a H_∞ control system without active control. The function $L^{(\cdot)}(\cdot) : \mathscr{M}\times\mathbb{R}^d \to \mathbb{R}$ corresponds to the *output* or *response* in H_∞ control system. The parameter γ is the H_∞ *attenuation bound* and the value function V is the *available storage*.

Denote by S_t the evolution semigroup associated to Problem 6.1. That is, for a function $\phi : \mathbb{R}^d \to \mathbb{R}$ and $x \in \mathbb{R}^d$, we have

$$S_t[\phi](x) = \sup_{\mathbf{u}\in W_t}\sup_{\mu\in\mathscr{D}_t} J(x,t;\mathbf{u},\mu) + \phi(\mathbf{x}(t))$$

where

$$\mathscr{D}_t \doteq \{\mu : [0,t] \to \mathscr{M} : \text{measurable}\} \ , \tag{6.4}$$

$$W_t \doteq L_2^{\text{loc}}([0,t];\mathbb{R}^k) \ , \tag{6.5}$$

and $\mathbf{x}(\cdot) : [0,t] \to \mathbb{R}^d$ absolutely continuous satisfies

$$\dot{\mathbf{x}}(s) = A^{\mu(s)}\mathbf{x}(s) + \sigma^{\mu(s)}\mathbf{u}(s) + l_2^{\mu(s)}, \ \ s \in [0,t]; \ \ \mathbf{x}(0) = x \ .$$

For every $m \in \mathscr{M}$, denote

$$\Sigma^m = \frac{1}{\gamma^2}\sigma^m(\sigma^m)' \ .$$

The corresponding Hamiltonian is:

$$H(x,p) = \max_{m\in\mathscr{M}}\{H^m(x,p)\}, \tag{6.6}$$

where for each m, H^m has the form:

$$H^m(x,p)=\frac{1}{2}x'D^m x+\frac{1}{2}p'\Sigma^m p+(A^m x)'p+(l_1^m)'x+(l_2^m)'p+\alpha^m \ . \tag{6.7}$$

The concept of H is constructed so as to resemble the Hamiltonian of some given nonlinear control problem which has a finite solution. We suppose that H is an approximation of $\tilde{H}$ and problem

$$0=-\tilde{H}(x,\nabla V),\ \ V(0)=0 \tag{6.8}$$

has finite value. Besides, the following assumptions are first made in [MDG08].

Assumption 6.1.

- Assume there exists unique viscosity solution $\tilde{V}$ to (6.8) in Q_K for some $K\in(0,\infty)$, where

  $$Q_K=\{\phi:\mathbb{R}^d\to\mathbb{R}:\phi \text{ is semiconvex and } 0\leqslant\phi(x)\leqslant K/2|x|^2,\ \ \forall x\in\mathbb{R}^d\}$$

  is the domain of semigroup $(S_t)_t$.

- Assume that

  $$H(x,p)\leqslant\tilde{H}(x,p),\quad x,p\in\mathbb{R}^d \ .$$

- Assume there exists $c_A>0$ such that

  $$x'A^m x\leqslant -c_A|x|^2,\quad \forall m\in\mathscr{M},x\in\mathbb{R}^d \ .$$

- Assume $H^1(x,p)$ has coefficients satisfying the following: $l_1^1=l_2^1=0$; $\alpha^1=0$; $D^1\succ 0$; and $\gamma^2/c_\sigma^2>c_D/c_A^2$, where c_D is such that $D^1\preccurlyeq c_D I_d$ and $c_\sigma:=|\sigma^1|$.

- Assume that system (6.3) is controllable in the sense that given $x,y\in\mathbb{R}^d$ and $T>0$, there exist processes $\mathbf{u}\in W$ and μ measurable with range in $\mathscr{M}$ such that $\mathbf{x}(T)=y$ when $\mathbf{x}(0)=x$ and one applies controls $\mathbf{u}$ and μ.

- Assume there exist $c_1,c_2<\infty$ such that for all $\varepsilon\in(0,1]$, $x\in\mathbb{R}^d$, and all ε-optimal pair $(\mu^\varepsilon,\mathbf{u}^\varepsilon)$ for problem 6.1 with initial state x, one has

  $$\|\mathbf{u}^\varepsilon\|^2_{L_2[0,T]}\leqslant c_1+c_2|x|^2,\ \ \forall T>0 \ .$$

The following theorem shows the existence of the value function and the convergence of the semigroup as time horizon tends to infinity.

Theorem 6.2 ([McE09])**.** *Under Assumption 6.1, the value function V defined in Problem 6.1 is the unique continuous solution of $V=S_T[V]$ in the class Q_K for any $T>0$. Besides, V is also the unique viscosity solution in the class Q_K of the static HJ PDE:*

$$0=-H(x,\nabla V),\ \ \forall x\in\mathbb{R}^d;\ \ V(0)=0 \ . \tag{6.9}$$

Further, given any $V_0\in Q_K$ such that $0\leqslant V_0\leqslant V$, we have

$$V=\lim_{T\to\infty}S_T[V_0] \tag{6.10}$$

uniformly on compact sets.

6.3 Principle of the algorithm

6.3.1 Single semigroup operator

For every $m \in \mathscr{M}$, define the semigroup $(S_t^m)_t$ as follows. For a function $\phi : \mathbb{R}^d \to \mathbb{R}$ and $x \in \mathbb{R}^d$, we have

$$S_t^m[\phi](x) := \sup_{\mathbf{u} \in W} J(x,t;\mathbf{u},m) + \phi(\mathbf{x}(t))$$

where $\mathbf{x}(\cdot) : [0,t] \to \mathbb{R}^d$ satisfies

$$\dot{\mathbf{x}}(s) = A^m \mathbf{x}(s) + \sigma^m \mathbf{u}(s) + l_2^m, \;\; s \in [0,t]; \;\; \mathbf{x}(0) = x \;\; .$$

Then it is clear that for all function ϕ, all $t > 0$ and $m \in \mathscr{M}$, we have $S_t[\phi] \leqslant S_t^m[\phi]$. Since we assume Q_K as the domain of S_t, we know that Q_K is also the domain of S_t^m.

6.3.2 Computation of single semigroup operator

The propagation of a quadratic function ϕ by S_τ^m reduces to solving a differential Riccati equation (DRE). Suppose there are only quadratic terms, i.e., $l_1^m = 0, l_2^m = 0, \alpha^m = 0$. Let $\phi(x) = \frac{1}{2} x^T P_0 x$, then $S_t^m[\phi](x) = \frac{1}{2} x^T P(t) x$, where $P(t)$ satisfies the following differential Riccati equation

$$\dot{P}(t) = (A^m)' P(t) + P(t) A^m + P(t) \Sigma^m P(t) + D^m, \;\; P(0) = P_0. \tag{6.11}$$

Moreover, it is well-known that one can recover the solution of a DRE from a system of Hamiltonian linear differential equations (see, e.g., [Rei72]). More specifically, the solution of (6.11) also satisfies $P(t) = Y(t)X(t)^{-1}$ and $(X(t), Y(t))$ are the solution of:

$$\begin{cases} \begin{pmatrix} \dot{X} \\ \dot{Y} \end{pmatrix} = \begin{pmatrix} -A^m & -\Sigma^m \\ D^m & (A^m)' \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} \\ \\ X(0) = I_d, \quad Y(0) = P_0 \end{cases} . \tag{6.12}$$

We denote by $\mathscr{A}$ the matrix coefficient in the above linear system. Note that the invertibility of $X(t)$ can be derived from the fact that the value function V is finite. Given a fixed time step $\tau > 0$, the fundamental solution $\exp(\mathscr{A}\tau)$ of the previous linear system satisfies:

$$\begin{pmatrix} X(\tau) \\ Y(\tau) \end{pmatrix} = \exp(\mathscr{A}\tau) \begin{pmatrix} I_d \\ P_0 \end{pmatrix} \;\; .$$

In the presence of linear or constant terms in the control system or in the quadratic function, the problem can be easily transformed into a purely quadratic one by adding a constant state variable. The above analysis shows that, given a fixed propagation time τ, computing $S_\tau^m[\phi]$, for every quadratic form ϕ, reduces to a matrix multiplication and an inverse operation, which can be done in $O(d^3)$ incremental time.

6.3.3 Max-plus based approximation

We review the basic steps of the max-plus algorithm proposed in [McE07, MDG08] to approximate the value function V. Choose an initial function V_0 inferior to the value function V. First we approximate V by $S_T[V_0]$ for some sufficiently large T. We then choose a time-discretization step $\tau > 0$

and a number of iterations N such that $T = N\tau$. For each $k = 0, \dots, N$, denote V_h^k the approximation of $S_{k\tau}[V_0]$. The algorithm iterates as follows. For $k = 1, \dots, N$,

$$S_{k\tau}[V_0] \simeq V_h^k = \tilde{S}_\tau[V_h^{k-1}] := \sup_{m \in \mathscr{M}} S_\tau^m[V_h^{k-1}].$$

It follows that if V_h^{k-1} is the pointwise maximum of $|n_{k-1}|$ quadratic functions, then V_h^k is the pointwise maximum of $M|n_{k-1}|$ quadratic functions. At the end of N iterations, we get our approximated value function represented by:

$$V \simeq S_{N\tau}[V_0] \simeq \{\tilde{S}_\tau\}^N[V_0] = \sup_{i_1, \cdots, i_N \in \mathscr{M}} S_\tau^{i_N} \cdots S_\tau^{i_1}[V^0].$$

If we choose V_0 as a quadratic function, then the approximated value function at the end of N iterations will be the supremum of $|\mathscr{M}|^N$ quadratic functions. The computational growth in the space dimension is cubic, as shown in Section 6.3.2. However, the number of quadratic forms grows by a factor of M at each iteration. This so-called *curse of complexity* can be reduced by performing a *pruning process* at each iteration of the algorithm to remove some quadratic functions. More precisely, let $F = \{1, 2, \dots, n_f\}$ and $\{\phi_i\}_{i \in F}$ be a finite set of quadratic functions and

$$\phi = \sup_{i \in F} \phi_i. \tag{6.13}$$

A pruning operation $\mathscr{P}$ applied to ϕ produces a subapproximation of ϕ by selecting a subset $J \subset F$:

$$\phi \simeq \mathscr{P}[\phi] = \sup_{j \in J} \phi_j.$$

If we take into account the pruning procedure, then the value function V is approximated by

$$V \simeq S_{N\tau}[V_0] \simeq \{\mathscr{P} \circ \tilde{S}_\tau\}^N[V_0] \ .$$

Remark 6.1. In [McE07] and [MDG08], the approximated semigroup is propagated in a dual space. Here we consider the *primal* curse of dimensionality free method: it is equivalent if no pruning is performed, but it avoids the use of dual representations.

6.3.4 Error bound of the algorithm

We restate the error bound of the algorithm proved by McEneaney in [McE09]. For this, we need to add two more additional assumptions.

Assumption 6.3. Assume

$$\sigma^m = \sigma, \ \ m \in \mathscr{M} \ .$$

Assumption 6.4. Assume there exist $\underline{T}$, $c_1 > 0$ such that for all $x \in \mathbb{R}^d$, all $\varepsilon \in (0, 1]$, and all $\mu^\varepsilon, \mathbf{u}^\varepsilon$ which are ε-optimal for Problem 6.1, one has

$$\int_0^T L^{\mu^\varepsilon(t)}(\mathbf{x}^\varepsilon(t))\, dt \geqslant c_1 \int_0^T |\mathbf{x}^\varepsilon(t)|^2\, dt, \ \ \forall T \geqslant \underline{T} \ ,$$

where $\mathbf{x}^\varepsilon$ satisfies $\dot{\mathbf{x}}^\varepsilon(t) = A^{\mu^\varepsilon(t)}\mathbf{x}^\varepsilon(t) + \sigma^{\mu^\varepsilon(t)}\mathbf{u}^\varepsilon(t) + l_2^{\mu^\varepsilon(t)}, \mathbf{x}(0) = x$.

Theorem 6.5. *[McE09] Under Assumption 6.1 and Assumption 6.4, given any $V_0 \in Q_K$ such that $0 \leqslant V_0 \leqslant V$, there is $K_1 > 0$ such that we have*

$$0 \leqslant V(x) - S_T[V_0](x) \leqslant K_1/T(1+|x|^2), \ \ \forall x \in \mathbb{R}^d, T > \underline{T}. \tag{6.14}$$

Under Assumption 6.1 and 6.3, there is a constant $K_2 > 0$ such that for sufficiently small $\tau > 0$

$$0 \leqslant S_T[V_0](x) - \{\tilde{S}_\tau\}^N[V_0](x) \leqslant K_2(M+1)^4(1+|x|^2)(1+T)\tau, \ \ \forall x \in \mathbb{R}^d, T > 0. \tag{6.15}$$

We will see in Chapter 7 that Assumption 6.3 is not necessary to obtain (6.15) (see Section 7.8.1).

6.4 SDP based pruning algorithms

We have seen that the number of quadratic functions grows exponentially with respect to the number of iterations N. To reduce this curse of complexity, a pruning procedure is needed at each iteration. Some SDP relaxation based pruning method was proposed in [MDG08] to reduce the number of quadratic forms. After a quick review of their pruning algorithm, we discuss improvements of this pruning method, still partly SDP based, but now exploiting the combinatorial nature of the problem. Let $F = \{1, 2, \ldots, n_f\}$ and $\{\phi_i\}_{i \in F}$ be a finite set of quadratic functions. Suppose that n_f is too large and we want to select some of the quadratic functions in order to approximation the function ϕ defined in (6.13).

6.4.1 SDP based pruning method

We review in this subsection the pruning algorithm proposed in [MDG08]. Roughly speaking, to each basis function $\phi_j(x)$ we associate an *importance metric* :

$$\nu_j = \max_{x \in \mathbb{R}^d} \min_{j' \neq j} (\phi_j(x) - \phi_{j'}(x))/(1+|x|^2) \ \ . \tag{6.16}$$

Then ν_j is the normalized L_∞ error caused by pruning the function $\phi_j(x)$. In some sense the bigger ν_j is, the more useful the function $\phi_j(x)$ is. In particular, when $\nu_j \leqslant 0$ the function $\phi_j(x)$ is dominated by the others and it can be pruned without generating any approximation error. Let

$$Q_j^{j'} = \frac{1}{2}\begin{bmatrix} c_j - c_{j'} & b_j^T - b_{j'}^T \\ b_j - b_{j'} & A_j - A_{j'} \end{bmatrix} = Q_j - Q_{j'} \ \ .$$

The problem in (6.16) is equivalent to:

$$\nu_j = \max_{\nu \in \mathbb{R}; y \in \mathbb{R}^{d+1}} \{\nu : y_1 \neq 0; \|y\| = 1; y^T Q_j^{j'} y \geqslant \nu, \forall j' \neq j\}. \tag{6.17}$$

This nonconvex QCQP(quadratically constrained quadratic program) [BV04] has its SDP relaxation given by:

$$\overline{\nu}_j = \max_{\nu \in \mathbb{R}, Y \succcurlyeq 0} \left\{ \nu \ \middle| \ \begin{array}{ll} Y_{11} > 0; & \mathrm{Tr}(Y) = 1; \quad Y \succcurlyeq 0; \\ \mathrm{Tr}(Y Q_j^{j'}) \geqslant \nu, & \forall j' \neq j. \end{array} \right\}. \tag{6.18}$$

Then $\overline{\nu}_j$ is an upper bound of the importance metric ν_j. In [MDG08], the authors proposed to sort all the upper bounds $\{\nu_j : j \in F\}$ and to pick up the k first ones. We call their method the *sort upper bound* method.

6.4.2 Reduction of pruning to k-center and k-median problems for a Bregman type distance

We first give a general formulation for the pruning problem appearing in the curse of dimensionality free methods. To measure the approximation error, we introduce a Bregman type distance $\mathrm{dist}_\phi(x;j)$ between each point $x \in \mathbb{R}^d$ and each basis function $\phi_j(\cdot)$, such that for all $x \in \mathbb{R}^d$ the following two conditions hold:

$$\begin{aligned}&\exists j_0 \in F, \text{ s.t. } \mathrm{dist}_\phi(x;j_0) = 0;\\ &i,j \in F, \ \mathrm{dist}_\phi(x;i) \leqslant \mathrm{dist}_\phi(x;j) \Leftrightarrow \phi_j(x) \leqslant \phi_i(x)\end{aligned}$$

In other words, the distance $\mathrm{dist}_\phi(x;j)$ measures the loss at point x caused when approximating $\phi(\cdot)$ by $\phi_j(\cdot)$. For example, the simplest choice is to let $\mathrm{dist}_\phi(x;j) = \phi(x) - \phi_j(x)$. Consider a compact set $X \subset \mathbb{R}^d$ on which we measure the loss. One may minimize the total loss (L_1 metric) or the maximal loss (L_∞ metric) on X.

- L_1 metric and k-median problem

$$\delta_k^1(\phi) = \min_{S \subset F, |S|=k} \int_X [\min_{j \in S} \mathrm{dist}_\phi(x;j)]dx \ . \tag{6.19}$$

- L_∞ metric and k-center problem

$$\delta_k^\infty(\phi) = \min_{S \subset F, |S|=k} \max_{x \in X} [\min_{j \in S} \mathrm{dist}_\phi(x;j)] \ . \tag{6.20}$$

We recognize in (6.19) and (6.20) the classical k-median and the k-center facility location problem with continuous demand area and discrete service points. The facility location problem, discrete or continuous, is known to be NP-hard even with euclidean distance. Besides, we remark that a subproblem of Problem (6.19) is the volume computation for polytopes, which is known to be $\#P$-hard. To the best of our knowledge, the only few references that discuss this general class of location problems replace the continuous demand with a discrete one with large number of points, see [DD97]. In the following, we consider a specific case and propose a method based on SDP relaxation to generate discrete points.

6.4.3 Refinements of SDP based pruning method

We consider the following normalized Bregman type distance function, following [MDG08],

$$\mathrm{dist}_\phi(x;j) = \frac{\phi(x) - \phi_j(x)}{1 + |x|^2} \ .$$

The SDP relaxation (6.18) provides not only an upper bound on the importance metric but also a rather simple way to generate feasible solutions.

Suppose that $(\overline{Y}, \overline{v})$ is a solution of program (6.18). We use the randomization technique [Fer00] to get feasible points: we pick up y as a Gaussian random variable satisfying $y \sim \mathcal{N}(0, \overline{Y})$. Then over this distribution, in (6.17) the constraints are satisfied and the maximum is reached on average. By sampling y a sufficient number of times, we get a y close to the optimal solution such that the inequality constraints in (6.17) are all satisfied. Then, setting $x = (y_2/y_1, \ldots, y_{d+1}/y_1)'$ provides a

lower bound of (6.16). The proposed procedure provides in practice a good lower bound, although there is no theoretical guarantee in the present generality.

For each $j \in F$, we sample an equal number of points following the optimal solution of the corresponding SDP program. By this randomization technique, we get a discrete set X' which in some sense reflect rather well the importance of each basis function. We replace the compact set X by this discrete set X' and seek to minimize the total loss on X'. This gives the discrete k-median problem:

$$\delta = \min_{S \subset F, |S|=k} \sum_{x \in X'} [\min_{j \in S} \mathrm{dist}_\phi(x; j)] \ . \tag{6.21}$$

This central problem in combinatorial optimization has seen a succession of papers designing approximations algorithms.

Based on the above observation, we propose the following pruning algorithms.

6.4.3.a 'sort lower bound' For each basis function $\phi_j(x)$ we calculate the lower bound $\underline{\nu}_j$ of the importance metric by:

$$\underline{\nu}_j = \max_{x \in X'} \min_{j' \neq j} (\phi_j(x) - \phi_{j'}(x))/(1+|x|^2) \ .$$

Then the *sort lower bound* method consists in sorting all of the lower bounds $\{\underline{\nu}_j, j \in F\}$ and keep the k first ones.

Our two last pruning methods are merely two heuristics for the k-median problem (6.21).

6.4.3.b 'J-V facility location' Lin and Vitter [LV92] proved that the constant factor approximation for general k-median problem is NP-hard. For metric distance, Jain and Vazirani [Vaz01] proposed a primal-dual 6-approximation algorithm. This algorithm is interesting not only due to its constant factor, but also because it is combinatorial (there is no need to solve a linear program). Although the constant factor approximation no longer holds for the present Bregman type distances (which are not metric in the usual sense), we implemented the primal-dual algorithm for the sake of comparison. The execution time of the primal-dual algorithm is $O(m(\log m)(\log |X'|L))$, here $m = |X'|n_f$ and $L = \max\{\mathrm{dist}_\phi(x; j) : x \in X', j \in F\}$. In the present context, it is used as an heuristic (without any bound estimates).

6.4.3.c 'greedy facility location' The fourth method is the greedy heuristic. Remember that the function to be minimized in the facility location problem is supermodular, which implies that the greedy heuristic has a bound estimate (even without the triangular inequality on the distance function). Let δ be the optimal value of (6.21) and δ_G be the value of a particular solution constructed by the greedy heuristic, then we have [NWF78]:

$$\delta_G \leqslant (1-\alpha^k)\delta + \alpha^k(\max_{j \in F} \sum_{x \in X'} \mathrm{dist}_\phi(x; j)), \tag{6.22}$$

where $\alpha = \frac{k-1}{k}$. The execution time of the greedy heuristic is $O(km)$.

6.5 Experimental results

6.5.1 Problem instance

To compare with the sort upper bound pruning technique (Section 6.4) proposed in [MDG08], we use the same instance tested in [MDG08] with parameters chosen so that the problem shows a complex behavior. The dimension d of the instance is 6 and the number of discrete controls M is 6.

6.5.2 Backsubstitution error

Without the exact value function, we do not have a direct error estimation. Recall that the value function V is the unique viscosity solution of the following HJ equation:

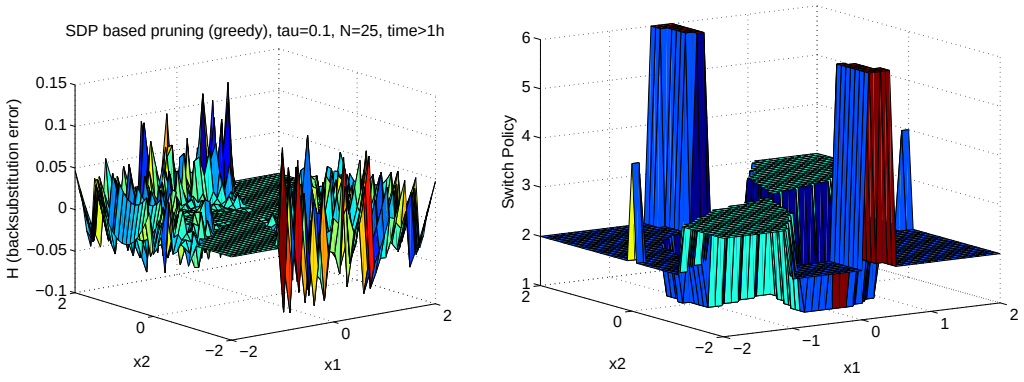
$$0 = -H = -\max_{m\in\mathscr{M}}\{H^m(x,\nabla V)\}, \tag{6.23}$$

where H^m is defined in (6.7). The value of Hamiltonian is then used to measure the approximation and we refer to it as the *backsubstitution error*.

6.5.3 Numerical results

We try different time step: $\tau = 0.1$ and$\tau = 0.05$. The overpruning threshold is the same as in [MDG08]: at iteration k we keep at most $20+6k$ quadratic functions. All of our results[1] are shown along the x_1-x_2 axes with the 4 other coordinates of x set to 0.

Figure 6.1 shows the value of Hamiltonian H at the end of 25 iterations, with $\tau = 0.1$ and using the greedy pruning algorithm (see Section 6.4.3.c). Comparing with the error plot shown in [MDG08], which is in the same scale but has a peak of error of order 1 (versus 0.15 here), we see that the primal max-plus basis method yields a small improvement.



(a) Backsubstitution error

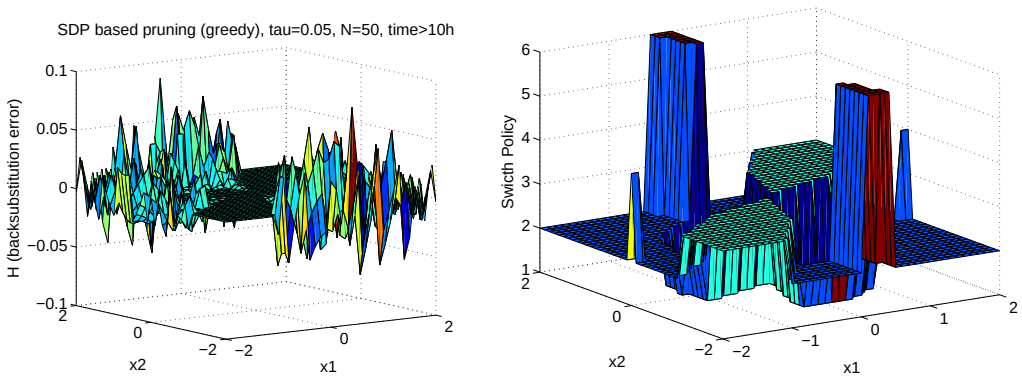
(b) Optimal control policy

Figure 6.1: Visualization of backsubstitution error and control policy (switch) on the plane x_1-x_2, square $[-2,2]\times[-2,2]$, $\tau = 0.1$, $N = 25$, with *greedy* pruning method

Figure 6.2 shows the backsubstitution error at the end of 50 iterations, with smaller discretization step $\tau = 0.05$ and using the greedy facility location algorithm (see Section 6.4.3.c).

Figure 6.3 compares the four pruning methods with $\tau = 0.1$ and $\tau = 0.05$. They both show that the *sort lower bound* and the *greedy facility location* pruning method are better than the two others.

[1]The code was mostly written in Matlab (version 7.11.0.584), calling YALMIP (version 3) and SeDuMi (version 1.3) for the resolution of SDP programs. The computation of the distance function and Jain & Vazirani's primal dual algorithm were written in C++. The results were obtained on a single core of an Intel quad core running at 2.66GHz, with 8Gb of memory.



(a) Backsubstitution error

(b) Optimal control policy

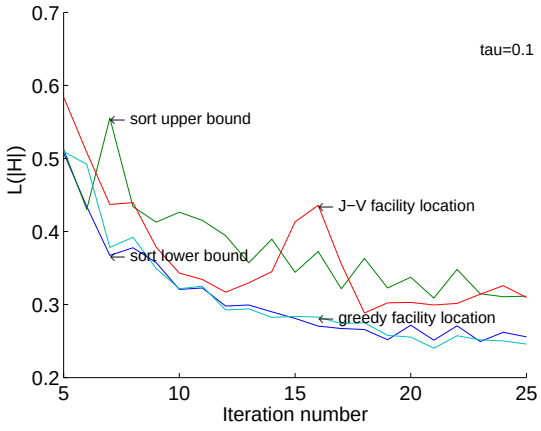
Figure 6.2: Visualization of backsubstitution error and control policy (switch) on the plane x_1-x_2, square $[-2,2]\times[-2,2]$, $\tau = 0.05$, $N = 50$, with *greedy* pruning method

Table 6.1: CPU time

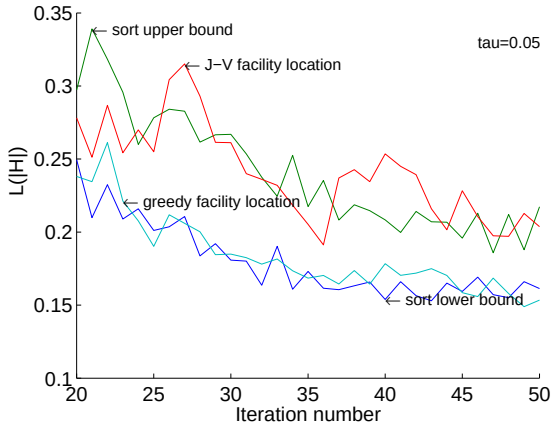
τ=0.2, K=25	Total time	Propagation	SDP	Pruning
sort lower	1.04h	1.85%	98.15%	0.00%
sort upper	1.34h	1.52%	98.43%	0.05%
J-V p-d	1.38h	1.45%	89.47%	9.08%
greedy	1.43h	1.63%	97.84%	0.53%

6.5.4 Discussion

Our experimental results confirm that the total approximation error comes both from the approximation error of the Lax-Oleinik semi-group and from the pruning error. The error of approximation of the semi-group can be improved by decreasing the discretization-time step-size τ, while the pruning error depends on the pruning techniques and the number of basis functions kept at each iteration. We introduced here refined pruning techniques, still SDP based and combining facility location algorithms and semidefinite relaxations, which improve the final precision (see Figure 6.3). However, these pruning techniques remain time-consuming (see Table 6.1), in particular, when τ becomes small, the pruning appears to be the bottleneck. Therefore, new ideas are needed to develop more efficient methods. Besides, our experiments also show that the error appears smaller than the bound of $O(\sqrt{\tau})$ established in [MK10]. In next chapter, we show that under an additional assumption, the error bound is indeed of order $O(\tau)$.



(a) $\tau = 0.1$



(b) $\tau = 0.05$

Figure 6.3: Comparison of the four pruning techniques by the evolution of the discrete L_1 norm of backsubstitution error on the rectangle $[-2,2] \times [-2,2]$ of the $x_1 - x_2$ plane, with respect to the number of iterations.

CHAPTER 7

An improved convergence analysis of the max-plus curse of dimensionality free method

In the previous chapter, we reviewed McEneaney's curse of dimensionality free method, which applies to the Hamilton Jacobi equations where the Hamiltonian takes the form of a (pointwise) maximum of affine/quadratic functions. In this chapter, we focus on the convergence analysis of the method, restricted to the case when the Hamiltonian is the pointwise maximum of pure quadratic forms (without affine terms). In previous works of McEneaney and Kluberg, the approximation error of the method was shown to be $O(1/(N\tau))+O(\sqrt{\tau})$ where τ is the time discretization step and N is the number of iterations. Here we use the contraction result for the indefinite Riccati flow in Thompson's part metric, established in Chapter 2, to show that under different technical assumptions, still covering an important class of problems, the error is only of order $O(e^{-\alpha N\tau})+O(\tau)$ for some $\alpha > 0$. Besides, our approach allows to incorporate the pruning error in the analysis and we show that if the pruning error is $O(\tau^2)$, than the same approximation error order holds. This allows us to tune the precision of the pruning procedure, which in practice is a critical element of the method.

This chapter is based on the preprint [Qu13a]. An abridged version of this chapter is included in the ECC conference proceeding [Qu13b].

7.1 Introduction

We study the error bound of McEneaney's curse of dimensionality method presented in the previous chapter. We consider the *pure quadratic* case, i.e., the Hamiltonian Jacobi equation takes the form:

$$0 = -H(x, \nabla V) = -\max_{m \in \mathscr{M}} \{H^m(x, \nabla V)\} \tag{7.1}$$

where $\mathscr{M} = \{1, 2, \cdots, M\}$ and

$$H^m(x,p) = (A^m x)'p + \frac{1}{2} x' D^m x + \frac{1}{2} p' \Sigma^m p \ . \tag{7.2}$$

This Hamilton-Jacobi equation corresponds to a linear quadratic switched optimal control problem (Section 7.2.1) where the control switches between several linear quadratic systems. The solution V of equation (7.1) is the value function of the corresponding infinite horizon switched optimal control problem. The method consists of two successive approximations (see Section 6.3). First we approximate the infinite horizon problem by a finite horizon problem. Then we approximate the value function of the finite horizon switched optimal control problem by choosing an optimal strategy which does not switch on small intervals.

We denote by $(S_t)_{t \geqslant 0}$ and $(S_t^m)_{t \geqslant 0}$ for all $m \in \mathscr{M}$ respectively the semigroup corresponding to H and H^m for all $m \in \mathscr{M}$. Let V_0 be a given initial function and $T > 0$ be the finite horizon. The first approximation uses $S_T[V_0]$ to approximate V and introduces the finite-horizon truncation error at point $x \in \mathbb{R}^d$:

$$\varepsilon_0(x, T, V_0) := V(x) - S_T[V_0](x).$$

Let $\tau > 0$ be a small time step and $N > 0$ such that $T = N\tau$. Denote by $\tilde{S}_\tau$ the semigroup of the optimal control problem where the control does not switch on the interval $[0, \tau]$. The second approximation approximates $S_T[V_0]$ by $\{\tilde{S}_\tau\}^N[V_0]$ where

$$\tilde{S}_\tau = \sup_{m \in \mathscr{M}} S_\tau^m.$$

The error at point x of this time discretization approximation is denoted by:

$$\varepsilon(x, \tau, N, V_0) := S_T[V_0](x) - \{\tilde{S}_\tau\}^N[V_0](x).$$

The total error at a point x is then simply $\varepsilon_0(x, T, V_0) + \varepsilon(x, \tau, N, V_0)$. The computational cost is $O(M^N d^3)$, with a cubic growth in the state dimension d. In this sense it is considered as a curse of dimensionality free method. However, we see that the computational cost is bounded by a number exponential in the number of iterations, which is referred to as the curse of complexity. In practice, a pruning procedure denoted by $\mathscr{P}_\tau$ removing at each iteration a number of functions less useful than others is needed in order to reduce the curse of complexity. We denote the error at point x of the time dicretization approximation incorporating the pruning procedure by:

$$\varepsilon^{\mathscr{P}_\tau}(x, \tau, N, V_0) = S_T[V_0](x) - \{\mathscr{P}_\tau \circ \tilde{S}_\tau\}^N[V_0](x).$$

7.1.1 Main contributions

In this chapter, we analyze the growth rate of $\varepsilon_0(x,T,V_0)$ as T tends to infinity and the growth rate of $\varepsilon^{\mathscr{P}_\tau}(x,\tau,N,V_0)$ as τ tends to 0, incorporating a pruning procedure $\mathscr{P}_\tau$ of error $O(\tau^r)$ with $r>1$. The error $\varepsilon(x,\tau,N,V_0)$ in the absence of pruning can be obtained by taking $r=+\infty$.

We show that under technical assumptions (Assumption 7.1 and 7.2),

$$\sup_{x\neq 0}\varepsilon_0(x,T,V_0)/|x|^2 = O(e^{-\alpha T}), \quad \text{as } T\to+\infty$$

uniformly for all initial quadratic functions $V_0(x)=\frac{1}{2}x'Px$ where P is a matrix in a certain compact (Theorem 7.4). We also show that given a pruning procedure generating an error $O(\tau^r)$ with $r>1$,

$$\sup_{x\neq 0}\varepsilon^{\mathscr{P}_\tau}(x,\tau,N,V_0)/|x|^2 = O(\tau^{\min\{1,r-1\}}), \quad \text{as } \tau\to 0$$

uniformly for all $N\in\mathbb{N}$ and V_0 as above (Theorem 7.5). As a direct corollary, we have

$$\sup_{x\neq 0}\varepsilon(x,\tau,N,V_0)/|x|^2 = O(\tau), \quad \text{as } \tau\to 0$$

uniformly for all $N\in\mathbb{N}$ and V_0 as above.

7.1.2 Comparison with earlier estimates

McEneaney and Kluberg showed in [MK10, Thm 7.1] that under Assumption 7.1, for a given V_0,

$$\sup_{x}\varepsilon_0(x,T,V_0)/(1+|x|^2) = O(\frac{1}{T}), \quad \text{as } T\to+\infty \tag{7.3}$$

They also showed [MK10, Thm 6.1] that if in addition to Assumption 7.1, the matrices Σ^m are all identical for $m\in\mathscr{M}$, then for a given V_0,

$$\sup_{x}\varepsilon(x,\tau,N,V_0)/(1+|x|^2) = O(\sqrt{\tau}), \quad \text{as } \tau\to 0 \tag{7.4}$$

uniformly for all $N\in\mathbb{N}$. Their estimates imply that to get a sufficiently small approximation error ε we can use a horizon $T=O(1/\varepsilon)$ and a discretization step $\tau=O(\varepsilon^2)$. Thus asymptotically the computational cost is:

$$O(M^{O(1/\varepsilon^3)}d^3), \quad \text{as } \varepsilon\to 0.$$

The same reasoning applied to our estimates shows a considerably smaller asymptotic growth rate of the computational cost (Corollary 7.11):

$$O(M^{O(-\log(\varepsilon)/\varepsilon)}d^3), \quad \text{as } \varepsilon\to 0$$

McEneaney and Kluberg [MK10] gave a technically difficult proof of the estimates (7.3) and (7.4), assuming that all the Σ^m's are the same. They conjectured that the latter assumption can at least be released for a subclass of problems. This is supported by our results, showing that for the subclass of problems satisfying Assumption 7.2, this assumption can be omitted. To this end, we use a totally different approach. Our main idea is to use Thompson's part metric to measure the error. In Chapter 2, we showed (Corollary 2.13) that the indefinite Riccati flows has a strict local contraction property in Thompson's part metric under some technical assumptions. This local contraction result

on the indefinite Riccati flow constitutes an essential part of our proofs. Indeed Assumption 7.2 is made to guarantee the strict local contraction property of the indefinite Riccati flows. We shall see in Corollary 7.8 that when all the cost functions are the same, then Assumption 7.2 can be dispensed with.

Our approach derives a tighter estimate of $\varepsilon_0(x,T,V_0)$ and $\varepsilon(x,\tau,N,V_0)$ compared to previous results as well as an estimate of $\varepsilon^{\mathscr{P}_\tau}(x,\tau,N,V_0)$ incorporating the pruning procedure. This new result justifies the use of pruning procedure of error $O(\tau^2)$ without increasing the asymptotic total approximation error order.

The chapter is organized as follows. In Section 7.2, we recall the switched linear quadratic control problem and the max-plus approximation method. In Section 7.3, we recall the contraction results on the indefinite Riccati flow as well as the extension of Thompson's part metric to the space of supremum of quadratic functions. In Sections 7.4 and 7.5, we present the estimates of the two approximation errors and part of the proofs. In Section 7.7, we show the proofs of some technical lemmas. Finally in Section 7.8, we give some remarks and some numerical illustrations of the theoretical estimates.

7.2 Problem statement

We recall briefly the problem class and present some basic concepts and necessary assumptions. The reader can find more details in [McE07].

7.2.1 Problem class

We consider Problem 6.1 without affine terms, i.e., the following infinite horizon switched optimal control problem:

Problem 7.1.

$$V(x) = \sup_{\mathbf{u}\in W}\sup_{\mu\in\mathscr{D}_\infty}\sup_{T<\infty} J(x,T;\mathbf{u},\mu)$$

where

$$\begin{aligned} J(x,T;\mathbf{u},\mu) &= \int_0^T \frac{1}{2}\mathbf{x}(s)' D^{\mu(s)}\mathbf{x}(s) - \frac{\gamma^2}{2}|\mathbf{u}(s)|^2\, ds \ , \\ \mathscr{D}_\infty &= \{\mu : [0,\infty) \to \mathscr{M} : \text{measurable}\} \ , \\ W &\doteq L_2^{\text{loc}}([0,\infty);\mathbb{R}^k) \ , \end{aligned}$$

and the state dynamics are given by

$$\dot{\mathbf{x}}(s) = A^{\mu(s)}\mathbf{x}(s) + \sigma^{\mu(s)}\mathbf{u}(s);\ \ \mathbf{x}(0) = x \in \mathbb{R}^d \ . \tag{7.5}$$

As in Chapter 6, we denote by $(S_t)_t$ the evolution semigroup associated to Problem 7.1. That is, for a function $\phi : \mathbb{R}^d \to \mathbb{R}$ and $x \in \mathbb{R}^d$, $S_t[\phi](x)$ is the value function at point x of the finite horizon optimal control problem with terminal reward ϕ:

Problem 7.2.

$$S_t[\phi](x) = \sup_{\mathbf{u}\in W_t}\sup_{\mu\in\mathscr{D}_t} J(x,t;\mathbf{u},\mu) + \phi(\mathbf{x}(t))$$

where W_t and $\mathscr{D}_t$ are defined in (6.4) *and $\mathbf{x}(\cdot) : [0,t] \to \mathbb{R}^d$ satisfies*

$$\dot{\mathbf{x}}(s) = A^{\mu(s)}\mathbf{x}(s) + \sigma^{\mu(s)}\mathbf{u}(s),\ \ s \in [0,t];\ \ \mathbf{x}(0) = x \ .$$

For $m \in \mathscr{M}$, define the semigroup $(S_t^m)_t$ associated to the linear quadratic problem indexed by m. More specifically, for a function $\phi : \mathbb{R}^d \to \mathbb{R}$ and $x \in \mathbb{R}^d$, we have:

$$S_t^m[\phi](x) := \sup_{\mathbf{u} \in W_t} J(x,t;\mathbf{u},m) + \phi(\mathbf{x}(t))$$

where $\mathbf{x}(\cdot) : [0,t] \to \mathbb{R}^d$ satisfies

$$\dot{\mathbf{x}}(s) = A^m \mathbf{x}(s) + \sigma^m \mathbf{u}(s), \;\; s \in [0,t]; \;\; \mathbf{x}(0) = x \;\; .$$

As in Chapter 6, for $m \in \mathscr{M}$, denote

$$\Sigma^m := \frac{1}{\gamma^2} \sigma^m (\sigma^m)' \;\; ,$$

and

$$\Phi^m(P) = (A^m)'P + PA^m + P\Sigma^m P + D^m.$$

For each $m \in \mathscr{M}$, denote by $M_\cdot^m(\cdot)$ the flow associated to the Riccati equation

$$\dot{P} = \Phi^m(P).$$

Then it is standard that (see [YZ99])

$$\phi(x) = \frac{1}{2} x'Px, \;\; \forall x \in \mathbb{R}^d \;\; \Rightarrow S_t^m[\phi](x) = \frac{1}{2} x' M_t^m(P) x, \;\; \forall m \in \mathscr{M}, x \in \mathbb{R}^d \;\; . \tag{7.6}$$

The corresponding Hamiltonian can be written as:

$$H(x,p) = \max_{m \in \mathscr{M}} H^m(x,p)$$

where for each m, H^m takes the form

$$H^m(x,p) = \frac{1}{2} x' D^m x + \frac{1}{2} p' \Sigma^m p + (A^m x)' p. \tag{7.7}$$

We made some assumptions on the matrix parameters in Chapter 6, in order to guarantee the existence of the value function of the infinite horizon problem. Here the assumptions are somewhat different since we restrict to the case without affine terms. As in [McE07], we make the following assumptions to Problem 7.1.

Assumption 7.1.

- There exists $c_A > 0$ such that:

$$x'A^m x \leqslant -c_A |x|^2, \quad \forall x \in \mathbb{R}^d, m \in \mathscr{M}$$

- There exists $c_\sigma > 0$ such that:

$$|\sigma^m| \leqslant c_\sigma, \quad \forall m \in \mathscr{M}$$

- All D^m are positive definite, symmetric, and there is c_D such that:

$$x'D^m x \leqslant c_D |x|^2, \quad \forall x \in \mathbb{R}^d, m \in \mathscr{M},$$

and

$$c_A^2 > \frac{c_D c_\sigma^2}{\gamma^2}$$

7.2.2 Steady HJ equation

For any $\delta \in (0,\gamma)$, define

$$G_\delta := \{V \text{ semiconvex } : V(x) \leqslant \frac{c_A(\gamma-\delta)^2}{c_\sigma^2}|x|^2, \ \forall x\}. \tag{7.8}$$

Then the value function V is the unique viscosity solution of the following corresponding HJ PDE in the class G_δ for sufficiently small δ [McE07]:

$$0 = -H(x,\nabla V) = -\max_{m\in\mathscr{M}} H^m(x,\nabla V). \tag{7.9}$$

where H^m is defined in (7.7). It was shown in [McE07] that for δ sufficiently small and $V_0 \in G_\delta$,

$$\lim_{T\to\infty} S_T[V_0] = V \tag{7.10}$$

uniformly on compact sets.

7.2.3 Max-plus based approximation errors

We refer the reader to Section 6.3.3 the basic steps of the curse of dimensionality algorithm proposed in [McE07] to approximate the value function V. As pointed out in [MK10], the approximation error comes from two parts. The first error source is due to the approximation of the infinite horizon problem by a finite horizon problem. At a point $x \in \mathbb{R}^d$, this error is denoted by:

$$\varepsilon_0(x,T,V_0) := V(x) - S_T[V_0](x) \ . \tag{7.11}$$

The second source of error is caused by the approximation of the semigroup by a time-discretization. At a point $x \in \mathbb{R}^d$, the latter error is denoted by:

$$\varepsilon(x,\tau,N,V_0) := S_{N\tau}[V_0](x) - \{\tilde{S}_\tau\}^N[V_0](x),$$

Recall that

$$\tilde{S}_\tau = \sup_{m\in\mathscr{M}} S_\tau^m.$$

In Section 6.3.3 we mentioned that in practice a pruning procedure is needed so as to reduce the number of quadratic functions. If we take into account the pruning procedure, then the second error source should be written as:

$$\varepsilon^{\mathscr{P}_\tau}(x,\tau,N,V_0) = S_{N\tau}[V_0](x) - \{\mathscr{P}_\tau \circ \tilde{S}_\tau\}^N[V_0](x), \tag{7.12}$$

where $\mathscr{P}_\tau$ represents a given pruning rule. We mark the subscript τ since it is expected that the pruning procedure be adapted with the time step τ.

7.3 Contraction properties of the indefinite Riccati flow

As already noted in Section 7.1, the essential ingredient of our proof is the local contraction property of the indefinite Riccati flow, presented in Chapter 2. Below is the additional assumption needed to apply this new contraction result:

Assumption 7.2. There is $m_D > 0$ such that

$$x' D^m x \geqslant m_D |x|^2, \quad \forall x \in \mathbb{R}^n, m \in \mathscr{M} \ .$$

Besides,

$$\lambda_1 < \lambda_2$$

where

$$\lambda_1 = \frac{\gamma^2 (c_A - \sqrt{c_A^2 - c_D c_\sigma^2/\gamma^2})}{c_\sigma^2} \ , \tag{7.13}$$

and

$$\lambda_2 := \sqrt{m_D \gamma^2 / c_\sigma^2} \ . \tag{7.14}$$

Remark 7.1. Note that when $m_D = c_D$, Assumption 7.2 is automatically satisfied if Assumption 7.1 holds. To see this, it suffices to remark that for all $a > b > 0$ such that $b^2 > a$ we have:

$$b - \sqrt{b^2 - a} < \sqrt{a} \ .$$

The condition $m_D = c_D$ implies exactly that all the matrices D^m equal to a multiple of identity matrix:

$$D^m = c_D I, \ \ \forall m \in \mathscr{M} \ .$$

The main ingredient to make our proofs is the following theorem, which is direct from Corollary 2.13.

Theorem 7.3. *Under Assumptions 7.1 and 7.2, for any $\lambda \in (0, \lambda_2)$, there is $\alpha > 0$ such that for all $P_1, P_2 \in (0, \lambda I]$,*

$$d_T(M_t^m(P_1), M_t^m(P_2)) \leqslant e^{-\alpha t} d_T(P_1, P_2), \ \ \forall t \geqslant 0, m \in \mathscr{M}.$$

Remark 7.2. Assumption 7.2 is better understood by considering a special case of Problem 6.1 with dimension equal to 1 and switching number equal to 1, satisfying Assumption 7.1. The coefficients D^1, Σ^1 and A^1 can now be replaced respectively by the scalars c_D, c_σ^2/γ^2 and $-c_A$. The Riccati differential equation associated to this special case is then:

$$\dot{p} = \Phi^1(p) \tag{7.15}$$

where

$$\Phi^1(p) = \frac{c_\sigma^2}{\gamma^2} p^2 - 2 c_A p + c_D, \ \ \forall p \in \mathbb{R} \ .$$

Theorem 7.3 can be interpreted as follows: the scalar Riccati flow associated to (7.15) is a strict contraction in $(0, \lambda_2)$, uniformly on compact sets, where λ_2 is defined in (7.14). Besides, it turns out that the scalar value λ_1 defined in (7.13) is the stable equilibrium point of the Riccati differential equation (7.15). Thus Assumption 7.2 requires that the stable equilibrium point of (7.15) be less than the right end point of the contraction interval of the flow associated to (7.15). This is automatically satisfied in this special case without switching because we can take $m_D = c_D$ (see Remark 7.1). In Figure 7.1, we show the plot of the function $\Phi^1(p) = (p-5)(p-1)$. The stable equilibrium point $\lambda_1 = 1$ is marked in blue and the contraction limit point $\lambda_2 = \sqrt{5}$ is marked in red.

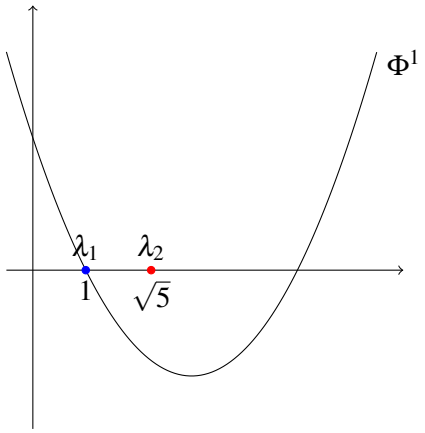


Figure 7.1: A scalar Riccati dynamic function. The stable equilibrium point is in blue (λ_1). The right end point of the contraction interval is in red (λ_2).

However, still in dimension 1, if the switching control number $|\mathscr{M}|$ is greater than 1, than λ_1 gives an upper bound on the stable equilibrium points and λ_2 gives a lower bound on the end points of contraction intervals. Assumption 7.2 is made to guarantee that the maximal stable equilibrium point is less than the minimal end point of all the contraction intervals. In Figure 7.2(a), we show an example of three Riccati dynamic functions such that all the stable equilibrium points (in blue) are less than the end points (in red) of contraction intervals. In Figure 7.2(b), we give an example not satisfying Assumption 7.2.

Remark 7.3. Under Assumption 7.2, we can choose $\varepsilon_1 > 0$ sufficiently small so that

$$M^m_{t_0}(0) \succcurlyeq \varepsilon_1 I, \quad \text{for some } t_0 > 0, m \in \mathscr{M}. \tag{7.16}$$

Since $\Phi^m(0) = D^m \succcurlyeq m_D I_d$ for all $m \in \mathscr{M}$, we can let ε be sufficiently small such that $\Phi^m(\varepsilon_1 I) \succcurlyeq 0$ for all $m \in \mathscr{M}$. Besides, for any $\lambda \in [\lambda_1, \lambda_2)$, we have $\Phi^m(\lambda I) \preccurlyeq 0$ for all $m \in \mathscr{M}$. Then it follows from Lemma 2.11 that:

$$M^m_t(P_0) \in [\varepsilon_1 I, \lambda I], \ \ \forall m \in \mathscr{M}, t \geqslant 0, P_0 \in [\varepsilon_1 I, \lambda I]. \tag{7.17}$$

7.3.1 Extension of the contraction result to the space of functions

Now we extend the definition of Thompson's part metric to the space of non-negative functions. For two functions $f, g : \mathbb{R}^n \to \mathbb{R}$, we consider the standard partial order "$\leqslant$" by:

$$f \leqslant g \Leftrightarrow f(x) \leqslant g(x), \ \ \forall x \in \mathbb{R}^n,$$

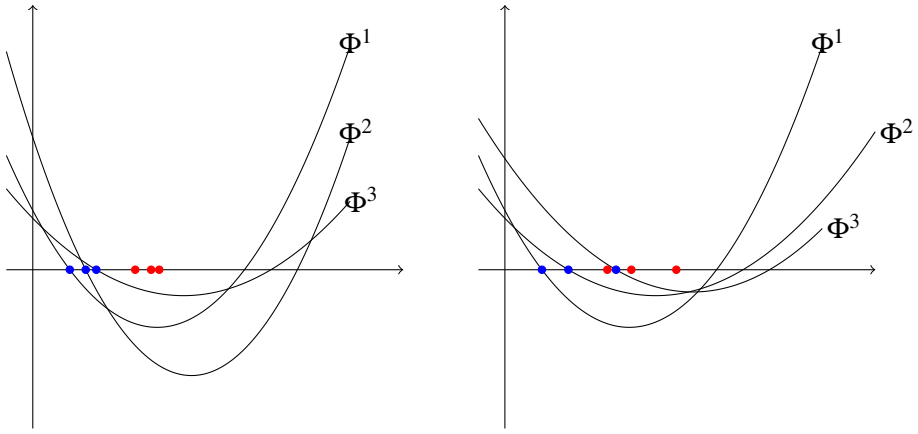
which coincides with the Loewner order on the set of quadratic forms. Similarly, for $f, g : \mathbb{R}^d \to \mathbb{R}_+$ we define

$$M(f/g) := \inf\{t > 0 : f \leqslant tg\}$$

We say that f and g are comparable if $M(f/g)$ and $M(g/f)$ are finite. In that case, we can define the *"Thompson metric"* between $f, g : \mathbb{R}^d \to \mathbb{R}_+$ by:

$$d_T(f,g) = \log(\max\{M(f/g), M(g/f)\}). \tag{7.18}$$

Then the following lemma can be easily proved using the definition:



(a) (b)

Figure 7.2: Plot of the scalar Riccati dynamic functions associated to an one dimensional instance satisfying Assumption 7.2 (left). Variant in which Assumption 7.2 is not satisfied (right).

Lemma 7.1. *Let $f, g : \mathbb{R}^d \to \mathbb{R}_+$ be given by pointwise maxima of non-negative functions*

$$f := \sup_{i\in I} f_i, \quad g := \sup_{i\in I} g_i$$

Then

$$d_T(f,g) \leqslant \sup_{i\in I} d_T(f_i, g_i). \tag{7.19}$$

The following result is a consequence of the order-preserving character of the Riccati flow and of the contraction property in Theorem 7.3.

Lemma 7.2. *Under Assumptions 7.1 and 7.2, let $\lambda \in [\lambda_1, \lambda_2)$ and $\varepsilon_1 > 0$ such that* (7.17) *holds. Then there is $\alpha > 0$ such that for any two functions V_1 and V_2 of the form:*

$$V_1(x) = \sup_{j\in J} \frac{1}{2} x' P_j x, \qquad V_2(x) = \frac{1}{2} x' Q x,$$

where $Q, P_j \in [\varepsilon_1 I, \lambda I]$ for all $j \in J$, we have

$$d_T(S^{i_N}_{t/N} \cdots S^{i_1}_{t/N}[V_1], S^{i_N}_{t/N} \cdots S^{i_1}_{t/N}[V_2]) \leqslant e^{-\alpha t} \log(\frac{\lambda}{\varepsilon})$$

for all $t \geqslant 0$, $N \in \mathbb{N}$ and $(i_1, \cdots, i_N) \in \mathscr{M}^N$.

Proof. For all $P, Q \in [\varepsilon_1 I, \lambda I]$, by (7.17) and Theorem 7.3 we have

$$d_T(M^{i_N}_{t/N} \cdots M^{i_1}_{t/N}(P), M^{i_N}_{t/N} \cdots M^{i_1}_{t/N}(Q)) \leqslant e^{-\alpha t} d_T(P, Q)$$

for all $t \geqslant 0$, $N \in \mathbb{N}$ and $(i_1, \cdots, i_N) \in \mathscr{M}^N$. Now by the max-plus linearity of the semigroup, Lemma 7.1 and the relationship between the semigroup and the flow (7.6), we get

$$\begin{aligned} & d_T(S^{i_N}_{t/N} \cdots S^{i_1}_{t/N}[V_1], S^{i_N}_{t/N} \cdots S^{i_1}_{t/N}[V_2]) \\ & \leqslant \sup_{j \in J} d_T(M^{i_N}_{t/N} \cdots M^{i_1}_{t/N}(P_j), M^{i_N}_{t/N} \cdots M^{i_1}_{t/N}(Q)) \\ & \leqslant e^{-\alpha t} \sup_{j \in J} d_T(P_j, Q) \leqslant e^{-\alpha t} \log(\frac{\lambda}{\varepsilon}). \end{aligned}$$

□

7.4 Finite horizon error estimate

We first study the finite horizon truncation error $\varepsilon_0(x, T, V_0)$ in (7.11). Below is one of our main results:

Theorem 7.4. *Under Assumptions 7.1 and 7.2, let $\lambda \in [\lambda_1, \lambda_2)$ and $\varepsilon_1 > 0$ such that* (7.16) *and* (7.17) *hold. There exist $\alpha > 0$ and $K > 0$ such that,*

$$\varepsilon_0(x, T, V_0) \leqslant K e^{-\alpha T} |x|^2, \ \ \forall x,$$

for all $T > 0$ and $V_0(x) = \frac{1}{2} x' P_0 x$ with $P_0 \in [\varepsilon_1 I, \lambda I]$.

The remaining part of the section is devoted to the proof of the above theorem. We shall need the following technical lemma. The proof is deferred to Section 7.7.1.

Lemma 7.3 (Approximation by piecewise constant controls)**.** *Let $V_0 : \mathbb{R}^d \to \mathbb{R}$ be a given locally Lipschitz function. For any $T > 0$ we have*

$$S_T[V_0] = \sup_N \sup_{i_1, \cdots i_N} S^{i_N}_{T/N} \cdots S^{i_1}_{T/N}[V_0].$$

From now on we make Assumptions 7.1 and 7.2. We also fix $\lambda \in [\lambda_1, \lambda_2)$ and $\varepsilon_1 > 0$ satisfying (7.16) and (7.17).

Remark 7.4. Since the interval $[\varepsilon_1 I, \lambda I]$ is invariant by any operator $\{S^m_\tau\}_{\tau \geqslant 0, m \in \mathscr{M}}$, it is direct from Lemma 7.3 that

$$\frac{\varepsilon_1}{2} |x|^2 \leqslant S_T[V_0](x) \leqslant \frac{\lambda}{2} |x|^2, \ \ \forall T > 0 \tag{7.20}$$

for all $V_0(x) = \frac{1}{2} x' P x$ with $P \in [\varepsilon_1 I, \lambda I]$.

Corollary 7.4. *The value function V is a pointwise supremum of quadratic functions*

$$V(x) = \sup_{j \in J} \frac{1}{2} x' P_j x$$

where $P_j \in [\varepsilon_1 I, \lambda I]$ for all $j \in J$.

Proof. By definition, we have:

$$V(x) = \sup_{T>0} S_T[0](x), \ \ \forall x.$$

By (7.16), there is $t_0 > 0$ and $m \in \mathscr{M}$ such that

$$M^m_{t_0}(0) \geqslant \varepsilon_1 I.$$

Besides, by the monotonicity of the semigroup,

$$S_T[S^m_{t_0}[0]](x) \leqslant S_T[S_{t_0}[0]](x), \ \ \forall x, T > 0$$

and

$$S_T[0](x) \leqslant S_T[S^m_{t_0}[0]](x), \ \ \forall x, T > 0.$$

Since

$$V(x) = \sup_{T>0} S_T[0](x) = \sup_T S_T[S_{t_0}[0]](x), \ \ \forall x \ \ ,$$

we get that:

$$\sup_T S_T[S^m_{t_0}[0]](x) \leqslant V(x) \leqslant \sup_T S_T[S^m_{t_0}[0]](x), \ \ \forall x \ \ .$$

Hence by Lemma 7.3:

$$V = \sup_T S_T[S^m_{t_0}[0]] = \sup_T \sup_N \sup_{i_1,\cdots i_N} S^{i_N}_{T/N} \cdots S^{i_1}_{T/N} S^m_{t_0}[0]. \tag{7.21}$$

Now using the invariance of the interval $[\varepsilon_1 I, \lambda I]$ in (7.17), we know that

$$M^{i_N}_{T/N} \cdots M^{i_1}_{T/N} M^m_{t_0}(0) \in [\varepsilon_1 I, \lambda I],$$

for all $T > 0$, $N \in \mathbb{N}$ and $i_1, \ldots, i_N \in \mathscr{M}$. Consequently V is a pointwise maximum of quadratic functions $\frac{1}{2} x' P_j x$ with $P_j \in [\varepsilon_1 I, \lambda I]$. □

Using the above lemma we show that:

Proposition 7.5. *There is $\alpha > 0$ such that for all $V_0(x) = \frac{1}{2} x' P_0 x$ with $P_0 \in [\varepsilon_1 I, \lambda I]$,*

$$d_T(V, S_T[V_0]) \leqslant e^{-\alpha T} \log(\frac{\lambda}{\varepsilon}), \ \ \forall T > 0.$$

Proof. By Corollary 7.4, the value function V is a pointwise supremum of quadratic functions:

$$V(x) = \sup_{j \in J} \frac{1}{2} x' P_j x$$

where $P_j \in [\varepsilon_1 I, \lambda I]$ for all $j \in J$. Let any $V_0(x) = \frac{1}{2} x' P_0 x$ with $P_0 \in [\varepsilon_1 I, \lambda I]$. By Corollary 7.2, we have:

$$d_T(S^{i_N}_{T/N} \cdots S^{i_1}_{T/N}[V], S^{i_N}_{T/N} \cdots S^{i_1}_{T/N}[V_0]) \leqslant e^{-\alpha T} \log(\frac{\lambda}{\varepsilon})$$

for all $T \geqslant 0$, $N \in \mathbb{N}$ and $(i_1, \cdots, i_N) \in \mathscr{M}^N$. We also know from Lemma 7.3 that

$$V = S_T[V] = \sup_N \sup_{i_1,\ldots,i_N} S^{i_N}_{T/N} \ldots S^{i_1}_{T/N}[V],$$

and that

$$S_T[V_0] = \sup_N \sup_{i_1,\dots,i_N} S^{i_N}_{T/N} \dots S^{i_1}_{T/N}[V_0].$$

Therefore by Lemma 7.1,

$$\begin{aligned} & d_T(V, S_T[V_0]) = d_T(S_T[V], S_T[V_0]) \\ & \leqslant \sup_N \sup_{i_1,\cdots i_N} d_T(S^{i_N}_{T/N} \cdots S^{i_1}_{T/N}[V], S^{i_N}_{T/N} \cdots S^{i_1}_{T/N}[V_0]) \\ & \leqslant e^{-\alpha T} \log(\tfrac{\lambda}{\varepsilon}). \end{aligned}$$

□

Now we have all the necessary elements to prove Theorem 7.4.

Proof of Theorem 7.4. Let any $V_0(x) = \frac{1}{2} x' P_0 x$ with $P_0 \in [\varepsilon_1 I, \lambda I]$. By Proposition 7.5 and (7.18), there is $\alpha > 0$ such that

$$V(x) \leqslant e^{e^{-\alpha T} \log(\lambda/\varepsilon)} S_T[V_0](x), \ \ \forall T > 0, x \in \mathbb{R}^d.$$

Thus there is constant $L > 0$ such that

$$V(x) \leqslant (1 + L e^{-\alpha T}) S_T[V_0](x), \ \ \forall T > 0, x \in \mathbb{R}^d$$

This leads to

$$\varepsilon_0(x, T, V_0) \leqslant L e^{-\alpha T} S_T[V_0](x) \leqslant \frac{\lambda L}{2} e^{-\alpha T} |x|^2, \ \ \forall T > 0, x \in \mathbb{R}^d.$$

where the last inequality follows from (7.20). It is clear that the constant $K = \frac{\lambda L}{2}$ is independent of $P_0 \in [\varepsilon_1 I, \lambda I]$. □

7.5 Discrete-time approximation error estimate

In this section we analyze the discrete-time approximation error $\varepsilon^{\mathscr{P}_\tau}(x, \tau, N, V_0)$. We say that $\mathscr{P}_\tau$ is a pruning procedure generating an error $O(\tau^r)$ if there is $L > 0$ such that for all function f of the form (6.13),

$$\mathscr{P}_\tau[f] \leqslant f \leqslant (1 + L\tau^r) \mathscr{P}_\tau[f]. \tag{7.22}$$

The special case without pruning procedure can be recovered by considering $r = +\infty$. Our main result is:

Theorem 7.5. *Let $r > 1$. Suppose that for each $\tau > 0$ the pruning operation $\mathscr{P}_\tau$ generates an error $O(\tau^r)$ (see (7.22)). Under Assumptions 7.1 and 7.2, let $\lambda \in [\lambda_1, \lambda_2)$ and $\varepsilon_1 > 0$ such that (7.17) holds. Then there exist $\tau_0 > 0$ and $L > 0$ such that*

$$\varepsilon^{\mathscr{P}_\tau}(x, \tau, N, V_0) \leqslant L \tau^{\min\{1, r-1\}} |x|^2, \ \ \forall x,$$

for all $N \in \mathbb{N}$, $\tau \leqslant \tau_0$ and $V_0(x) = \frac{1}{2} x' P_0 x$ with $P_0 \in [\varepsilon_1 I, \lambda I]$.

The remaining part of the section is devoted to the proof of Theorem 7.5. We first state a technical lemma which is proved in Section 7.7.2.

Lemma 7.6. *Let $\mathscr{K} \subset \mathrm{S}_d$ be a compact convex subset. There exist $\tau_0 > 0$ and $L > 0$ such that*

$$S_\tau[V_0](x) \leqslant \tilde{S}_\tau[V_0](x) + L\tau^2|x|^2, \ \ \forall x,$$

for all $\tau \in [0, \tau_0]$ and $V_0(x) = \frac{1}{2}x'P_0x$ with $P_0 \in \mathscr{K}$.

Now we take into account the pruning procedure and analyze the error of the following approximation

$$S_\tau \simeq \mathscr{P}_\tau \circ \tilde{S}_\tau.$$

Below is a direct consequence of Lemma 7.6 and (7.17).

Corollary 7.7. *Let ε, λ, r and $\mathscr{P}_\tau$ be as in Theorem 7.5. Then there exist $\tau_0 > 0$ and $L > 0$ such that:*

$$S_\tau[V_0](x) \leqslant (1 + L\tau^{\min\{2,r\}})\mathscr{P}_\tau \circ \tilde{S}_\tau[V_0](x), \ \ \forall x,$$

for all $\tau \in [0, \tau_0]$ and $V_0(x) = \frac{1}{2}x'P_0x$ with $P_0 \in [\varepsilon_1 I, \lambda I]$.

We are ready to give a proof of Theorem 7.5:

Proof of Theorem 7.5. Denote $s = \min\{2, r\}$. Let any $\lambda' > 0$ such that

$$\lambda < \lambda' < \lambda_2.$$

Denote $\delta = \lambda'/\lambda$. Consider the two compact convex subsets $\mathscr{K}_0 = [\varepsilon_1 I, \lambda I]$ and $\mathscr{K}_1 = [\varepsilon_1 I, \lambda' I]$. It is easily verified that:

$$\Phi^m(\lambda' I) \leqslant 0, \ \ \forall m \in \mathscr{M}.$$

Therefore for all $P_0 \in \mathscr{K}_0$, $P_1 \in \mathscr{K}_1$, $t \geqslant 0$ and $m \in \mathscr{M}$,

$$M_t^m(P_0) \in \mathscr{K}_0, \quad M_t^m(P_1) \in \mathscr{K}_1. \tag{7.23}$$

By Corollary 7.7, there is τ_0 and $L > 0$ such that for all $\tau \in [0, \tau_0]$ and $V_0 = \frac{1}{2}x'Px$ with $P \in \mathscr{K}_1$:

$$S_\tau[V_0] \leqslant (1 + L\tau^s)\mathscr{P}_\tau \circ \tilde{S}_\tau[V_0]. \tag{7.24}$$

Let $\tau_0 > 0$ be sufficiently small such that:

$$(1 + L\tau^s)^{\frac{1}{1-e^{-\alpha\tau}}} \leqslant \delta, \quad \forall \tau \in [0, \tau_0].$$

Let any $V_0(x) = \frac{1}{2}x'P_0x$ with $P_0 \in \mathscr{K}_0$ and $\tau \in [0, \tau_0]$, we are going to prove by induction on $N \in \mathbb{N}$ the following inequalities:

$$S_{N\tau}[V_0] \leqslant (1 + L\tau^s)^{1+e^{-\alpha\tau}+\cdots+e^{-(N-1)\alpha\tau}}\{\mathscr{P}_\tau \circ \tilde{S}_\tau\}^N[V_0], \quad \forall N \in \mathbb{N}.$$

The case $N = 1$ is already given in (7.24). Suppose that the above inequality is true for some $k \in \mathbb{N}$, that is,

$$S_{k\tau}[V_0] \leqslant L_k\{\mathscr{P}_\tau \circ \sup_m S_\tau^m\}^k[V_0]$$

where $L_k = (1 + L\tau^s)^{1+e^{-\alpha\tau}+\cdots+e^{-(k-1)\alpha\tau}}$. We denote by $I_k \subset \mathscr{M}^k$ the subset such that

$$\{\mathscr{P}_\tau \circ \sup_m S_\tau^m\}^k[V_0] = \sup_{(i_1,\cdots,i_k)\in I_k} S_\tau^{i_k} \cdots S_\tau^{i_1}[V_0].$$

Thus,

$$S_{k\tau}[V_0] \leqslant \sup_{(i_1,\cdots,i_k)\in I_k} L_k S_\tau^{i_k}\cdots S_\tau^{i_1}[V_0]. \tag{7.25}$$

From (7.23), we know that for all $(i_1,\cdots,i_k)\in I_k$

$$M_\tau^{i_k}\cdots M_\tau^{i_1}(P_0)\in\mathscr{K}_0. \tag{7.26}$$

Besides,

$$1\leqslant L_k\leqslant (1+L\tau^s)^{\frac{1}{1-e^{-\alpha\tau}}}\leqslant\delta.$$

Thus for all $(i_1,\cdots,i_k)\in I_k$,

$$L_k(M_\tau^{i_k}\cdots M_\tau^{i_1}(P_0))\in\mathscr{K}_1 \tag{7.27}$$

Recall that

$$L_k S_\tau^{i_k}\cdots S_\tau^{i_1}[V_0](x)=\frac{L_k}{2}x'(M_\tau^{i_k}\cdots M_\tau^{i_1}(P_0))x,$$

then by applying (7.25) and (7.24), we obtain that

$$\begin{aligned}
&S_{(k+1)\tau}[V_0]=S_\tau[S_{k\tau}[V_0]]\\
&\leqslant \sup_{(i_1,\cdots,i_k)\in I_k} S_\tau[L_k S_\tau^{i_k}\cdots S_\tau^{i_1}[V_0]]\\
&\leqslant \sup_{(i_1,\cdots,i_k)\in I_k} (1+L\tau^s)\mathscr{P}_\tau\circ\tilde{S}_\tau[[L_k S_\tau^{i_k}\cdots S_\tau^{i_1}[V_0]]]
\end{aligned} \tag{7.28}$$

Now by Theorem 7.3, there is $\alpha>0$ such that for all $P_1,P_2\in\mathscr{K}_1$ and $m\in\mathscr{M}$

$$d_T(M_\tau^m(P_1),M_\tau^m(P_2))\leqslant e^{-\alpha\tau}d_T(P_1,P_2)$$

Therefore from (7.26) and (7.27) we get that for any $(i_1,\ldots,i_k)\in I_k$ and $m\in\mathscr{M}$

$$\begin{aligned}
&d_T(M_\tau^m M_\tau^{i_k}\cdots M_\tau^{i_1}(P_0),M_\tau^m[L_k(M_\tau^{i_k}\cdots M_\tau^{i_1}(P_0))])\\
&\leqslant e^{-\alpha\tau}d_T(M_\tau^{i_k}\cdots M_\tau^{i_1}(P_0),L_k(M_\tau^{i_k}\cdots M_\tau^{i_1}(P_0)))=e^{-\alpha\tau}\log L_k.
\end{aligned}$$

This implies that

$$M_\tau^m[L_k(M_\tau^{i_k}\cdots M_\tau^{i_1}(P_0))]\leqslant L_k^{e^{-\alpha\tau}}M_\tau^m M_\tau^{i_k}\cdots M_\tau^{i_1}(P_0),\ \ \forall m\in\mathscr{M}$$

which is,

$$S_\tau^m[L_k S_\tau^{i_k}\cdots S_\tau^{i_1}[V_0]]\leqslant L_k^{e^{-\alpha\tau}}S_\tau^m S_\tau^{i_k}\cdots S_\tau^{i_1}[V_0],\ \ \forall m\in\mathscr{M}.$$

Therefore we deduce from the inequality (7.28):

$$\begin{aligned}
S_{(k+1)\tau}[V_0] &\leqslant (1+L\tau^s)\mathscr{P}_\tau[\sup_{m\in\mathscr{M},(i_1,\cdots,i_k)\in I_k} S_\tau^m[L_k S_\tau^{i_k}\cdots S_\tau^{i_1}[V_0]]]\\
&\leqslant (1+L\tau^s)L_k^{e^{-\alpha\tau}}\mathscr{P}_\tau[\sup_{m\in\mathscr{M},(i_1,\cdots,i_k)\in I_k} S_\tau^m S_\tau^{i_k}\cdots S_\tau^{i_1}[V_0]]\\
&=(1+L\tau^s)^{1+e^{-\alpha\tau}+\cdots+e^{-k\alpha\tau}}\{\mathscr{P}_\tau\circ\tilde{S}_\tau\}^{k+1}[V_0].
\end{aligned}$$

Thereby we proved that

$$S_{N\tau}[V_0]\leqslant(1+L\tau^s)^{\frac{1}{1-e^{-\alpha\tau}}}\{\mathscr{P}\circ\tilde{S}_\tau\}^N[V_0],\quad\forall N\in\mathbb{N}.$$

Note that

$$\lim_{\tau\to 0^+}\frac{(1+L\tau^s)^{\frac{1}{1-e^{-\alpha\tau}}}-1}{\tau^{s-1}}=\frac{L}{\alpha},$$

from which we deduce the existence of τ_0 and $K>0$ such that for all $\tau\in[0,\tau_0]$, $N\in\mathbb{N}$ and $V_0(x)=\frac{1}{2}x'Px$ with $P\in[\varepsilon_1 I,\lambda I]$

$$\{S_\tau\}^N[V_0]\leqslant(1+K\tau^{s-1})\{\mathscr{P}\circ\tilde{S}_\tau\}^N[V_0].$$

which leads to

$$\varepsilon^{\mathscr{P}_\tau}(x,\tau,N,V_0)\leqslant K\tau^{\min\{1,r-1\}}\{\mathscr{P}\circ\tilde{S}_\tau\}^N[V_0]\leqslant\frac{K\lambda}{2}\tau^{\min\{1,r-1\}}|x|^2.$$

□

Remark 7.5. It should be pointed out that the crucial point is having $\alpha>0$. If this is not the case ($\alpha=0$), then the iteration (7.28) only leads to:

$$d_T(S_{N\tau}[V_0],\{\mathscr{P}_\tau\circ\tilde{S}_\tau\}^N[V_0])\leqslant LN\tau^s,\ \ \forall N\in\mathbb{N}.$$

7.6 A special case

For every $m\in\mathscr{M}$, let V^m be the quadratic function associated to the infinite horizon linear quadratic optimal control problem indexed by m, namely,

$$V^m=\lim_{T\to\infty}S_T^m[0]\ \ .$$

We next show that if all the Lagrangian functions are the same, then Assumption 7.2 is not necessary to obtain the same error bounds as in Theorem 7.4 and Theorem 7.5.

Corollary 7.8. *Under Assumption 7.1, if the Lagrangian functions are all the same, i.e., $D^m=D$ for all $m\in\mathscr{M}$, then there exist $\alpha>0$, $\tau_0>0$, $L>0$ and $K>0$ such that*

$$\varepsilon_0(x,T,V^m)\leqslant Ke^{-\alpha T}|x|^2,\quad \varepsilon^{\mathscr{P}_\tau}(x,\tau,N,V^m)\leqslant L\tau^{\min\{1,r-1\}}|x|^2,\ \ \forall x,$$

for all $T>0$, $N\in\mathbb{N}$, $\tau\leqslant\tau_0$ and $m\in\mathscr{M}$.

Proof. If all the matrices D^m are equal to a same matrix D, then by taking $y=D^{\frac{1}{2}}x$ Problem 7.1 is equivalent to the following switching optimal control problem:

Problem 7.3.

$$W(y)=\sup_{\mathbf{u}\in W}\sup_{\mu\in\mathscr{D}_\infty}\sup_{T<\infty}\int_0^T\frac{1}{2}|\mathbf{y}(s)|^2-\frac{\gamma^2}{2}|\mathbf{u}(s)|^2\,ds$$

where the state dynamics are given by

$$\dot{\mathbf{y}}(s)=D^{\frac{1}{2}}A^{\mu(s)}D^{-\frac{1}{2}}\mathbf{y}(s)+D^{\frac{1}{2}}\sigma^{\mu(s)}\mathbf{u}(s);\ \ \mathbf{y}(0)=y\in\mathbb{R}^d\ \ . \tag{7.29}$$

By Remark 7.1, the matrix parameters of Problem 7.3 satisfy Assumption 7.1 and Assumption 7.2, hence the error bounds obtained in Theorem 7.4 and Theorem 7.5 hold for Problem 7.3. Therefore the same error bounds hold as well for Problem 7.1 with all $D^m=D$. □

Corollary 7.8 should be compared with Theorem 6.1 and 7.1 in [MK10]. Recall that $\varepsilon(x,\tau,N,V^m)$ equals to the discrete-time approximation error $\varepsilon^{\mathscr{P}_\tau}(x,\tau,N,V^m)$ when $r=+\infty$.

Theorem 7.6 ([MK10])**.** *Under Assumption 7.1, there exists $K>0$ such that*

$$\varepsilon_0(x,T,V^m)\leqslant K(1+|x|^2)/T,\ \ \forall x\in\mathbb{R}^d\ \ ,$$

for all $T>0$ and $m\in\mathscr{M}$. If in addition, the dynamics are all the same, i.e., $\sigma^m=\sigma$ for all $m\in\mathscr{M}$, then there exist $\tau_0>0$ and $L>0$ such that

$$\varepsilon(x,\tau,N,V^m)\leqslant L(\tau+\sqrt{\tau})(1+|x|^2),\ \ \forall x\in\mathbb{R}^d\ \ ,$$

for all $N\in\mathbb{N}$, $\tau\leqslant\tau_0$ and $m\in\mathscr{M}$.

7.7 Proofs of the technical lemmas

7.7.1 Proof of Lemma 7.3

For two functions $\mu,\nu\in\mathscr{D}_T$ we consider the metric $d(\mu,\nu)$ defined by the measure of subset on which the two controls μ and ν differ from each other:

$$d(\mu,\nu)=\int_0^T 1_{\mu\neq\nu}dt. \tag{7.30}$$

The proof of Lemma 7.3 needs the next lemma. It shows that the objective function is continued on the variable $\mu\in\mathscr{D}_T$ with respect to the metric d defined in (7.30).

Lemma 7.9. *Let $V_0:\mathbb{R}^d\to\mathbb{R}$ be a locally Lipschitz function. Let $x\in\mathbb{R}^n$ and $T>0$. Given $\mu\in\mathscr{D}_T$ and $\mathbf{u}\in W_T$, for any $\varepsilon>0$, there is $\delta_0>0$ such that*

$$|J(x,T;\mathbf{u},\mu)+V_0(\mathbf{x}(T))-J(x,T;\mathbf{u},\tilde{\mu})-V_0(\tilde{\mathbf{x}}(T))|\leqslant\varepsilon,$$

for all $\tilde{\mu}\in\mathscr{D}_T$ such that $d(\mu,\tilde{\mu})\leqslant\delta_0$ and $(\mathbf{x},\mathbf{u},\mu)$, $(\tilde{\mathbf{x}},\mathbf{u},\tilde{\mu})$ satisfying (7.5)*.*

Proof. Let any $\tilde{\mu}\in\mathscr{D}_T$ and denote:

$$\delta=d(\mu,\tilde{\mu}).$$

Let $\mathbf{x}$ and $\tilde{\mathbf{x}}$ be respectively the solutions to (7.5) under the control $(\mathbf{u},\mu)$ and $(\mathbf{u},\tilde{\mu})$ and with initial state $\mathbf{x}(0)=x$. Thus

$$\mathbf{x}(t)-\tilde{\mathbf{x}}(t)=\int_0^t A^{\mu(s)}\mathbf{x}(s)+\sigma^{\mu(s)}\mathbf{u}(s)-(A^{\tilde{\mu}(s)}\tilde{\mathbf{x}}(s)+\sigma^{\tilde{\mu}(s)}\mathbf{u}(s))ds,\ \ \forall t\in[0,T].$$

Denote

$$L=\max(\max_m\|A^m\|,\max_m|\sigma^m|,\max_m\|D^m\|,(\int_0^T(|\mathbf{x}(s)|+|\mathbf{u}(s)|)^2ds)^{1/2}).$$

We have:

$$\begin{array}{ll}
|\mathbf{x}(t)-\tilde{\mathbf{x}}(t)| & \leqslant\int_0^t|A^{\mu(s)}\mathbf{x}(s)-A^{\tilde{\mu}(s)}\mathbf{x}(s)|+|A^{\tilde{\mu}(s)}\mathbf{x}(s)-A^{\tilde{\mu}(s)}\tilde{\mathbf{x}}(s)|+|\sigma^{\mu(s)}-\sigma^{\tilde{\mu}(s)}||\mathbf{u}(s)|ds\\
&\leqslant\int_0^t L|\mathbf{x}(s)-\tilde{\mathbf{x}}(s)|ds+\int_0^t 1_{\mu\neq\tilde{\mu}}(\|A^{\mu(s)}-A^{\tilde{\mu}(s)}\||\mathbf{x}(s)|+|\sigma^{\mu(s)}-\sigma^{\tilde{\mu}(s)}||\mathbf{u}(s)|)ds\\
&\leqslant\int_0^t L|\mathbf{x}(s)-\tilde{\mathbf{x}}(s)|ds+2L\int_0^t 1_{\mu\neq\tilde{\mu}}(|\mathbf{x}(s)|+|\mathbf{u}(s)|)ds\\
&\leqslant\int_0^t L|\mathbf{x}(s)-\tilde{\mathbf{x}}(s)|ds+2L(\int_0^t 1_{\mu\neq\tilde{\mu}}ds)^{1/2}(\int_0^t(|\mathbf{x}(s)|+|\mathbf{u}(s)|)^2)^{1/2}\\
&\leqslant\int_0^t L|\mathbf{x}(s)-\tilde{\mathbf{x}}(s)|ds+2L^2\delta^{\frac{1}{2}}, \qquad\qquad \forall t\in[0,T].
\end{array}$$

By Gronwall's Lemma,

$$|\mathbf{x}(t)-\tilde{\mathbf{x}}(t)| \leqslant 2L^2\delta^{\frac{1}{2}}e^{Lt} \leqslant L\delta^{\frac{1}{2}}, \quad \forall t \in [0,T].$$

Then

$$|\tilde{\mathbf{x}}(t)| \leqslant \sup_{t\in[0,T]} |\mathbf{x}(t)| + L\delta^{\frac{1}{2}} \leqslant L, \forall t \in [0,T].$$

Note that L is independent of $\tilde{\mu}$. Now by the local Lipschitz property of V_0 and the boundedness of $\mathbf{x}$ and $\tilde{\mathbf{x}}$, there is $L>0$ such that:

$$|V_0(\mathbf{x}(T)) - V_0(\tilde{\mathbf{x}}(T))| \leqslant L|\mathbf{x}(T)-\tilde{\mathbf{x}}(T)| \leqslant L\delta^{\frac{1}{2}}$$

Besides,

$$\begin{aligned}
&|\textstyle\int_0^T \mathbf{x}(t)' D^{\mu(t)}\mathbf{x}(t) - \tilde{\mathbf{x}}(t)' D^{\tilde{\mu}(t)}\tilde{\mathbf{x}}(t)dt| \\
&\leqslant \textstyle\int_0^T |\mathbf{x}(t)' D^{\mu(t)}(\mathbf{x}(t)-\tilde{\mathbf{x}}(t))| + |\tilde{\mathbf{x}}(t)' D^{\mu(t)}(\mathbf{x}(t)-\tilde{\mathbf{x}}(t))| + |\tilde{\mathbf{x}}(t)'(D^{\mu(t)}-D^{\tilde{\mu}(t)})\tilde{\mathbf{x}}(t)|dt \\
&\leqslant L\textstyle\int_0^T (|\mathbf{x}(t)-\tilde{\mathbf{x}}(t)| + 1_{\mu\neq\tilde{\mu}})dt \\
&\leqslant L(\delta^{\frac{1}{2}}+\delta)
\end{aligned}$$

Thus there is a constant L independent of $\tilde{\mu}$ such that:

$$|J(x,T;\mathbf{u},\mu) + V_0(\mathbf{x}(T)) - J(x,T;\mathbf{u},\tilde{\mu}) - V_0(\tilde{\mathbf{x}}(T))| \leqslant L(\delta^{\frac{1}{2}}+\delta)$$

whence for any $\varepsilon>0$ there is $\delta_0>0$ such that

$$|J(x,T;\mathbf{u},\mu) + V_0(\mathbf{x}(T)) - J(x,T;\mathbf{u},\tilde{\mu}) - V_0(\tilde{\mathbf{x}}(T))| \leqslant \varepsilon$$

for all $\tilde{\mu} \in \mathscr{D}_T$ such that $d(\mu,\tilde{\mu}) \leqslant \delta_0$. □

Using this, we can prove Lemma 7.3:

Proof of Lemma 7.3. Let V_0 be a locally Lipschitz function. Fix $x \in \mathbb{R}^d$. Let $\mu \in \mathscr{D}_\tau$ and $\mathbf{u} \in W_\tau$ be $\frac{\varepsilon}{2}$-optimal for the optimal control problem $S_\tau[V_0](x)$ (Problem 7.2), that is:

$$S_\tau[V_0](x) \leqslant J(x,\tau;\mathbf{u},\mu) + V_0(\mathbf{x}(\tau)) + \frac{\varepsilon}{2}. \tag{7.31}$$

By Lemma 7.9, there is $\delta_0>0$ such that:

$$|J(x,\tau;\mathbf{u},\mu) + V_0(\mathbf{x}(\tau)) - J(x,\tau;\mathbf{u},\tilde{\mu}) - V_0(\tilde{\mathbf{x}}(\tau))| \leqslant \frac{\varepsilon}{2} \tag{7.32}$$

for all $\tilde{\mu} \in \mathscr{D}_\tau$ such that $d(\mu,\tilde{\mu}) \leqslant \delta_0$. Now it remains to prove that there is at least one piecewise constant function $\tilde{\mu} \in \mathscr{D}_\tau$ such that $d(\mu,\tilde{\mu}) \leqslant \delta_0$. To this end, by Lusin's theorem [Fol99], there is a compact $K \subset [0,\tau]$ such that

$$\int_0^\tau 1_K(t)\,dt > \tau - \delta_0$$

and the restriction of μ on K is continuous, thus uniformly continuous. Let $\delta>0$ such that for all $t,s \in K$ and $|t-s| \leqslant \delta$,

$$|\mu(t)-\mu(s)| \leqslant \frac{1}{2}$$

which implies

$$\mu(t) = \mu(s).$$

Now let $N_0 \in \mathbb{N}$ such that $\frac{1}{N_0} < \delta$. We construct a piecewise constant function $\tilde{\mu} \in \mathscr{D}_\tau$ as following. For $i \in \{0, 1 \cdots, N_0 - 1\}$, let

$$\tilde{\mu}(\frac{i}{N_0}\tau) = \begin{cases} \mu(s), & \text{if there is } s \in K \cap [\frac{i}{N_0}\tau, \frac{i+1}{N_0}\tau) \\ 1, & \text{else} \end{cases}$$

and

$$\tilde{\mu}(t) = \tilde{\mu}(\frac{i}{N_0}\tau), \quad t \in [\frac{i}{N_0}\tau, \frac{i+1}{N_0}\tau).$$

Since $\mu(s) = \mu(t)$ for all $s,t \in K \cap [\frac{i}{N_0}\tau, \frac{i+1}{N_0}\tau)$, it follows that

$$\mu(t) = \tilde{\mu}(t), \quad \forall t \in K.$$

Thus

$$\int_0^\tau 1_{\mu \neq \tilde{\mu}} dt \leqslant \int_0^\tau 1 - 1_K dt \leqslant \delta_0.$$

So $d(\mu, \tilde{\mu}) \leqslant \delta_0$ and $\tilde{\mu}$ is constant on interval $[\frac{i}{N_0}\tau, \frac{i+1}{N_0}\tau)$ for all $i \in \{0, 1, \cdots, N_0 - 1\}$. Hence, by (7.32),

$$J(x,\tau;\mathbf{u},\mu) + V_0(\mathbf{x}(\tau)) \leqslant J(x,\tau;\mathbf{u},\tilde{\mu}) + V_0(\tilde{\mathbf{x}}(\tau)) + \frac{\varepsilon}{2} \leqslant \sup_{i_1,\cdots,i_{N_0}} S^{i_{N_0}}_{\tau/N_0} \cdots S^{i_1}_{\tau/N_0}[V_0](x) + \frac{\varepsilon}{2}$$

Now by (7.31), we get

$$S_\tau[V_0](x) \leqslant \sup_{i_1,\cdots,i_{N_0}} S^{i_{N_0}}_{\tau/N_0} \cdots S^{i_1}_{\tau/N_0}[V_0](x) + \varepsilon \leqslant \sup_N \sup_{i_1,\cdots,i_N} S^{i_N}_{\tau/N} \cdots S^{i_1}_{\tau/N}[V_0](x) + \varepsilon.$$

This is true for any $\varepsilon > 0$, we conclude that:

$$S_\tau[V_0](x) = \sup_N \sup_{i_1,\cdots,i_N} S^{i_N}_{\tau/N} \cdots S^{i_1}_{\tau/N}[V_0](x)$$

for all $x \in \mathbb{R}^d$. Thus

$$S_\tau[V_0] = \sup_N \sup_{i_1,\cdots,i_N} S^{i_N}_{\tau/N} \cdots S^{i_1}_{\tau/N}[V_0].$$

□

7.7.2 Proof of Lemma 7.6

The proof of Lemma 7.6 shall need the following estimates:

Lemma 7.10. *Let $\mathscr{K} \subset \mathrm{S}_d$ be a compact convex set. There exist $\tau_0 > 0$ and $L > 0$ such that*

$$\|M^m_\tau(P) - P - \tau\Phi^m(P_0)\| \leqslant L\tau^2 + L\tau\|P - P_0\|$$

for all $P, P_0 \in \mathscr{K}$, $\tau \in [0, \tau_0]$ and $m \in \mathscr{M}$.

Proof. Let $\tau_0 > 0$ such that for all $P \in \mathscr{K}$, $m \in \mathscr{M}$, the Riccati equation

$$\dot{P} = \Phi^m(P), \;\; P(0) = P,$$

has a solution in $[0, \tau_0]$. Therefore,

$$\tilde{\mathscr{K}} := \{M_t^m(P) : t \in [0, \tau_0], P \in \mathscr{K}, m \in \mathscr{M}\}$$

is compact. Besides, for $P \in \mathscr{K}$ and $m \in \mathscr{M}$, the function $M_\cdot^m(P) : [0, \tau_0] \to \mathrm{S}_d$ is twice differentiable in the variable t and it satisfies:

$$\dot{M}_t^m(P) = \Phi^m(M_t^m(P)), \quad \ddot{M}_t^m(P) = D\Phi^m(M_t^m(P)) \circ \Phi^m(M_t^m(P)), \quad t \in [0, \tau_0].$$

By the mean value theorem, for all $P, P_0 \in \mathscr{K}$ and $\tau \in [0, \tau_0]$

$$\|M_\tau^m(P) - P - \tau\Phi^m(P)\| \leqslant \sup_{t \in (0,\tau)} \|D\Phi^m(M_t^m(P)) \circ \Phi^m(M_t^m(P))\| \tau^2$$

and

$$\|\Phi^m(P) - \Phi^m(P_0)\| \leqslant \sup_{Q \in \mathscr{K}} \|D\Phi^m(Q)\| \|P - P_0\|.$$

Let

$$L = \max\{\sup_m \sup_{P \in \tilde{\mathscr{K}}} \|D\Phi^m(P) \circ \Phi^m(P)\|, \sup_m \sup_{P \in \mathscr{K}} \|D\Phi^m(P)\|\},$$

then we have

$$\|M_\tau^m(P) - P - \tau\Phi^m(P_0)\| \leqslant L\tau^2 + L\tau\|P - P_0\|$$

for all $P, P_0 \in \mathscr{K}$, $\tau \in [0, \tau_0]$ and $m \in \mathscr{M}$. □

Using Lemma 7.10 we give a proof of Lemma 7.6:

Proof of Lemma 7.6. Let any $0 < \delta < 1$ and $\tilde{\mathscr{K}} \subset \mathrm{S}_d$ be the compact convex set defined by:

$$\tilde{\mathscr{K}} := \overline{\mathrm{conv}}(\cup_{P_0 \in \mathscr{K}} \overline{B(P_0, \delta)}).$$

By Lemma 7.10, there exists $\tau_1, L_1 > 0$ such that for all $m \in \mathscr{M}$, $P, P_0 \in \tilde{\mathscr{K}}$ and $\tau \in [0, \tau_1]$

$$M_\tau^m(P) \leqslant P + \tau\Phi^m(P_0) + (L_1\tau^2 + L_1\tau\|P - P_0\|)I. \tag{7.33}$$

Let

$$\begin{aligned} L_2 &= \sup\{\|\Phi^m(P)\| : m \in \mathscr{M}, P \in \mathscr{K}\}, \\ L_0 &= \max(L_1, L_1L_2), \\ \tau_0 &= \min(\tfrac{\delta}{2L_2}, \sqrt{\tfrac{\delta}{2eL_0}}, \tfrac{1}{L_1}, \tau_1). \end{aligned}$$

Let any $N \in \mathbb{N}$, $(i_1, \cdots, i_N) \in \mathscr{M}^N$, $\tau \in [0, \tau_0]$ and $V_0(x) = \frac{1}{2}x'P_0x$ with $P_0 \in \mathscr{K}$. We are going to prove by induction on $k \in \{1, \cdots, N\}$ that:

$$M_{\tau/N}^{i_k} \ldots M_{\tau/N}^{i_1}(P_0) \leqslant P_0 + \frac{\tau}{N}\Phi_{i_1}(P_0) + \cdots + \frac{\tau}{N}\Phi_{i_k}(P_0) + L_0(1 + \frac{1}{N})^k \frac{\tau^2 k^2}{N^2} I \tag{7.34}$$

When $k = 1$, since $\frac{\tau}{N} \in [0, \tau_0]$ and $P_0 \in \tilde{\mathscr{K}}$, by (7.33) we get:

$$\begin{aligned} M_{\tau/N}^{i_1}(P_0) &\leqslant P_0 + \tfrac{\tau}{N}\Phi_{i_1}(P_0) + L_1(\tfrac{\tau}{N})^2 I \\ &\leqslant P_0 + \tfrac{\tau}{N}\Phi_{i_1}(P_0) + L_0(1 + \tfrac{1}{N})(\tfrac{\tau}{N})^2 I. \end{aligned}$$

Suppose that (7.34) is true for some $k \in \{1, \cdots, N-1\}$. That is:

$$M^{i_k}_{\tau/N} \cdots M^{i_1}_{\tau/N}(P_0) \leqslant P_0 + \Delta_k \tag{7.35}$$

where $\Delta_k = \frac{\tau}{N}\Phi_{i_1}(P_0) + \cdots + \frac{\tau}{N}\Phi_{i_k}(P_0) + L_0(1+\frac{1}{N})^k \frac{\tau^2 k^2}{N^2} I$. Since

$$\begin{aligned} \|\Delta_k\| &\leqslant \tfrac{k\tau}{N} L_2 + L_0(1+\tfrac{1}{N})^k \tfrac{\tau^2 k^2}{N^2} \\ &\leqslant \tau L_2 + L_0 e \tau^2 \leqslant \delta, \end{aligned}$$

we have that $P_0 + \Delta_k \in \tilde{\mathscr{K}}$ and by (7.33):

$$\begin{aligned} M^{i_{k+1}}_{\tau/N}(P_0+\Delta_k) &\leqslant P_0 + \Delta_k + \tfrac{\tau}{N}\Phi_{i_{k+1}}(P_0) + (L_1 \tfrac{\tau^2}{N^2} + L_1 \tfrac{\tau}{N}\|\Delta_k\|) I \\ &\leqslant P_0 + \tfrac{\tau}{N}\Phi_{i_1}(P_0) + \cdots + \tfrac{\tau}{N}\Phi_{i_k}(P_0) + L_0(1+\tfrac{1}{N})^k \tfrac{\tau^2 k^2}{N^2} I \\ &\quad + \tfrac{\tau}{N}\Phi_{i_{k+1}}(P_0) + L_1 \tfrac{\tau^2}{N^2} I + L_1 \tfrac{\tau}{N}[\tfrac{k\tau}{N} L_2 + L_0(1+\tfrac{1}{N})^k \tfrac{\tau^2 k^2}{N^2}] I \\ &= P_0 + \tfrac{\tau}{N}\Phi_{i_1}(P_0) + \cdots + \tfrac{\tau}{N}\Phi_{i_k}(P_0) + \tfrac{\tau}{N}\Phi_{i_{k+1}}(P_0) \\ &\quad \tfrac{\tau^2}{N^2}[L_0(1+\tfrac{1}{N})^k k^2 + L_1 + L_1 L_2 k + L_1 L_0 (1+\tfrac{1}{N})^k \tfrac{\tau k^2}{N}] I \\ &\leqslant P_0 + \tfrac{\tau}{N}\Phi_{i_1}(P_0) + \cdots + \tfrac{\tau}{N}\Phi_{i_k}(P_0) + \tfrac{\tau}{N}\Phi_{i_{k+1}}(P_0) \\ &\quad + \tfrac{\tau^2}{N^2}[L_0(1+\tfrac{1}{N})^k (k^2+k+1) + L_0(1+\tfrac{1}{N})^k \tfrac{k^2}{N}] I \\ &\leqslant P_0 + \tfrac{\tau}{N}\Phi_{i_1}(P_0) + \cdots + \tfrac{\tau}{N}\Phi_{i_k}(P_0) + \tfrac{\tau}{N}\Phi_{i_{k+1}}(P_0) \\ &\quad + \tfrac{\tau^2 (k+1)^2}{N^2} L_0 (1+\tfrac{1}{N})^{k+1} I \end{aligned}$$

Thus, by (7.35) and the monotonicity of the flow:

$$\begin{aligned} &M^{i_{k+1}}_{\tau/N} M^{i_k}_{\tau/N} \cdots M^{i_1}_{\tau/N}(P_0) \leqslant M^{i_{k+1}}_{\tau/N}(P_0 + \Delta_k) \\ &\leqslant P_0 + \tfrac{\tau}{N}\Phi_{i_1}(P_0) + \cdots + \tfrac{\tau}{N}\Phi_{i_k}(P_0) + \tfrac{\tau}{N}\Phi_{i_{k+1}}(P_0) + L_0(1+\tfrac{1}{N})^{k+1} \tfrac{\tau^2 (k+1)^2}{N^2} I. \end{aligned}$$

We conclude that:

$$M^{i_N}_{\tau/N} \cdots M^{i_1}_{\tau/N}(P_0) \leqslant P_0 + \tfrac{\tau}{N}\Phi_{i_1}(P_0) + \cdots + \tfrac{\tau}{N}\Phi_{i_N}(P_0) + e L_0 \tau^2 I$$

Denote:

$$g(x) = \sup_{m \in \mathscr{M}} \frac{1}{2}(x' P_0 x + x' \Phi^m(P_0) x).$$

By Lemma 7.10 we have that

$$P_0 + \tau \Phi^m(P_0) \leqslant M^m_\tau(P_0) + L_1 \tau^2 I, \quad \forall \tau \in [0, \tau_0], m \in \mathscr{M}.$$

That is

$$g(x) \leqslant S^m_\tau[V_0](x) + \frac{L_1}{2}\tau^2 |x|^2, \quad \forall \tau \in [0, \tau_0], m \in \mathscr{M}.$$

Therefore,

$$\begin{aligned} S^{i_N}_{\tau/N} \cdots S^{i_1}_{\tau/N}[V_0](x) &= \tfrac{1}{2} x' M^{i_N}_{\tau/N} \cdots M^{i_1}_{\tau/N}(P_0) x \\ &\leqslant \tfrac{1}{2}(x' P_0 x + \tfrac{\tau}{N} x' \Phi_{i_1}(P_0) x + \cdots + \tfrac{\tau}{N} x' \Phi_{i_N}(P_0) x + e L_0 \tau^2 |x|^2) \\ &\leqslant g(x) + \tfrac{e L_0}{2}\tau^2 |x|^2 \\ &\leqslant \sup_m S^m_\tau[V_0](x) + L\tau^2 |x|^2, \ \ \forall x \in \mathbb{R}^d \end{aligned}$$

where $L = \frac{e L_0 + L_1}{2}$ is clearly independent of V_0, N, $(i_1, \cdots, i_N)$ and $\tau \leqslant \tau_0$. We conclude that:

$$\sup_N \sup_{i_1, \cdots, i_N} S^{i_N}_{\tau/N} \cdots S^{i_1}_{\tau/N}[V_0](x) \leqslant \sup_m S^m_\tau[V_0](x) + L\tau^2 |x|^2$$

for all $\tau \in [0, \tau_0]$ and $V_0(x) = x' P_0 x$ with $P_0 \in \mathscr{K}$. Finally we apply Lemma 7.3 to obtain the desired result. □

7.8 Further discussions and a numerical illustration

7.8.1 Linear quadratic Hamiltonians

The contraction result being crucial to our analysis (see Remark 7.5), it is impossible to extend the results to the general case with linear terms as in [McE09]. However, the one step error analysis (Lemma 7.6) is not restricted to the pure quadratic Hamiltonian. The interested reader can verify that the one step error $O(\tau^2)$ still holds in the case of [McE09]. Then by simply adding up the errors to time T, we get that:

$$\varepsilon(x,\tau,N,V_0) \leqslant L(1+|x|^2)N\tau^2 = L(1+|x|^2)T\tau.$$

Note that the term $|x|^2$ is replaced by $(1+|x|^2)$ for the general Hamiltonian with linear terms. This estimate is of the same order as in [McE09] with much weaker assumption, especially the assumption on Σ^m.

7.8.2 A tighter bound on the complexity

From Theorem 7.4 and 7.5, we obtain a tighter bound on the complexity of the algorithm (compared with [MK10]):

Corollary 7.11. *Under Assumptions 7.1 and 7.2, to get an approximation of V of order ε, the number of iterations is*

$$O(\frac{-\log\varepsilon}{\varepsilon}), \quad \text{as } \varepsilon \to 0,$$

whence the number of arithmetic operations is:

$$O(M^{O(\frac{-\log\varepsilon}{\varepsilon})}d^3), \quad \text{as } \varepsilon \to 0. \tag{7.36}$$

7.8.3 A numerical illustration

In this subsection, we present some numerical experiments to illustrate our convergence analysis. We use the following 2-dimensional instance with 3 switching controls.

$$A^1 = \begin{bmatrix} -1.9 & 0.7 \\ 0.7 & -1.8 \end{bmatrix}, \quad A^2 = \begin{bmatrix} -1.5 & 0.15 \\ 0.15 & -1.5 \end{bmatrix}, \quad A^3 = \begin{bmatrix} -1.6 & 0.8 \\ 0.8 & -2.1 \end{bmatrix}$$
$$D^1 = \begin{bmatrix} 1.5 & 0.2 \\ 0.2 & 1.5 \end{bmatrix}, \quad D^2 = \begin{bmatrix} 1.3 & 0 \\ 0 & 1.6 \end{bmatrix}, \quad D^3 = \begin{bmatrix} 1.2 & 0.12 \\ 0.12 & 1.3 \end{bmatrix}$$
$$\Sigma^1 = \begin{bmatrix} 0.17 & -0.01 \\ -0.01 & 0.57 \end{bmatrix} \quad \Sigma^2 = \begin{bmatrix} 0.27 & 0.1 \\ 0.1 & 0.27 \end{bmatrix}, \quad \Sigma^3 = \begin{bmatrix} 0.27 & -0.01 \\ -0.01 & 0.27 \end{bmatrix}$$

It can be checked that Assumption 7.1 and 7.2 are satisfied by taking $c_A = 1.01$, $c_D = 1.7$, $m_D = 1.12$ and $c_\sigma^2/\gamma^2 = 0.571$.

We will divide the unit circle into pieces of length Δ and select the basis functions active at the discretized points. The pruning error obtained in this way is of the same order as Δ. Given a time step τ, a propagation horizon T and a length Δ, the curse of dimensionality method will result an approximation function $\tilde{V}$. Since we do not know the exact value function, we consider the normalized backsubstitution error

$$|H(\tilde{V})|_\infty := \sup_{|x|=1} H(x, \nabla\tilde{V}(x)) \ .$$

instead of the approximation error $V - \tilde{V}$, studied in this chapter. This can be justified by the fact that the backsubstitution error is of same order as the approximation error, i.e., there are constants $C_1 > C_2 > 0$ such that

$$C_2|H(\tilde{V})|_\infty \leqslant \sup_{|x|=1} V(x) - \tilde{V}(x) \leqslant C_1|H(\tilde{V})|_\infty \ ,$$

for all $0 \leqslant \tilde{V} \leqslant V$ given by the supremum of finitely many quadratic functions in G_0. The left inequality can be deduced from Lemme 8.6 and the right inequality can be derived from Proposition 8.5.

Remark 7.6. We implement the numerical method to show the consistency between theoretical estimates and observed experimental results. We need to run the algorithm for sufficiently small τ (from 0.05 to 0.0005) and for sufficiently large number of iterations N (up to 5000). Therefore the instance used is of small dimension d and with small number of values of the switched control $|\mathscr{M}|$.

Remark 7.7. We do not use the SDP based pruning algorithm (Chapter 6) for two reasons. First, the instance is of dimension 2 and all the basis functions are quadratic without affine terms thus homogeneous of degree 2. Therefore, we only need to discretize the unit circle, which is of dimension 1. The second reason is due to the difficulty in the control of the pruning error if we use the SDP relaxation technique and the heuristic algorithms.

In Figure 7.3(a) we show the plot of the logarithm of the backsubstitution error $\log(|H|_\infty)$ with respect to the propagation horizon T, for time step τ fixed to 0.005 and different divided length Δ. We see from Figure 7.3 that for every Δ, the logarithm of $|H|_\infty$ decreases linearly with respect to the propagation horizon T with a same rate. This coincides with the exponential decreasing error bound of the finite horizon truncation error in Theorem 7.4. The error stops decreasing after a certain propagation horizon because the semigroup approximation error can not be reduced by extending the propagation horizon. In Figure 7.3(a), for $\Delta =$ 4e-3 (blue curve), $\Delta =$ 2e-3 (green curve) and $\Delta =$ 1e-3 (red curve) we observe an obvious oscillation of the backsubstitution error after it stops decreasing. The oscillation magnitude decreases as the discretized length Δ decreases and becomes invisible to the naked eye for Δ sufficiently small. Thus it is clear that the oscillation occurs when the pruning error is too large that it covers the semigroup approximation error. Besides, for all Δ smaller than 7e-4 we obtain a same curve without visible oscillation. This indicates that the pruning error is now sufficiently small that it is dominated by the semigroup approximation error and improving further the pruning accuracy will no longer decrease the backsubstitution error.

In Figure 7.3(b), we show the plot of the logarithm of the backsubstitution error $\log(|H|_\infty)$ with respect to the propagation horizon T, for fixed $\Delta =$ 1e-4 and different time step τ. Note that the fixed $\Delta =$ 1e-4 is sufficiently small for all the time steps used so that the backsubstitution error does not oscillate after it stops decreasing. As we mentioned, for every τ, there is a *horizon threshold* T_0 after which the backsubstitution error stops decreasing and we refer to this error level as the *final backsubstitution error*.

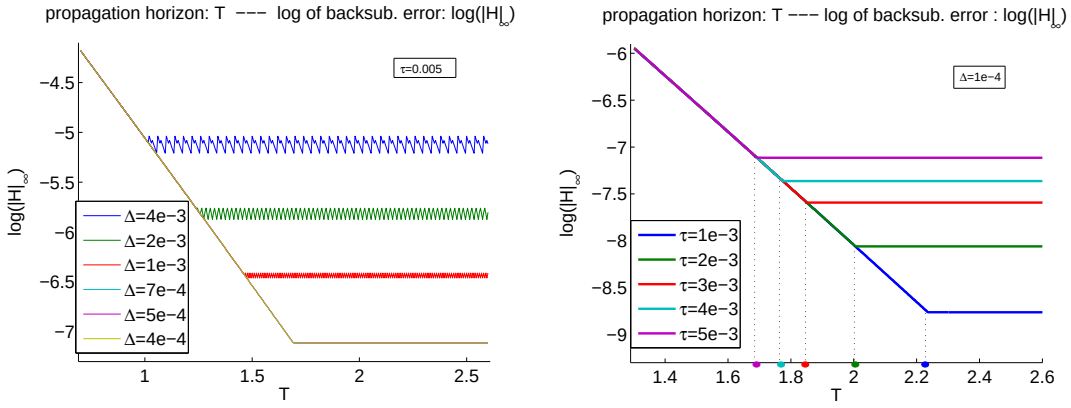
The final backsubstitution error decreases as τ decreases. By Theorem 7.5, the final error decreases at least linearly to the time step τ. In Figure 7.4(a) we see a linear decreasing rate of the final error with respect to the time step τ. This experimental result leads to believe that our error bound $O(\tau)$ is optimal.

In Figure 7.4(b) we show the plot of horizon threshold T_0 obtained by different time step τ. We see that T_0 grows less than $O(-\log(\tau))$. This coincides once again with the exponential decreasing rate of the finite truncation error in Theorem 7.4.

The reader is referred to Table 7.1 for the running time of the algorithm.

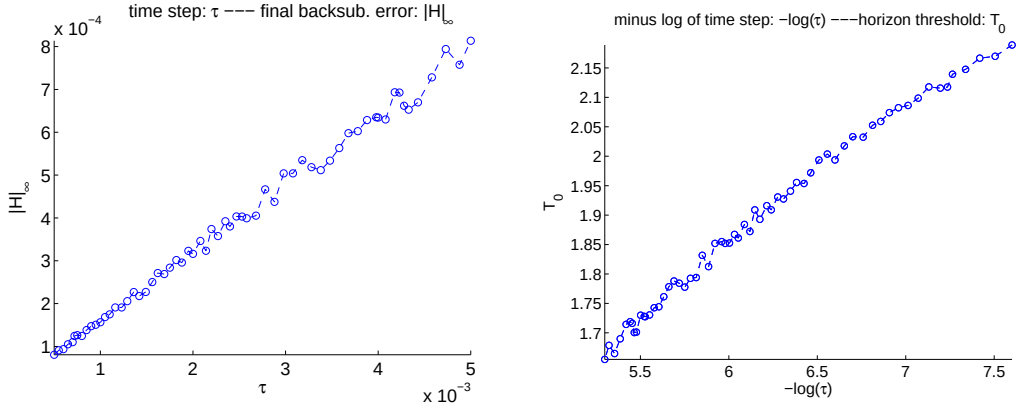
Table 7.1: computational time table

τ	Δ	T_0	cpu time	$\|H\|_\infty$
0.005	8e-4	1.65	60s	8.1e-04
0.004	5.7e-4	1.72	96s	6.4e-4
0.003	4.4e-4	1.79	223s	5e-4
0.002	3e-4	1.91	495s	3.1e-4
0.001	1.5e-4	2.07	2.6e+3s	1.5e-4
9e-4	1.3e-4	2.08	3.4e+3s	1.4e-4
6e-4	8e-5	2.16	1.21e+4s	9.3e-5
5e-4	7.7e-5	2.19	3.19e+4s	8e-5



(a) T–$\log(|H|_\infty)$ (b) T–$\log(|H|_\infty)$

Figure 7.3: Plot of the log backsubstitution error $\log(|H|_\infty)$ w.r.t. the horizon T for time step τ fixed to 0.005 and divided length Δ varying from 4e-3 to 4e-4 (left). Plot of the log backsubstitution error $\log(|H|_\infty)$ w.r.t. the horizon T for divided length Δ fixed to 1e-4 and time step τ varying from 0.001 to 0.005 (right).



(a) τ–$|H|_\infty$ (b) $\log(1/\tau)$–T_0

Figure 7.4: Plot of the final backsubstitution error $|H|_\infty$ w.r.t. the time step τ (left). Plot of the horizon threshold T_0 w.r.t $\log(1/\tau)$ (right).

CHAPTER 8

A new max-plus based algorithm for infinite horizon control problems

In Chapter 6, we introduced McEneaney's curse of dimensionality free method, which applies to the class of infinite horizon switched optimal control problems for which the Hamiltonian takes the form of a supremum of quadratic functions. The method can then be used to approximate any infinite horizon optimal control problem with semiconvex Hamiltonians. The geometric growth rate of the number of basis functions of the method requires an efficient pruning algorithm in the practical implementation. We presented several SDP based pruning algorithms in the same chapter. While the method provides rather good approximation in a reasonable time inaccessible by classical grid based method, the large computational effort required by the SDP based pruning procedure appears to be the bottleneck of the method if we want to reduce the discretization step thus increase the precision order.

In this chapter, we develop a new max-plus based randomized algorithm for the same class of infinite horizon optimal control problems. The major difference between the new algorithm and the previous SDP based curse of dimensionality free method is that, instead of adding a large number of functions and then pruning the less useful ones, the new algorithm finds quickly (linearly in the current number of basis functions), by a randomized procedure, useful quadratic functions and adds only those functions to the set of basis functions. The efficiency of the new algorithm is reflected by a comparison of the running time of the SDP based method and that of the new algorithm for a backsubstitution error of the same order. Experimental results show a speedup around 10 for small instances (small dimension and small number of values of the switch control), up to 100 for instances of more number

of switches. We also observe that, the maximal precision order which can be reached in a reasonable running time by the new algorithm is much better than what can be done by the SDP based algorithm. Besides, with the new randomized algorithm we are now able to deal with instances of more number of switches for which the previous SDP based curse of dimensionality method can not reduce the initial backsubstitution error in a reasonable running time. This will allow us, in the future work, to consider more general infinite horizon optimal control problems with semiconvex Hamiltonians, because the latter one can be approximated fairly well by the supremum of a large number of linear quadratic functions. Finally we give a first convergence result, showing the consistency of the new algorithm.

8.1 Introduction

We have seen (Chapter 6 and 7) that McEneaney's curse of dimensionality free method applies to the class of infinite horizon optimal control problems where the Hamiltonian is given or approximated by the supremum of finitely quadratic functions. The corresponding HJ PDE is written as:

$$0 = -H(x, \nabla V) = -\max_{m \in \mathscr{M}} H^m(x, \nabla V)$$

where

$$H^m(x,p) = (A^m x)'p + \frac{1}{2}x'D^m x + \frac{1}{2}p'\Sigma^m p \ .$$

For simplicity we consider here the pure quadratic case such that Assumption 7.1 holds. Using Lemma 7.3 we know that:

$$V = \sup_T S_T[0] = \sup_T \sup_N \sup_{i_1,\cdots i_N} S^{i_N}_{T/N} \cdots S^{i_1}_{T/N}[0]. \tag{8.1}$$

In essence, in McEneaney's curse of dimensionality free method, we choose a sufficiently large $T > 0$, a sufficiently large integer $N \in \mathbb{N}$ to approximate the value function V as follows:

$$V \simeq \sup_{i_1,\cdots i_N} S^{i_N}_{T/N} \cdots S^{i_1}_{T/N}[0] \ .$$

Each basis function is a quadratic form and can be obtained by solving a sequence of Riccati equations. However, the number of basis functions is exponential to the number of iterations N. To keep the approximation tractable, we need to execute a pruning operation at each iteration, keeping only those basis functions that contribute most to the approximation. The pruning algorithm first introduced in [MDG08] is based on solving the same number of semidefinite programs as the number of basis functions. Further refinement and adaptation of the algorithm in [GMQ11], [SGJM10] and [Sri12] also require to solve a large amount of SDP programs for the pruning phase. Without pruning, the method suffers from the curse of complexity. However, the so far developed pruning procedure is time consuming (see the computation time distribution reported in Table 6.1), and it turns out to be the bottleneck of the method as the required accuracy increases.

The pruning algorithms select in essence "useful" basis functions. The usefulness of a basis function refers to the improvement it can bring to the approximation of the value function. A way to get around the pruning procedure is then to develop a method adding only the most useful basis functions. The next question is then how to identify quickly a basis function which brings a significant improvement to the approximation. For this purpose, first we observe that for a matrix $P \in \mathrm{S}^+_d$, a state point $x_0 \in \mathbb{R}^d$ and an index $m \in \mathscr{M}$, if

$$H^m(x_0, Px_0) > 0 \ ,$$

then for sufficiently small $t > 0$, we have (Lemma 8.2):

$$x_0' M_t^m(P) x_0 > x_0' P x_0 \ . \tag{8.2}$$

Next suppose that we have a subapproximation $\tilde{V} \leqslant V$ which is the supremum of finitely many quadratic forms:

$$\tilde{V}(x) = \sup_{i \in I} \frac{1}{2} x' P_i x, \ \ x \in \mathbb{R}^d \ .$$

Then at every differentiability point x of $\tilde{V}$, we have:

$$H(x, \nabla \tilde{V}) = \max\{H^m(x, P_i x) : m \in \mathscr{M}, \frac{1}{2} x' P_i x = \tilde{V}(x)\} \ . \tag{8.3}$$

Hence if there is a differentiability point x_0 such that $H(x_0, \nabla \tilde{V}) > 0$ then we can find, in linear time (with respect to the number of basis functions), the indices $i \in I$ and $m \in \mathscr{M}$ reaching the maximum in (8.3) at point $x = x_0$ and satisfying

$$H^m(x_0, P_i x_0) > 0.$$

By the above analysis, we can find quickly (by dichotomy for example) a time step $t > 0$ such that the strict inequality (8.2) holds, so that the new quadratic form associated to the matrix $M_t^m(P_i)$ improves the approximation at least at point x_0.

We show that if $\tilde{V} \leqslant V$ is not equal to V, then the set of differentiability points x_0 such that $H(x_0, \nabla \tilde{V}) > 0$ is of positive measure (Proposition 8.11). To find such points, we randomly generate a number of points and select those at which the value of Hamiltonian is positive. This also takes a linear computation time (with respect to the current number of basis functions) for each sampled point. The advantages of this method are as follows. First adding a basis function only requires a number of arithmetical operations that is linear to the number of basis functions. Thus, such an elementary step can be done quickly. Secondly, the time discretization step is automatically adapted by the method, and each added basis function is guaranteed to be useful in some region of the state space, at least at the step when we add it. Numerical experiments (Section 8.4) show that to reach a back-substitution of the same order, the new algorithm takes much less time than the SDP based pruning algorithm (103s vs >10h for the instance used in [MDG08]). We might make a connection between this randomized method for infinite horizon optimal control problems and the so-called "point-base value iteration" [CLZ97, ZZ01] for Partially Observable Markov Decision Process (POMDP). Although developed in very different settings, the two methods share the idea that improving quickly the approximate value function at randomly generated witness points leads to a better performance than improving uniformly but slowly the approximated value function.

The present chapter is organized as follows. In Section 8.2 we present the key observations leading to the new algorithm. In Section 8.3 we state the pseudocode of the algorithm. In Section 8.4 we show the experimental results obtained by the new algorithm and by the SDP based pruning algorithm, for 11 instances of dimension from 4 to 15. Finally, Section 8.5 contains a proof of the almost sure convergence of the randomized algorithm, under some technical assumptions.

8.2 Main ideas of the algorithm

We state the essential observations leading to the new algorithm for solving Problem 7.1, all the notations follow from Section 7.2. We make Assumption 7.1 throughout the Chapter. We shall see

that the method can be extended directly to the more general class of problems to which McEneaney's curse of dimensionality applies. Let $\delta > 0$ be sufficiently small such that G_δ defined in (7.8) is the domain of the semigroup $(S_t)_t$ and that

$$\lim_{T\to+\infty} S_T[V_0] = V, \;\; V_0 \in G_\delta \;\; . \tag{8.4}$$

Denote

$$\bar{P} = \frac{c_A(\gamma-\delta)^2}{c_\sigma^2} I_d \;\; ,$$

and let

$$\mathscr{G} = [0, \bar{P}] \;\; ,$$

be the set of positive semidefinite matrices bounded by $\bar{P}$ in the Loewner order. Then it can be checked by basic calculus that

$$\Phi^m(0) \succcurlyeq 0, \;\; \Phi^m(\bar{P}) \preccurlyeq 0 \;\; , \;\; \forall m \in \mathscr{M} \;\; .$$

Hence using Lemma 2.11 we know that for each $m \in \mathscr{M}$, the interval $\mathscr{G} \subset \mathrm{S}_d^+$ is invariant by the flow $M_\cdot^m(\cdot)$ associated to the function Φ^m.

We show an extension of formula (8.1).

Lemma 8.1. *Let $P_1 \in \mathscr{G}$ which defines a quadratic function:*

$$V_0(x) = \frac{1}{2} x' P_1 x, \;\; x \in \mathbb{R}^d \;\; .$$

If $V_0 \leqslant V$, then

$$V(x) = \sup_{T>0} \sup_N \sup_{i_N,\ldots,i_1} \frac{1}{2} x' \big(M_{T/N}^{i_N} \ldots M_{T/N}^{i_1}(P_1) \big) x, \quad \forall x \in \mathbb{R}^d \;\; . \tag{8.5}$$

Proof. By the monotonicity of the semigroup we know that if $V_0 \leqslant V$ and $T > 0$, then

$$S_T[V_0] \leqslant S_T[V] = V \;\; .$$

Now by (8.4), we get

$$\sup_{T>0} S_T[V_0] = V \;\; .$$

Next by Lemma 7.3, this can be written as:

$$V = \sup_T \sup_N \sup_{i_1,\cdots i_N} S_{T/N}^{i_N} \cdots S_{T/N}^{i_1}[V_0] \;\; .$$

Using the relation between the semigroup and the Riccati flow (7.6), we get immediately (8.5). □

Lemma 8.1 shows that V is the supremum of quadratic functions, each of which is obtained by solving a sequence of Riccati equations. The new algorithm selects only a finite number of them to approximate V. The selection principle is based on the following basic observation.

Lemma 8.2. *Let $P \in \mathrm{S}_d$. If there are $x \in \mathbb{R}^d$ and $m \in \mathscr{M}$ satisfying:*

$$H^m(x, Px) > 0 \;\; ,$$

then there is $t_0 > 0$ such that for all $0 < t \leqslant t_0$:

$$x' M_t^m(P) x > x' P x.$$

Proof. Define the continuously differentiable function $f : [0,T) \times \mathbb{R}^d \to \mathbb{R}$ by

$$f(t) = \frac{1}{2} x' M_t^m(P) x.$$

Then

$$f'(0) = H^m(x_0, Px_0) > 0.$$

Thus there is $t_0 > 0$ such that $f'(t) > 0$ for all $t \in [0,t_0]$. Therefore $f(t) > f(0)$ for all $0 < t \leqslant t_0$. □

Definition 8.1. Let $\{P_i \in \mathrm{S}_d^+ : i \in I\}$ be a finite set of positive semidefinite matrices. We say that a function $\tilde{V}$ is associated to $\{P_i : i \in I\}$ if

$$\tilde{V}(x) = \sup_{i \in I} \frac{1}{2} x' P_i x, \ \ \forall x \in \mathbb{R}^d \ .$$

Define $\delta_H(\tilde{V})$ as the maximal backsubstitution error on the unit sphere if we approximate V by $\tilde{V}$, i.e.,

$$\delta_H(\tilde{V}) := \sup\{H(x,p) : x \in S_d(1), p \in \partial \tilde{V}(x)\} \tag{8.6}$$

where $S_d(1)$ denotes the unit sphere in $\mathbb{R}^d$ and ∂ denotes the subdifferential operator.

Corollary 8.3. *Let $\{P_i \in \mathrm{S}_d^+ : i \in I\}$ be a finite set of positive semidefinite matrices and $\tilde{V}$ be the function associated to $\{P_i : i \in I\}$:*

$$\tilde{V}(x) = \sup_{i \in I} \frac{1}{2} x' P_i x, \ \ x \in \mathbb{R}^d \tag{8.7}$$

such that $\tilde{V} \leqslant V$. If $\delta_H(\tilde{V}) > 0$, then there is $x_0 \in S_d(1)$, $t_0 > 0$, $i_0 \in I$ and $m_0 \in \mathscr{M}$ such that:

$$\tilde{V}(x_0) < \frac{1}{2} x_0' M_t^{m_0}(P_{i_0}) x_0, \ \ t \in (0,t_0]$$

Besides,

$$\frac{1}{2} x' M_t^{m_0}(P_{i_0}) x \leqslant V(x), \ \ \forall x \in \mathbb{R}^d \ . \tag{8.8}$$

Proof. Remark that if $\delta_H(\tilde{V}) > 0$, then there is $x_0 \in S_d(1)$ such that

$$\sup_{q \in \partial \tilde{V}(x_0)} H(x_0, q) > 0, \tag{8.9}$$

By (A.2), we know that:

$$\partial \tilde{V}(x_0) = \mathrm{conv}\{P_i x_0 : \frac{1}{2} x_0' P_i x_0 = \tilde{V}(x_0)\}. \tag{8.10}$$

The function $H(x_0, \cdot)$ is convex and the set $\partial \tilde{V}(x_0)$ is a polytope thus the supremum in (8.9) is attained at an extreme point of the set $\partial \tilde{V}(x_0)$. By (8.10), there is $i_0 \in \arg\max_{i \in I} x_0' P_i x_0$ such that

$$\sup_{q \in \partial \tilde{V}(x_0)} H(x_0, q) = H(x_0, P_{i_0} x_0) \ .$$

Let

$$m_0 \in \arg\max_{m\in\mathscr{M}} H^m(x_0, P_{i_0}x_0) \ ,$$

so that

$$\sup_{q\in\partial\tilde{V}(x_0)} H(x_0,q) = H^{m_0}(x_0,P_{i_0}x_0) \ .$$

Then by (8.9),

$$H^{m_0}(x_0,P_{i_0}x_0) > 0.$$

Now by virtue of Lemma 8.2 we know that there is $t_0 > 0$ such that for all $0 < t \leqslant t_0$

$$x_0' M_t^{m_0}(P_{i_0})x_0 > x_0' P_{i_0} x_0$$

Since $i_0 \in \arg\max_{i\in I} x_0' P_i x_0$, we know that $\frac{1}{2}x_0' P_{i_0}x_0 = \tilde{V}(x_0)$. Finally the last inequality (8.8) follows from the monotonicity of the semigroup. □

Corollary 8.3 implies that there is a feasible way to improve a subapproximation at those points where the backsubstitution error is positive. We next show that $\delta_H(\tilde{V}) > 0$ as long as $\tilde{V}$ is not the exact value function V. Moreover, the normalized approximation error can be bounded through the normalized maximal backsubstitution error $\delta_H(\tilde{V})$.

First let us recall a useful lemma:

Lemma 8.4. *([McE98, Lemma 2.3],[MK10, Lemma 6.7]) Under Assumption 7.1, there is a constant $K > 0$, independent of (x,T), such that for all $\varepsilon \leqslant 1$ and all ε-optimal solution $\mathbf{x}^\varepsilon(\cdot) : [0,T] \to \mathbb{R}^d$ of the optimal control problem $S_T[0](x)$ (Problem 7.2), the following inequality holds:*

$$\int_O^T |\mathbf{x}^\varepsilon(s)|^2 ds \leqslant K(|x|^2 + 1).$$

Proposition 8.5. *Under Assumption 7.1, let $K > 0$ be the constant appearing in Lemma 8.4. Then for all $\tilde{V}$ associated to a finite set of positive semidefinite matrices $\{P_i : i \in I\}$ satisfying $0 \leqslant \tilde{V} \leqslant V$, we have:*

$$V(x) \leqslant \tilde{V}(x) + K\delta_H(\tilde{V})(1+|x|^2), \ \ \forall x \in \mathbb{R}^d \ .$$

Proof. Let any $x \in \mathbb{R}^d$ and $T > 0$. Let $0 < \varepsilon \leqslant 1$ and $(\mathbf{u}^\varepsilon(\cdot), \mu^\varepsilon(\cdot), \mathbf{x}^\varepsilon(\cdot))$ be an ε-optimal pair of the optimal control problem $S_T[0](x)$ (Problem 7.2). Since $\tilde{V}$ is a subsmooth function, we know that:

$$\begin{aligned}
&\tilde{V}(\mathbf{x}^\varepsilon(T)) - \tilde{V}(x) \\
&= \int_0^T \sup_{p\in\partial\tilde{V}(\mathbf{x}^\varepsilon(t))} p'\dot{\mathbf{x}^\varepsilon}(t)\, dt \\
&= \int_0^T \sup_{p\in\partial\tilde{V}(\mathbf{x}^\varepsilon(t))} p'(A^{\mu^\varepsilon(t)}\mathbf{x}^\varepsilon(t) + \sigma^{\mu^\varepsilon(t)}\mathbf{u}^\varepsilon(t))dt
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&S_T[0](x)\\
&\leqslant \int_0^T \frac{1}{2}(\mathbf{x}^\varepsilon(t))' D^{\mu^\varepsilon(t)}\mathbf{x}^\varepsilon(t) - \frac{\gamma^2}{2}|\mathbf{u}^\varepsilon(t)|^2 dt + \varepsilon\\
&\leqslant \tilde{V}(x) + \tilde{V}(\mathbf{x}^\varepsilon(T)) - \tilde{V}(x) + \int_0^T \frac{1}{2}(\mathbf{x}^\varepsilon(t))' D^{\mu^\varepsilon(t)}\mathbf{x}^\varepsilon(t) - \frac{\gamma^2}{2}|\mathbf{u}^\varepsilon(t)|^2 dt + \varepsilon \qquad \text{(since } \tilde{V} \geqslant 0\text{)}\\
&= \tilde{V}(x) + \int_0^T \sup_{p\in\partial\tilde{V}(\mathbf{x}^\varepsilon(t))} p'(A^{\mu^\varepsilon(t)}\mathbf{x}^\varepsilon(t) + \sigma^{\mu^\varepsilon(t)}\mathbf{u}^\varepsilon(t)) + \frac{1}{2}(\mathbf{x}^\varepsilon(t))' D^{\mu^\varepsilon(t)}\mathbf{x}^\varepsilon(t) - \frac{\gamma^2}{2}|\mathbf{u}^\varepsilon(t)|^2 dt + \varepsilon\\
&\leqslant \tilde{V}(x) + \int_0^T \sup_{p\in\partial\tilde{V}(\mathbf{x}^\varepsilon(t))} H(\mathbf{x}^\varepsilon(t), p) dt + \varepsilon
\end{aligned}
$$

Now note that Problem 7.1 is without affine terms, and so H is positively homogeneous of degree 2. By (8.10) we see that the subdifferential of $\tilde{V}$ is positively homogeneous of degree 1. Hence for all $x \in \mathbb{R}^d$,

$$\sup_{p\in\partial\tilde{V}(x)} H(x,p) = \sup_{p\in\partial\tilde{V}(x/|x|)} H(x/|x|,p)|x|^2 \ .$$

Hence,

$$S_T[0](x) \leqslant \tilde{V}(x) + \delta_H(\tilde{V}) \int_0^T |\mathbf{x}^\varepsilon(t)|^2 dt + \varepsilon$$

Now we use Lemma 8.4 and let ε tend to 0 to get:

$$S_T[0](x) \leqslant \tilde{V}(x) + K\delta_H(\tilde{V})(1+|x|^2) \ .$$

Since this is true for arbitrary $x \in \mathbb{R}^d$ and $T > 0$, using (8.1) we get:

$$V(x) \leqslant \tilde{V}(x) + K\delta_H(\tilde{V})(1+|x|^2), \ \ \forall x \in \mathbb{R}^d \ .$$

□

8.3 Algorithm

Based on the above analysis, we propose the following *max-plus randomized descent algorithm* (Algorithm 1).

Algorithm 1 Max-plus randomized descent algorithm

Parameter: a threshold $\vartheta > 0$;
Input: a finite set of matrices $\mathscr{P}^0 \subset \mathscr{G}$ such that the associated function $\tilde{V}_0$ is inferior to V;
for $k = 0, 1, 2, \ldots$ **do**
  Randomly choose a point x_k on the unit sphere $S_d(1)$;
  $P_k \leftarrow \arg\max\{x' P x : P \in \mathscr{P}^k\}$;
  $q_k \leftarrow P_k x_k$;
  $h_k \leftarrow \max\{H^m(x_k, q_k) : m \in \mathscr{M}\}, \quad m_k \leftarrow \arg\max\{H^m(x_k, q_k) : m \in \mathscr{M}\}$;
  if $h_k > \vartheta$, **then**
    $Q \leftarrow \Psi(P_k, m_k, x_k)$;
    $\mathscr{P}^{k+1} \leftarrow \mathscr{P}^k \cup \{Q\}$;
  else
    $\mathscr{P}^{k+1} \leftarrow \mathscr{P}^k$;
  end if
end for

Algorithm 2 Function $\Psi(P, m, x)$ // propagate the matrix P by the m-th Riccati flow

Parameter: a constant $C > 0$;
Input: a matrix P; an index $m \in \mathscr{M}$; a state point $x \in \mathbb{R}^d$;
$q \leftarrow H^m(x, Px)$;
$t \leftarrow 2q/C$;
$Q \leftarrow M_t^m(P)$;
Output: matrix Q.

8.3.1 Parameters and distribution law

Algorithm 1 requires two parameters C and ϑ where ϑ is a precision parameter. The constant C should satisfy:

$$C \geqslant \sup\{\|D\Phi^m(P)\big(\Phi^m(P)\big)\|, \ \ m \in \mathscr{M}, \ \ P \in \mathscr{G}\} \ , \tag{8.11}$$

where hereinafter $\|\cdot\|$ denotes the spectral norm in the space of symmetric matrices. We also need to precise a distribution law on the unit sphere for the random generation in Algorithm 1. Intuitively we prefer those points with large backsubstitution error. However we do not have a priori information on the distribution of the backsubstitution error. Thus we should treat each point on the unit sphere equally and the random generation of points follows the uniform distribution on the unit sphere.

8.3.2 Initial input matrices

The initial set of matrices $\mathscr{P}^0$ can be chosen as $\{P^1, \ldots, P^M\}$ where for every $m \in \mathscr{M}$, the matrix P^m is the solution in $\mathscr{G}$ of the algebraic Riccati equation:

$$(A^m)' P + P A^m + P \Sigma^m P + D^m = 0 \ .$$

Indeed, Assumption 7.1 guarantees that for every $m \in \mathscr{M}$, the value function of the infinite horizon linear quadratic optimal control problem indexed by m is exactly the quadratic function associated to

the matrix P^m. More precisely,

$$\lim_{T\to+\infty} S_T^m[0](x) = \frac{1}{2}x'P^m x, \ \ \forall x \in \mathbb{R}^d, m \in \mathscr{M} \ \ .$$

Since

$$S_T^m[0] \leqslant S_T[0] \ \ \forall T \geqslant 0 \ \ , m \in \mathscr{M}$$

we deduce that

$$\tilde{V}_0 = \sup_{m\in\mathscr{M}} \frac{1}{2}x'P^m x \leqslant V(x), \ \ \forall x \in \mathbb{R}^d \ \ ,$$

thus $\tilde{V}_0$ defined in this way satisfy the constraint require in Algorithm 1.

8.3.3 Complexity analysis

We add the number of operations at each line. It is clear that at iteration k, the number of arithmetical operations is $O(|\mathscr{P}^k| + |\mathscr{M}| + d^3)$.

8.3.4 Practical issues

In practice, we randomly generate a large number of points uniformly distributed on the unit sphere and keep those with large backsubstitution error values. This can be seen as a learning procedure of the backsubstitution error distribution.

Also, the evaluation of the bound in (8.11) is often coarse, leading to a big parameter C and so, the time step estimated in line 4 of Algorithm 2 is too small (of order 10^{-5}). In practice, we start from a proper time step and divide it by two until we find a sufficiently small time step such that the propagation improves strictly the approximation at some point. This procedure will stop after a finite number of searches, actually the number of division is bounded by $\log_2(q/C)$.

8.3.5 Extension to other switched infinite horizon optimal control problem

Although Algorithm 1 is designed for solving Problem 7.1, the same idea can be extended to other class of switched infinite horizon optimal control problems to which McEneaney's curse of dimensionality free methods apply. More precisely, such class of problems can be described as follows. First, the Hamiltonian is given or approximated by a supremum of finitely many simpler Hamiltonians:

$$H(x,p) = \sup_{m\in\mathscr{M}} H^m(x,p) \ \ .$$

Secondly, there is a set of basis functions $\mathscr{B}$ such that for each $m \in \mathscr{M}$ and $t \geqslant 0$, the semigroup S_t^m associated to H^m preserves the structure of basis functions, i.e.,

$$S_t^m[\phi] \in \mathscr{B}, \ \ , \forall \phi \in \mathscr{B} \ \ ,$$

and $S_t^m[\phi]$ is easily computable. If the above two conditions are satisfied, then under some other necessary assumptions for the existence of the solution, Algorithm 1 can be directly adapted to solving the static HJ PDE:

$$H(x,\nabla V) = 0, \ \ \forall x \in \mathbb{R}^d \ \ ; V(0) = 0 \ \ .$$

From this observation, it is clear that we can easily adapt Algorithm 1 solve Problem 6.1, sharing the same structure with Problem 7.1 but with affine terms.

Sridharan et al. [SGJM10] applied the curse of dimensionality free techniques to quantum systems. They obtained an approximate solution for an optimal gate synthesis problem which can be formulated as an optimal control problem with the special unitary group $SU(4)$ as the state space. This leads to a problem of dimension 15 that is computationally intractable by grid based approaches. We briefly recall the problem and refer the reader to the original paper for more details.

Determine the bounds on the number of one and two qubit gates required to perform a desired unitary operation is a problem of special interest in quantum algorithms. One approach to this task of constructing an optimal circuit is to to find a least path-length trajectory on a Riemannian manifold. Let $\{-iH_1,\ldots,-iH_M\}$ correspond to the set of available one and two qubit Hamiltonians. The span of the set $\{-iH_1,\ldots,-iH_M\}$ and all brackets thereof is assumed to be the Lie algebra of the special unitary group $SU(4)$. Let $\{e_1,\ldots,e_M\}$ be the standard basis vectors of $\mathbb{R}^M$. The geodesic distance between U_0 and the identity element I is approximated by the following switched optimal cost control problem:

$$C_0(U_0) = \inf_{v\in\mathscr{V}_g}\{\int_0^{t_{U_0}(v)} \sqrt{v(s)'Rv(s)}ds\},$$

where the state dynamics are given by

$$\frac{dU}{dt} = -i\{\sum_{k=1}^{M} v_k(t)H_k\}U,\ \ U\in SU(4)\ \ , \tag{8.12}$$

and the control space is

$$\mathscr{V}_g = \{v(\cdot)\mid v_k:[0,\infty)\to\{e_1,\ldots,e_M\}\text{ measurable }\}.$$

Moreover, $t_{U_0}(v)$ denotes the time to reach the identity element starting from U_0:

$$t_{U_0}(v) = \inf\{t>0 : U(0)=U_0, U(t)=I, U:[0,t]\to SU(4)\ \text{ satisfying (8.12)}\}$$

The HJ equation of the above optimal cost control problem is given by

$$H(U,\nabla C) = 0,\ \ \forall U\in SU(4)\ \ , C(I)=0\ \ , \tag{8.13}$$

with

$$H(U,p) := \sup_{k\in\{1,\ldots,M\}} H^k(U,p)\ \ ,$$

where

$$H^k(U,p) = \text{trace}[-ipH_kU] - \sqrt{e_k'Re_k}\ \ .$$

The basis functions are chosen to be the affine functions on the space $SU(4)$. For every $k\in\{1,\ldots,M\}$, the semigroup associated to H^k preserves the affine structure. In [SGJM10], the authors applied McEneaney's curse of dimensionality free method with an SDP based pruning strategy to solve 8.13. Our randomized algorithm can be easily adapted to solving equation 8.13.

8.4 Experimental results

In this section, we report the experimental results. A convergence proof will be given in Section 8.5. We compare the numerical results obtained respectively by the SDP based curse of dimensionality free method (Chapter 6) and by the max-plus randomized descent algorithm (Algorithm 1).

The pruning algorithm used is the greedy algorithm (see Section 6.4.3.c), which was shown to be the most efficient (see Figure 6.3) compared to the other ones. The code was mostly written in Matlab (version 7.11.0.584), calling YALMIP (version 3) and SeDuMi (version 1.3) for the resolution of SDP programs. The results were obtained on a single core of an Intel quad core running at 3.10GHz, with 8Gb of memory.

We test the SDP based curse of dimensionality free method and the max-plus randomized algorithm for 11 instances, with dimension d varying from 4 to 15 and the number of discrete controls M varying from 6 to 50. The instances are numbered from 1 to 11. The instance No.4 is the same 6-dimensional instance used in [MDG08] and also in Chapter 6. The instance No. 11 is the 15-dimensional quantum instance with state in $SU(4)$ used in [SGJM10]. All the other 9 instances are generated randomly, all following in the class of Problem 7.2.1 without affine terms. The instances are available upon request and will appear on line later.

Table 8.1 exposes the experimental results for the 11 instances, obtained by the two different methods. The first column gives the number of the instance. The second and third column display the dimension d and the number of the discrete controls M. The fourth column is the (normalized) initial backsubstitution error. The fifth column represents the final backsubstitution error obtained by the SDP based COD-free method after the execution time in the sixth column, or by the max-plus randomized algorithm after the execution time in the eighth column. The seventh and ninth column are respectively the corresponding number of basis functions at the end of execution of the two algorithms. The mark '-' means that the error order in the fifth column can not be obtained by the method in less than 20 hours.

For every instance numbered from 1 to 10, we randomly generate a unitary matrix of the same dimension, and measure the backsubstitution error on the plane generated by the two first column vectors of the unitary matrix. We denote the 10 unitary matrices by $\{U_1,\ldots,U_{10}\}$, with the subscribed number corresponding to the number of instance. For instance No.4, the backsubstitution error reported in Table 8.1 is the maximal value of Hamiltonian of the approximation function on the rectangle $[-2,2]\times[-2,2]$ on the plane generated by e_1 and e_2, with e_1,e_2 the two first standard basis in $\mathbb{R}^6$. For instance No.11, the error reported is the maximal value of Hamiltonian of the approximation function on the plane generated by $\sigma_x\otimes I$ and $\sigma_z\otimes I$ in the Lie algebra of the special group $SU(4)$. For the 9 instances randomly generated, the backsubstitution error is homogeneous with degree 2 and we report the maximal normalized error on the planes.

In short, the two main messages that we get from Table 8.1 are as follows:

- First, the max-plus randomized algorithm can reach the same precision order obtained by the SDP based method with a speedup of order 10 up to 100.
- Second, for instances with more number of discrete controls M, whereas the SDP based curse of dimensionality free method fails to give much more accurate approximation than the initial value function V_0, the max-plus randomized algorithm still achieves to reduce the backsubstitution error in an acceptable running time.

For illustration, we choose some instances and show in Figures 8.1, 8.2, 8.3, 8.4 the 3D plots of the backsubstitution errors obtained by the two methods, on the randomly generated planes. Note that until now, we compare the two methods through the backsubstitution error, which by Proposition 8.5 provides a bound of the maximal normalized approximation error. Hence, from a theoretical point of view, a smaller backsubstitution error implies a smaller maximal approximation error. However, a smaller backsubstitution error does not imply a better subapproximation everywhere in the state space. What we observe in practice is that the randomized algorithm provides a subapproximation much

Table 8.1: Backsubstitution error table

Instance nb.	d	M	initial backsub. error	backsub. error	SDP based		randomized algo	
					cpu time	basis nb.	cpu time	basis nb.
1	4	10	1.2	0.8	280s	100	10s	345
1	4	10	1.2	0.007	-	-	527s	26605
2	4	20	4.5	0.14	-	-	184s	9925
2	4	20	4.5	0.024	-	-	2900s	75831
3	4	50	3.8	1.5	5923s	80	41s	120
3	4	50	3.8	0.1	-	-	155s	5938
3	4	50	3.8	0.034	-	-	3650s	99848
4	6	6	2.3	0.21	188s	170	27s	1231
4	6	6	2.3	0.12	1.5h	170	37s	1636
4	6	6	2.3	0.08	>10h	320	103s	3281
4	6	6	2.3	0.0257	-	-	1217s	10227
5	8	20	0.13	0.1	163s	50	16s	56
5	8	20	0.13	0.01	-	-	90s	8108
5	8	20	0.13	0.0045	-	-	1887s	32860
6	8	30	0.16	0.012	-	-	151s	13698
6	8	30	0.16	0.012	-	-	2994s	76393
7	8	50	0.21	0.01	-	-	120s	10011
7	8	50	0.21	0.0044	-	-	3124s	67515
8	12	20	0.7	0.01	-	-	1485s	44894
8	12	20	0.7	0.0052	-	-	5368s	93292
9	12	30	0.83	0.012	-	-	1322s	39987
9	12	30	0.83	0.007	-	-	6512s	99803
10	12	50	0.8	0.12	-	-	296s	10026
10	12	50	0.8	0.104	-	-	6619s	99992
11	15	6	159	79	>10h	1100	129s	3995
11	15	6	159	31	-	-	15438s	300000

better than the one obtained by the SDP based algorithm on some region in the state space. However, there exits also some region where the subapproximation obtained by the randomized algorithm is worse than the one by SDP based algorithm, see the plot of the difference between the approximation functions in Figures 8.1, 8.2, 8.3, 8.4.

8.5 Convergence result for Algorithm 1

In this section, we give a convergence proof of Algorithm 1, restricted to Problem 7.1.

Algorithm 1 generates a sequence of independent random variables $(x_1, x_2, \dots)$ uniformly distributed on the sphere $S_d(1)$. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be the standard probability space on which is defined the stochastic sequence $(x_1, x_2, \dots)$ equipped with the uniform measure, so that for all $k \in \mathbb{N}$, $t_1, \dots, t_k \in \mathbb{N}$ and borel sets $F_1, \dots, F_k \subseteq S_d(1)$, we have:

$$\mathbb{P}(x_{t_1} \in F_1, \dots, x_{t_k} \in F_k) = \prod_{i=1}^{k} \frac{A(F_i)}{A(S_d(1))} \quad . \tag{8.14}$$

Here $A(\cdot)$ denotes the area.

The output of Algorithm 1 $(\mathscr{P}^1, \mathscr{P}^2, \dots)$ is a nondecreasing sequence of sets of semidefinite matrices which determines then a nondecreasing sequence of functions $(\tilde{V}_1, \tilde{V}_2, \dots)$ as follows:

$$\tilde{V}_k(x) = \sup\{\frac{1}{2} x' P x : P \in \mathscr{P}^k\}, \quad x \in \mathbb{R}^d, \ k = 1, 2, 3, \dots$$

At each iteration $k \in \mathbb{N}$, the function $\tilde{V}_k$ is a subapproximation of the value function V. Denote by e_k the approximation error at iteration k measured by the sup-norm distance on the unit sphere:

$$e_k := \sup\{V(x) - \tilde{V}_k(x) : x \in S_d(1)\} \quad . \tag{8.15}$$

The main result of this section is:

Theorem 8.1. *Let $K > 0$ be the constant in Lemma 8.4. Under Assumption 7.1, surely Algorithm 1 stops adding new quadratic function after a finite number of iterations and almost surely the approximation error obtained by Algorithm 1 is bounded by $K\vartheta$, i.e.:*

$$\mathbb{P}(\lim_{n \to +\infty} e_n \leqslant K\vartheta) = 1 \quad , \tag{8.16}$$

where $\{e_n\}_{n \in \mathbb{N}}$ is defined in 8.15.

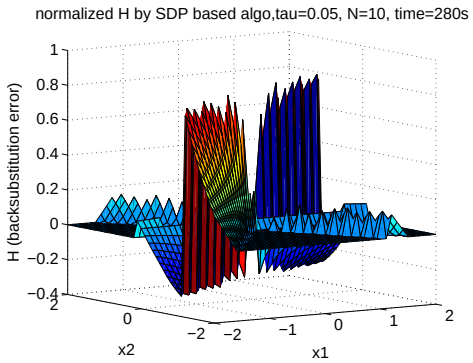
The proof shall need three technical results.

The first technical result (Proposition 8.7) states that each time we add a new quadratic function, the maximal increased value on the unit sphere is not smaller than ϑ^2/C.

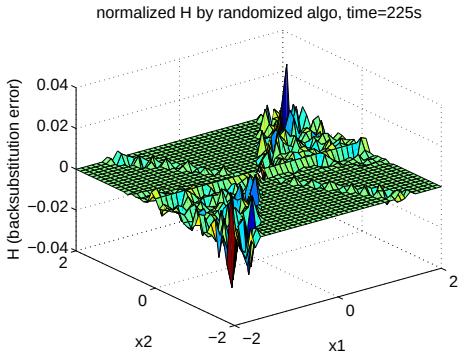
The second technical result (Lemma 8.9) states that for every possible sequence $(x_1, x_2, \dots,)$, the corresponding sequence of subapproximation functions converges uniformly on the unit sphere.

The third technical result (Proposition 8.11) states that for any $\delta > 0$ strictly less than the maximal normalized backsubstitution error of a subapproximation function $\tilde{V}$, there is a set of positive area on the unit sphere, containing only differentiability points of $\tilde{V}$ with Hamiltonian value larger than δ.

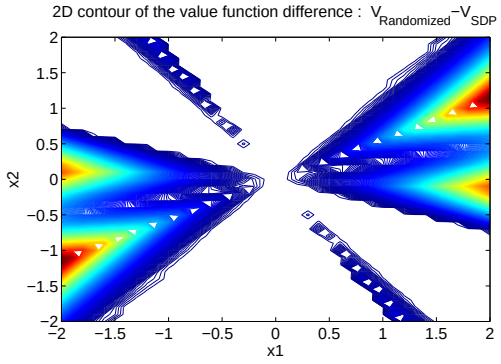
The first and second technical results imply that for every sequence $(x_1, x_2, \dots,)$, Algorithm 1 only adds finitely many basis functions, because otherwise the sequence of subapproximation functions cannot converge uniformly. The third technical result implies that almost surely, the maximal



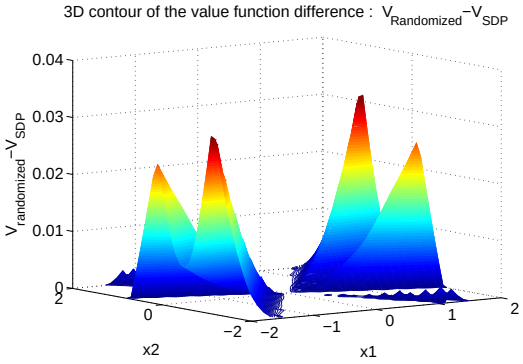
(a) Normalized backsubstitution error, obtained by SDP based algorithm in 280 seconds



(b) Normalized backsubstitution error, obtained by the randomized algorithm in 225 seconds

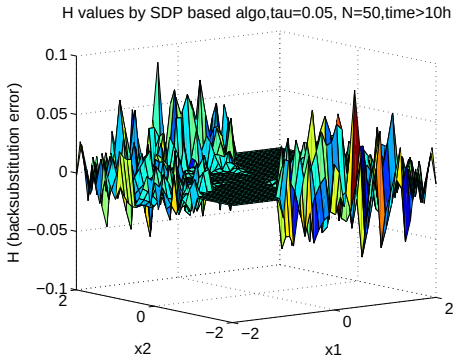


(c) 2D contour of the difference between the two sub-approximation functions obtained by the two algorithms

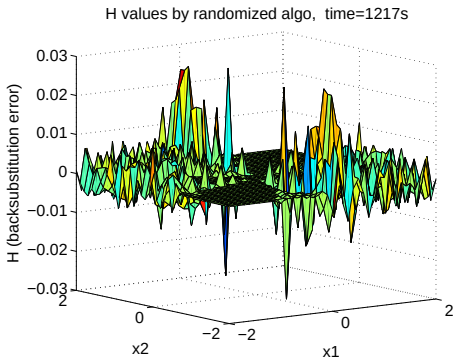


(d) 3D contour of the difference between the two sub-approximation functions obtained by the two algorithms

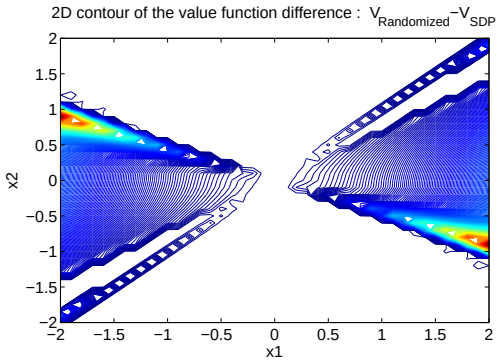
Figure 8.1: Instance number 1, dimension $d = 4$, number of discrete controls $M = 10$; The error is visualized on the plane generated by two vectors $U_1 e_1$ and $U_1 e_2$, where e_1 and e_2 are the first two standard basis vectors in $\mathbb{R}^4$. Recall that $V_{Randomized} \leqslant V$ and $V_{SDP} \leqslant V$ where V is the exact value function to approximate.



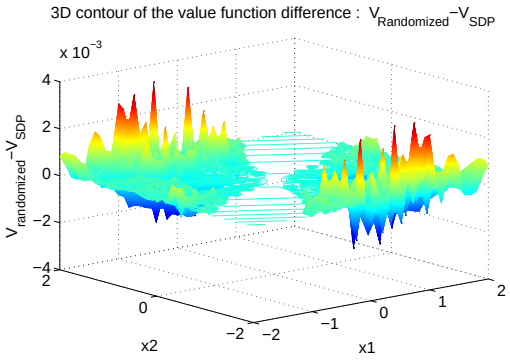
(a) Backsubstitution error, obtained by SDP based algorithm after more than 10 hours



(b) Backsubstitution error, obtained by the randomized algorithm in 1217 seconds

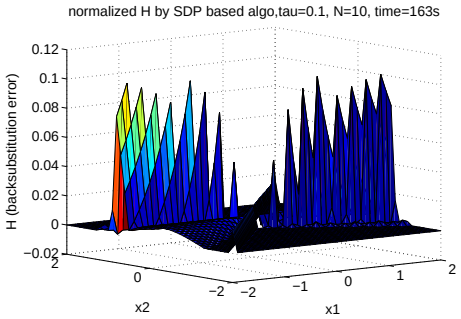


(c) 2D contour of the difference between the two sub-approximation functions obtained by the two algorithms

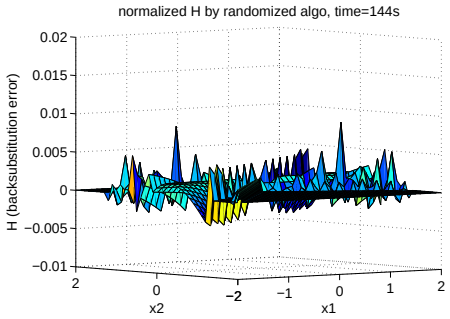


(d) 3D contour of the difference between the two sub-approximation functions obtained by the two algorithms

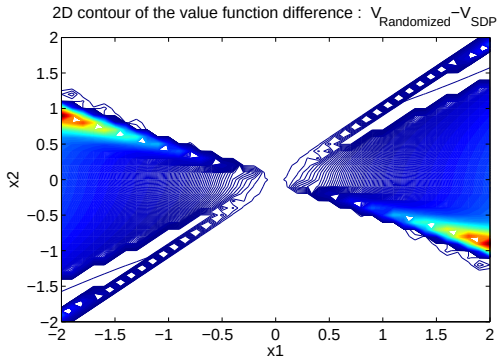
Figure 8.2: Instance number 4, dimension $d = 6$, number of discrete controls $M = 6$; The error is visualized on the plane generated by two vectors e_1 and e_2, where e_1 and e_2 are the first two standard basis vectors in $\mathbb{R}^6$. Recall that $V_{Randomized} \leqslant V$ and $V_{SDP} \leqslant V$ where V is the exact value function to approximate.



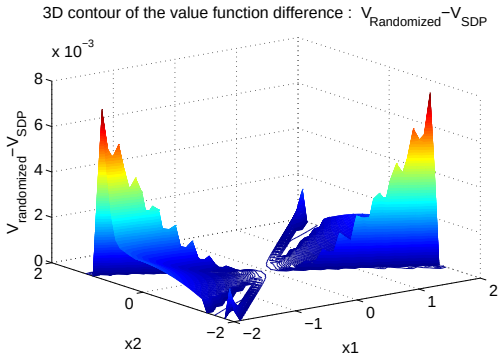
(a) Normalized backsubstitution error, obtained by SDP based algorithm in 163 seconds



(b) Normalized backsubstitution error, obtained by the randomized algorithm in 144 seconds

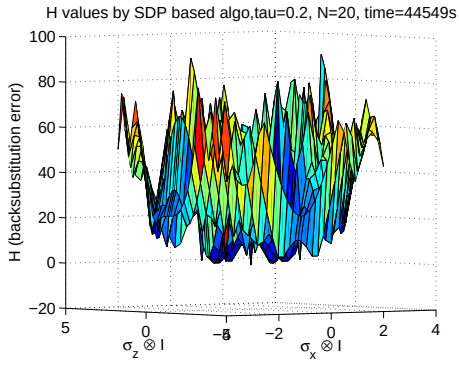


(c) 2D contour of the difference between the two sub-approximation functions obtained by the two algorithms

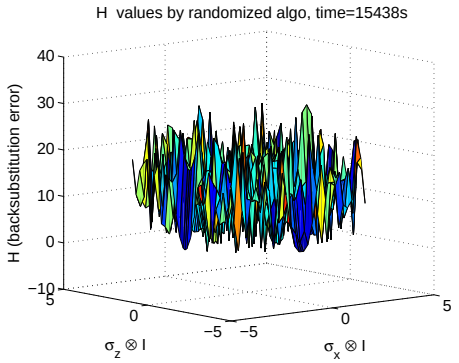


(d) 3D contour of the difference between the two sub-approximation functions obtained by the two algorithms

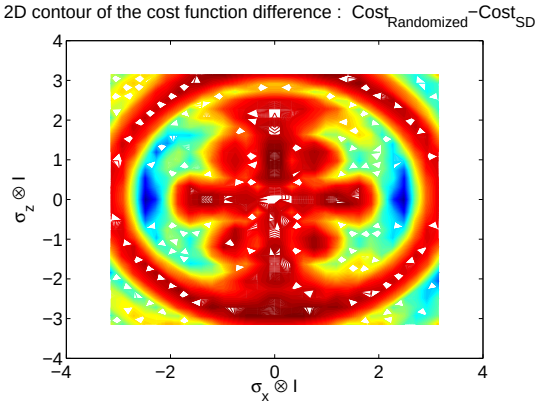
Figure 8.3: Instance number 5, dimension $d = 8$, number of discrete controls $M = 20$; The error is visualized on the plane generated by two vectors $U_5 e_1$ and $U_5 e_2$, where e_1 and e_2 are the first two standard basis vectors in $\mathbb{R}^8$. Recall that $V_{Randomized} \leqslant V$ and $V_{SDP} \leqslant V$ where V is the exact value function to approximate.



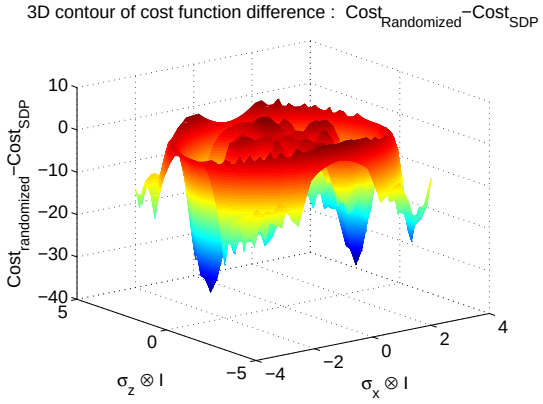
(a) Backsubstitution error, obtained by SDP based algorithm in 44549 seconds



(b) Backsubstitution error, obtained by the randomized algorithm in 15438 seconds



(c) 2D contour of the difference between the two sup-approximation cost functions obtained by the two algorithms



(d) 3D contour of the difference between the two sup-approximation cost functions obtained by the two algorithms

Figure 8.4: Instance number 11, dimension $d = 15$, number of discrete controls $M = 6$; The error is visualized on the plane generated by $\sigma_x \otimes I$ and $\sigma_z \otimes I$ in the Lie algebra of the special unitary group $SU(4)$. Recall that $Cost_{Randomized} \geqslant C_0$ and $Cost_{SDP} \geqslant C_0$ where C_0 is the exact cost function to approximate.

normalized backsubstitution error tends to a value less than the threshold ϑ, because otherwise Algorithm 1 will generate infinitely many times a point at which the backsubstitution error is larger than the threshold ϑ and thus add infinitely many quadratic basis functions.

Finally we use Proposition 8.5 to show that as long as the maximal approximation error on the unit sphere at iteration n is not sufficiently small ($e_n > K\vartheta$), the maximal backsubstitution error on the sphere at iteration n is larger than the threshold ($\delta_H(\tilde{V}_n) > \vartheta$). It follows that almost surely the maximal approximation error on the unit sphere tends to a sufficiently small value ($\lim_{n\to+\infty} e_n \leqslant K\vartheta$).

We present in the next three subsections the proof for the latter three technical results. The proof of Theorem 8.1 is in Section 8.5.4.

In what follows, $(x_1,\dots,x_n,\dots)$ denotes the sequence of independent random points on the unit sphere, drawn with the uniform distribution. The sequence $(\mathscr{P}^1,\dots,\mathscr{P}^n,\dots)$ denotes the corresponding output of Algorithm 1.

8.5.1 Preparation for the proof of Theorem 8.1: first part

In this subsection, we show (Proposition 8.7) that each time we add a new quadratic function, the maximal increased value on the unit sphere is not smaller than ϑ^2/C.

Below is an extended version of Lemma 8.2.

Lemma 8.6. *Let $P \in \mathscr{G}$. If there are $x_0 \in S_d(1)$ and $m \in \mathscr{M}$ satisfying:*

$$q_0 := H^m(x_0, Px_0) > 0,$$

then setting t_0 to $\frac{2q_0}{C}$ we have:

$$\frac{1}{2}x_0' M^m_{t_0}(P)x_0 \geqslant \frac{1}{2}x_0' Px_0 + q_0^2/C.$$

Proof. As in the proof of Lemma 8.2, we denote by $f : [0,T) \times \mathbb{R}^d \to \mathbb{R}$ the function defined by

$$f(t) = \frac{1}{2}x_0' M^m_t(P)x_0, \;\; t \in [0,T).$$

It follows that:

$$f'(t) = \frac{1}{2}x_0' \Phi^m(M^m_t(P))x_0, \;\; t \in [0,T),$$

and

$$f''(t) = \frac{1}{2}x_0^\top D\Phi^m(M^m_t(P)) \circ (\Phi^m(M^m_t(P)))x_0, \;\; t \in [0,T).$$

Therefore,

$$f'(0) = H^m(x_0, Px_0) = q_0 \;\; . \tag{8.17}$$

Recall that $M_t(P) \in \mathscr{G}$ for all $t \in [0,T)$. Thus by (8.11),

$$f''(t) \leqslant C/2, \;\; t \in [0,T). \tag{8.18}$$

Now by mean value theorem, (8.17) and (8.18), we have:

$$f(t) \geqslant f(0) + tq_0 - \frac{C}{4}t^2, \;\; \forall t \in (0,T).$$

Now take $t_0 = 2q_0/C$ we get:

$$f(t_0) \geqslant f(0) + q_0^2/C = \frac{1}{2}x_0' P x_0 + q_0^2/C \ .$$

□

Let $n, l \in \mathbb{N}$. The approximation function $\tilde{V}_{n+l}$ is greater than or equal to the approximation function $\tilde{V}_n$. Denote by $w_{n,n+l}$ the maximal increased value on the unit sphere from iteration n to iteration $n+l$:

$$w_{n,n+l} := \sup\{\tilde{V}_{n+l}(x) - \tilde{V}_n(x) : x \in S_d(1)\}. \tag{8.19}$$

Proposition 8.7. *Let $n \in \mathbb{N}$ and $k \in \mathbb{N}$ be such that $\mathscr{P}^{n+k} \neq \mathscr{P}^n$, then*

$$w_{n,n+k} > \vartheta^2/C \ .$$

Proof. If $|\mathscr{P}^{n+k}| \neq |\mathscr{P}^n|$, then at least a new quadratic basis function is added after step n and before step $n+k$. Without loss of generality suppose that a new quadratic basis function is added at step $n+1$. This only happens when the Boolean test at line 8 in Algorithm 1 is verified. That is, there is $P \in \mathscr{P}^n$, $x \in S_d(1)$ and $m \in \mathscr{M}$ such that

$$P = \arg\max\{x'Qx : Q \in \mathscr{P}^n\}, \ \ H^m(x, Px) > \vartheta \ .$$

The newly added quadratic function is associated to the matrix propagated by the m-th Riccati flow $\Psi(P,m,x)$ (see Algorithm 2). By Lemma 8.6, it is clear that

$$\frac{1}{2}x'\Psi(P,m,x)x > \frac{1}{2}x'Px + \vartheta^2/C = \vartheta^2/C + \tilde{V}_n(x) \ .$$

Hence

$$w_{n,n+k} > \vartheta^2/C \ .$$

□

8.5.2 Preparation for the proof of Theorem 8.1: second part

In this subsection, we show (Lemma 8.9) that each sequence of subapproximation functions converges uniformly on the unit sphere.

Lemma 8.8. *Let $\{P_i : i \in I\}$ be a finite set of matrices contained in the interval $\mathscr{G} = [0, \bar{P}]$. Let $\tilde{V}$ be the function:*

$$\tilde{V}(x) = \sup_{i \in I} \frac{1}{2}x'P_i x, \ \ x \in \mathbb{R}^d \ .$$

Then $\tilde{V}$ is a Lipschitz function on the unit sphere with Lipschitz constant $\|\bar{P}\|$.

Proof. For all $x, y \in S_d(1)$,

$$\begin{aligned}
\tilde{V}(x) - \tilde{V}(y) &\leqslant \sup_{i\in I} \frac{1}{2} x' P_i x - \frac{1}{2} y' P_i y \\
&= \frac{1}{2} \sup_{i\in I} x' P_i (x-y) + y' P_i (x-y) \\
&\leqslant \frac{1}{2} \sup_{i\in I} |x||P_i(x-y)| + |y||P_i(x-y)| \\
&= \sup_{i\in I} |P_i(x-y)| \\
&\leqslant \sup_{i\in I} \|P_i\||x-y|
\end{aligned}$$

Now for all $i \in I$, since $0 \preccurlyeq P_i \preccurlyeq \bar{P}$, we know that $\|P_i\| \leqslant \|\bar{P}\|$. Therefore,

$$\tilde{V}(x) - \tilde{V}(y) \leqslant \|\bar{P}\||x-y| \ .$$

□

Lemma 8.9. *Every sequence of subapproximation functions* $(\tilde{V}_1, \tilde{V}_2, \cdots,)$ *generated by Algorithm 1 converges uniformly on the unit sphere* $S_d(1)$.

Proof. Let $(\tilde{V}_1, \tilde{V}_2, \cdots,)$ be a sequence of subapproximation functions generated by Algorithm 1. Recall that for all n the function $\tilde{V}_n$ is associated to a matrix contained in $\mathscr{G}$. Since $\mathscr{G}$ in bounded, it is clear that the functions $\{\tilde{V}_1, \ldots, \tilde{V}_n, \ldots,\}$ are uniformly bounded on unit sphere. The equicontinuity property of the sequence follows directly from Lemma 8.8. Now applying the Arzelà-Ascoli theorem, we know that there is a subsequence of functions converging uniformly on the unit sphere. We also know that the sequence $(\tilde{V}_1, \tilde{V}_2, \ldots)$ is nondecreasing (pointwise). Therefore the sequence converges uniformly on the unit sphere. □

The following lemma is a direct corollary of Lemma 8.9.

Lemma 8.10. *For all* $\delta > 0$,

$$\lim_{n\to+\infty} \mathbb{P}(\sup_{l>0} w_{n,n+l} > \delta) = 0.$$

Proof. Let $\delta > 0$. By Lemma 8.9, for every sequence $(x_1, x_2, \ldots,)$, the corresponding sequence of approximation functions $(\tilde{V}_1, \tilde{V}_2, \ldots,)$ converges uniformly on the unit sphere. Therefore, for all $\omega \in \Omega$, there is $n \in \mathbb{N}$ such that for all $l \in \mathbb{N}$:

$$\sup\{\tilde{V}_{n+l}(x) - \tilde{V}_n(x) : x \in S_d(1)\} \leqslant \delta.$$

This can be summarized by:

$$\Omega = \{\omega : \bigcup_{n\in\mathbb{N}} \bigcap_{l\in\mathbb{N}} w_{n,n+l} \leqslant \delta\}.$$

Therefore

$$\mathbb{P}(\bigcap_{n\in\mathbb{N}} \{\bigcup_{l\in\mathbb{N}} w_{n,n+l} > \delta\}) = 0,$$

which is equivalent to:

$$\lim_{n\to+\infty} \mathbb{P}(\bigcup_{l\in\mathbb{N}} \{w_{n,n+l} > \delta\}) = 0.$$

□

8.5.3 Preparation for the proof of Theorem 8.1: third part

Let $\mathscr{P}$ be a finite set included in $\mathscr{G}$ and $\tilde{V}$ be the function associated to $\mathscr{P}$. Let $\delta > 0$. Define

$$\mathscr{A}(\tilde{V},\delta) := \underset{P\in\mathscr{P}}{\cup} \mathscr{B}(P,\delta) \ , \tag{8.20}$$

where for $P \in \mathscr{P}$, the set $\mathscr{B}(P,\delta)$ describes the points on the unit sphere where the matrix P is the only active one and the backsubstitution error is strictly bigger than δ, i.e.,

$$\mathscr{B}(P,\delta) := \{x \in S_d(1) : x'Px > \max\{x'Qx : Q \in \mathscr{P}, Q \neq P\},\ \ H(x,Px) > \delta\}$$

The main object of this subsection is to show that for all $0 < \delta < \delta_H(\tilde{V})$, the area of $\mathscr{A}(\tilde{V},\delta)$ is positive (Proposition 8.11). We remark that $\mathscr{A}(\tilde{V},\delta)$ is an open set relative to $S_d(1)$.

Proposition 8.11. *For all $\delta < \delta_H(\tilde{V})$, the area of $\mathscr{A}(\tilde{V},\delta)$ is positive:*

$$A(\mathscr{A}(\tilde{V},\delta)) > 0 \ .$$

Proof. Let $\delta < \delta_H(\tilde{V})$. Since $\mathscr{A}(\tilde{V},\delta)$ is open, we only need to show that the set is not empty. Recall that

$$\delta_H(\tilde{V}) := \sup\{H(x,p) : x \in S_d(1), p \in \partial\tilde{V}(x)\} \ .$$

Then there is $x_0 \in S_d(1)$ such that

$$\sup\{H(x_0,p) : p \in \partial\tilde{V}(x_0)\} > \frac{\delta + \delta_H(\tilde{V})}{2} \ .$$

We assume that there is more than one matrix active at point x_0 (otherwise $x_0 \in \mathscr{A}(\tilde{V},\delta)$). By (A.2) and the convexity of H in p:

$$\sup\{H(x_0,p) : p \in \partial\tilde{V}(x_0)\} = \max\{H(x_0,Px_0) : P \in \mathscr{P}, \frac{1}{2}x_0'Px_0 = \tilde{V}(x_0)\} \ . \tag{8.21}$$

Let $P \in \mathscr{P}$ be a matrix reaching the maximum in the latter formula. Then by the continuity of the function H, there is $\varepsilon > 0$ such that

$$H(x,Px) > \delta, \ \ \forall x \in B(x_0;\varepsilon) \ . \tag{8.22}$$

If there is a set $\mathscr{A} \subset S_d(1) \cap B(x_0;\varepsilon)$ of positive area such that

$$\frac{1}{2}x'Px > \max\{\frac{1}{2}x'Qx : Q \in \mathscr{P}, Q \neq P\}, \ \ \forall x \in \mathscr{A} \ , \tag{8.23}$$

then by definition $\mathscr{A} \subset \mathscr{A}(\tilde{V},\delta)$ and so the area of $\mathscr{A}(\tilde{V},\delta)$ is positive. Now suppose that there is no $\mathscr{A} \subset S_d(1) \cap B(x_0;\varepsilon)$ of positive area such that 8.23 holds. Define a function $W : \mathbb{R}^d \to \mathbb{R}$ by

$$W(x) = \sup\{\frac{1}{2}x'Qx, Q \in \mathscr{P}, Q \neq P\} \ \ \forall x \in \mathbb{R}^d.$$

Then V and W differs on a set of measure 0 in the neighborhood $B(x_0;\varepsilon)$. By the continuity of V and W we get:

$$W(x) = \tilde{V}(x), \ \ \forall x \in B(x_0;\varepsilon).$$

In other words, we can remove the quadratic function associated to P without changing the local structure of $\tilde{V}$ in a neighborhood around x_0. Thus there must be another matrix $Q \in \mathscr{P}$ reaching the maximum in (8.21). That explains why there is always a matrix $P \in \mathscr{P}$ and a set $\mathscr{A}$ of positive Lebesgue measure such that (8.22) and (8.23) hold at the same time. □

The next lemma shows that if the randomly generated point at iteration k is in the set $\mathscr{A}(\tilde{V}_k, \vartheta + \delta/2)$, then a new quadratic basis function is added and the maximal increased value on the unit sphere from iteration k to iteration $k+1$ is larger than some fixed value.

Lemma 8.12. *For all $\delta > 0$. If $x_k \in \mathscr{A}(\tilde{V}_k, \vartheta + \delta/2)$, then*

$$w_{k,k+1} > \vartheta^2/C \ .$$

Proof. If $x_k \in \mathscr{A}(\tilde{V}_k, \vartheta + \delta/2)$, then by definition, there is a matrix $P_k \in \mathscr{P}^k$ such that

$$x_k' P_k x_k > \max\{x_k' Q x_k : Q \in \mathscr{P}^k, Q \neq P\}, \ \ H(x_k, P_k x_k) > \vartheta \ .$$

Therefore, the output at line 5 of Algorithm 1 is P_k. The output of line 7 of Algorithm 1 is $h_k = H(x_k, P_k x_k) > \vartheta$. Thus the Boolean function at line 8 in Algorithm 1 is verified. Hence, a new matrix is added, i.e.

$$|\mathscr{P}^{k+1}| = |\mathscr{P}^k| + 1 \ .$$

By Proposition 8.7, this implies that $w_{k,k+1} > \vartheta^2/C$. □

8.5.4 Proof of Theorem 8.1

We are now ready to give a proof of the convergence of Algorithm 1 (Theorem 8.1).

Proof. By Lemma 8.9, for every possible sequence $(x_1, x_2, \dots,)$, the corresponding sequence of sub-approximation functions converges uniformly on the unit sphere. By Proposition 8.7, each time we add a new quadratic function, the maximal increased value on the unit sphere is not smaller than ϑ^2/C. Therefore, surely Algorithm 1 stops adding quadratic function after a finite number of iterations.

Let $\delta > 0$, $n \in \mathbb{N}$ and $l \in \mathbb{N}\backslash\{0\}$. By Proposition 8.7, we know that if $w_{n,n+l-1} \leqslant \vartheta^2/C$, then no quadratic function has been added after step n until step $n+l-1$. In that case,

$$\mathscr{A}(\tilde{V}_n, \vartheta + \delta/2) = \mathscr{A}(\tilde{V}_{n+l-1}, \vartheta + \delta/2) \ ,$$

because $\tilde{V}_n = \tilde{V}_{n+l-1}$. Therefore, if

$$w_{n,n+l-1} \leqslant \vartheta^2/C, \ \ x_{n+l-1} \in \mathscr{A}(\tilde{V}_n, \vartheta + \delta/2) \ ,$$

then

$$x_{n+l-1} \in \mathscr{A}(\tilde{V}_{n+l-1}, \vartheta + \delta/2) \ .$$

If the latter condition is true, then by Lemma 8.12,

$$w_{n,n+l} \geqslant w_{n+l-1,n+l} > \vartheta^2/C.$$

In other words,

$$\begin{aligned} &\{w_{n,n+l-1} \leqslant \vartheta^2/C\} \cap \{x_{n+l-1} \in \mathscr{A}(\tilde{V}_n, \vartheta + \delta/2)\} \\ &\subseteq \{w_{n,n+l-1} \leqslant \vartheta^2/C\} \cap \{w_{n,n+l} > \vartheta^2/C\} \ . \end{aligned}$$

Hence,

$$\bigcup_{0 \leqslant k \leqslant l-1} \{x_{n+k} \in \mathscr{A}(\tilde{V}_n, \vartheta + \delta/2)\} \subseteq \{w_{n,n+l} > \vartheta^2/C\} \ .$$

Then we have:

$$\begin{aligned}
&\mathbb{P}(\{e_n > K(\vartheta+\delta)\} \cap \{w_{n,n+l} > \vartheta^2/C\}) \\
&\geqslant \mathbb{P}(\underset{0\leqslant k\leqslant l-1}{\cup} \{x_{n+k} \in \mathscr{A}(\tilde{V}_n, \vartheta+\delta/2)\} \cap \{e_n > K(\vartheta+\delta)\}) \\
&= \mathbb{P}(\{e_n > K(\vartheta+\delta)\}) - \mathbb{P}(\{\underset{0\leqslant k\leqslant l-1}{\cap} \{x_{n+k} \notin \mathscr{A}(\tilde{V}_n, \vartheta+\delta/2)\} \cap \{e_n > K(\vartheta+\delta)\}) \ .
\end{aligned}$$

Now denote:

$$\alpha(x_1,\dots,x_{n-1}) := A(\mathscr{A}(\tilde{V}_n, \vartheta+\delta/2))/A(S_d(1)) \ .$$

Then by the independence of random variables $(x_1, x_2, \dots)$, we get:

$$\begin{aligned}
&\mathbb{P}(\underset{0\leqslant k\leqslant l-1}{\cap} x_{n+k} \notin \mathscr{A}(\tilde{V}_n, \vartheta+\delta/2) \cap \{e_n > K(\vartheta+\delta)\}) \\
&= \mathbb{E}(1_{e_n > K(\vartheta+\delta)}(1-\alpha(x_1,\dots,x_{n-1}))^l) \ .
\end{aligned}$$

Hence we have:

$$\mathbb{P}(\{e_n > K(\vartheta+\delta)\} \cap \{w_{n,n+l} > \vartheta^2/C\}) \tag{8.24}$$

$$\geqslant \mathbb{P}(\{e_n > K(\vartheta+\delta)\}) - \mathbb{E}(1_{e_n > K(\vartheta+\delta)}(1-\alpha(x_1,\dots,x_{n-1}))^l) \ . \tag{8.25}$$

By Proposition 8.5, we know that

$$\Omega = \{e_n \leqslant K\delta_H(\tilde{V}_n)\}.$$

Thus

$$\{e_n > K(\vartheta+\delta)\} \subseteq \{\delta_H(\tilde{V}_n) > \vartheta+\delta\}.$$

By Proposition 8.11, we know that

$$\{\delta_H(\tilde{V}_n) > \vartheta+\delta\} \subseteq \{\alpha(x_1,\dots,x_{n-1}) > 0\}.$$

Therefore, we obtain that:

$$\{e_n > K(\vartheta+\delta)\} \subseteq \{\alpha(x_1,\dots,x_{n-1}) > 0\}. \tag{8.26}$$

Hence, we apply the dominated convergence theorem and obtain that:

$$\lim_{l\to+\infty} \mathbb{E}(1_{e_n > K(\vartheta+\delta)}(1-\alpha(x_1,\dots,x_{n-1}))^l) = 0 \ . \tag{8.27}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}(\{\sup_{l\in\mathbb{N}} w_{n,n+l} > \vartheta^2/C\}) \\
&= \lim_{l\to+\infty} \mathbb{P}(\{w_{n,n+l} > \vartheta^2/C\}) \\
&\geqslant \lim_{l\to+\infty} \mathbb{P}(\{e_n > K\sqrt{\vartheta+\delta}\} \cap \{w_{n,n+l} > \vartheta^2/C\}) \\
&\geqslant \mathbb{P}(e_n > K(\vartheta+\delta)) - \lim_{l\to+\infty} \mathbb{E}(1_{e_n > K(\vartheta+\delta)}(1-\alpha(x_1,\dots,x_{n-1}))^l) && \text{by (8.25)} \\
&= \mathbb{P}(e_n > K(\vartheta+\delta)). && \text{by (8.27)}
\end{aligned}$$

Now by Lemma 8.10, we know that

$$\lim_{n\to+\infty} \mathbb{P}(\sup_{l\in\mathbb{N}} w_{n,n+l} > \vartheta^2/C) = 0.$$

Hence we have:

$$\lim_{n\to+\infty} \mathbb{P}(e_n > K(\vartheta+\delta)) \leqslant \lim_{n\to+\infty} \mathbb{P}(\sup_{l\in\mathbb{N}} w_{n,n+l} > \vartheta^2/C) = 0.$$

Since $\delta > 0$ is arbitrary, this implies that

$$\mathbb{P}(\lim_{n\to+\infty} e_n \leqslant K\vartheta) = 1.$$

□

8.6 Conclusion and remarks

This chapter is constituted of the most recent work of the thesis. We developed a new max-plus based algorithm, which by experimental results improves the previous SDP based curse of dimensionality method both in terms of the speed and the accuracy. We gave a first convergence proof showing that the method is consistent.

With this new algorithm which allows now to handle infinite horizon switched optimal control problem with more number of switches, we will consider, in the future work, to extend the method to more general class of Hamiltonians, in particular, those which can be locally well approximated by the supremum of linear quadratic forms. For example, semiconvex Hamiltonians will satisfy this requirement. As we showed in Chapter 5, the curse of dimensionality is unavoidable if we approximate smooth semiconvex function by quadratic forms. Hence, the number of switches needed for a given precision order will increase exponentially with respect to the dimension d of the state space. This prevent us from extending the new algorithm to more general class of Hamiltonians in high dimensional case. On the other hand, we remarked a possible reduced need of number of basis functions in the case when the semiconvex function to approximate is flat, see Remark 5.7. This leads us to study the class of optimal control problems with semiconvex and "flat" Hamiltonians, to which we will apply the algorithm developed in the present chapter.

Besides, although we proved the convergence of the algorithm, the complexity (or the convergence speed) of the algorithm for a given precision is not clear at all at the moment. We shall explore more theoretical studies on the complexity of the algorithm, to see to what extent this new algorithm can reduce the curse of dimensionality.

APPENDIX $\mathcal{A}$

On the differential calculus of pointwise max of finitely many smooth functions

We gather some standard results concerning the differential caluclus of pointwise max of finitely many smooth functions. We refer the reader to the book of Rockafellar and Wets' [RW98] and of Clarke *et al.* [CLSW98] for more background. For a function $f : \mathbb{R}^d \to \mathbb{R}$, the proximal subdifferential of f at point $x \in \mathbb{R}^d$, denoted by $\hat{\partial} f(x)$, is defined as:

$$\hat{\partial} f(x) = \{v : f(y) \geqslant f(x) + \langle v, y-x\rangle + o(|y-x|),\ \ \forall y \in \mathbb{R}^d\}.$$

The (general) subdifferential of f at point $x \in \mathbb{R}^d$, written as $\partial f(x)$, is defined by:

$$\partial f(x) = \{\lim_i v_i : v_i \in \hat{\partial} f(x_i), x_i \to x\} \ .$$

If f is a convex function, then $\hat{\partial} f$ and ∂f coincide with the subgradient set of f, i.e.,

$$\hat{\partial} f(x) = \partial f(x) = \{v : f(y) \geqslant f(x) + \langle v, y-x\rangle,\ \ \forall y \in \mathbb{R}^d\} \ .$$

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ represented by:

$$f = \sup_{a \in A} f_a \tag{A.1}$$

where A is a compact and every function $f_a : \mathbb{R}^n \to \mathbb{R}$ is of class C^1. If further, the functions $(a,x) \to f_a(x)$ and $(a,x) \to Df_a(x)$ are continuous, then f is called *subsmooth* function in [RW98].

The following theorem is Theorem 10.31 in [RW98].

Theorem A.1 ([RW98])**.** *Let f be a subsmooth function of form* (A.1)*. The subdifferential of f at point $x \in \mathbb{R}^d$ is:*

$$\partial f(x) = conv\{Df_a(x) : f_a(x) = f(x), a \in A\}. \tag{A.2}$$

The one-side directional derivative of f at point x in the direction v is:

$$f'(x;v) = \sup_{a \in A, f_a(x) = f(x)} Df_a(x)v, \quad \forall x, v \in \mathbb{R}^n. \tag{A.3}$$

Bibliography

[AB99] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis. A Hitchiker's Guide*. Springer, 1999.

[AB09] David Angeli and Pierre-Alexandre Bliman. Convergence speed of unsteady distributed consensus: decay estimate along the settling spanning-trees. *SIAM J. Control Optim.*, 48(1):1–32, 2009.

[ABM11] Maria Soledad Aronna, J. Frederic Bonnans, and Pierre Martinon. A Shooting Algorithm for Optimal Control Problems with Singular Arcs. Research Report RR-7763, INRIA, October 2011.

[ACS00] E. Andruchow, G. Corach, and D. Stojanoff. Geometrical significance of Löwner-Heinz inequality. *Proc. Amer. Math. Soc.*, 128(4):1031–1037, 2000.

[AGL08] M. Akian, S. Gaubert, and A. Lakhoua. The max-plus finite element method for solving deterministic optimal control problems: basic properties and convergence analysis. *SIAM J. Control Optim.*, 47(2):817–848, 2008.

[Ali11] M. D. S. Aliyu. *Nonlinear H_∞-control, Hamiltonian systems and Hamilton-Jacobi equations*. CRC Press, Boca Raton, FL, 2011.

[AQV98] Marianne Akian, Jean-Pierre Quadrat, and Michel Viot. Duality between probability and optimization. In *Idempotency (Bristol, 1994)*, volume 11 of *Publ. Newton Inst.*, pages 331–353. Cambridge Univ. Press, Cambridge, 1998.

[ARCZ01] M. Ait Rami, X. Chen, and X.Y. Zhou. Discrete-time indefinite lq control with state and control dependent noises. In *Decision and Control, 2001. Proceedings of the 40th IEEE Conference on*, volume 2, pages 1249 –1250 vol.2, 2001.

[Bel52] Richard Bellman. On the theory of dynamic programming. *Proc. Nat. Acad. Sci. U. S. A.*, 38:716–719, 1952.

[Bel57] Richard Bellman. *Dynamic programming*. Princeton University Press, Princeton, N. J., 1957.

[Ber11] Dimitri P. Bertsekas. Approximate policy iteration: a survey and some new methods. *J. Control Theory Appl.*, 9(3):310–335, 2011.

[Bet01] John T. Betts. *Practical methods for optimal control using nonlinear programming*, volume 3 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[BGPS06] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE Trans. Inform. Theory*, 52(6):2508–2530, 2006.

[Bha03] Rajendra Bhatia. On the exponential metric increasing property. *Linear Algebra Appl.*, 375:211–220, 2003.

[Bir57] Garrett Birkhoff. Extensions of Jentzsch's theorem. *Trans. Amer. Math. Soc.*, 85:219–227, 1957.

[BMG12] J. Bonnans, Frédéric, Pierre Martinon, and Vincent Grélard. Bocop - A collection of examples. Rapport de recherche RR-8053, INRIA, August 2012.

[Bon69] J.-M. Bony. Principe du maximum, inégalité de harnack et unicité du problème de cauchy pour les opérateurs elliptiques dégénérés. *Annales de l'institut Fourier*, 19(1):277–304, 1969.

[Bör00] Károly Böröczky, Jr. Approximation of general smooth convex bodies. *Adv. Math.*, 153(2):325–341, 2000.

[Bou93] Philippe Bougerol. Kalman filtering with random coefficients and contractions. *SIAM J. Control Optim.*, 31(4):942–959, 1993.

[Brè67] L. M. Brègman. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 7:620–631, 1967.

[Bre70] Haïm Brezis. On a characterization of flow-invariant sets. *Comm. Pure Appl. Math.*, 23:261–263, 1970.

[BS08] Salman Beigi and Peter W. Shor. On the complexity of computing zero-error and holevo capacity of quantum channels. *arxiv:0709.2090*, 2008.

[BT89] Dimitri P. Bertsekas and John N. Tsitsiklis. *Parallel and distributed computation: numerical methods*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[BTEGN09] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2009.

[Bus73] P. J. Bushell. Hilbert's metric and positive contraction mappings in a Banach space. *Arch. Rational Mech. Anal.*, 52:330–338, 1973.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[BY12] Dimitri P. Bertsekas and Huizhen Yu. Q-learning and enhanced policy iteration in discounted dynamic programming. *Math. Oper. Res.*, 37(1):66–94, 2012.

[BZ07] O. Bokanowski and H. Zidani. Anti-dissipative schemes for advection and application to Hamilton-Jacobi-Bellman equations. *J. Sci. Compt*, 30(1):1–33, 2007.

[CCFP12] Simone Cacace, Emiliano Cristiani, Maurizio Falcone, and Athena Picarelli. A patchy dynamic programming scheme for a class of Hamilton-Jacobi-Bellman equations. *SIAM J. Sci. Comput.*, 34(5):A2625–A2649, 2012.

[CD83] I. Capuzzo Dolcetta. On a discrete approximation of the Hamilton-Jacobi equation of dynamic programming. *Appl. Math. Optim.*, 10(4):367–377, 1983.

[CFF04] E. Carlini, M. Falcone, and R. Ferretti. An efficient algorithm for Hamilton-Jacobi equations in high dimension. *Comput. Vis. Sci.*, 7(1):15–29, 2004.

[CFLS94] F. Camilli, M. Falcone, P. Lanucara, and A. Seghini. A domain decomposition method for Bellman equations. In *Domain decomposition methods in scientific and engineering computing (University Park, PA, 1993)*, volume 180 of *Contemp. Math.*, pages 477–483. Amer. Math. Soc., Providence, RI, 1994.

[CGQ04] G. Cohen, S. Gaubert, and J-P. Quadrat. Duality and separation theorem in idempotent semimodules. *Linear Algebra and Appl.*, 379:395–422, 2004.

[CL83] Michael G. Crandall and Pierre-Louis Lions. Viscosity solutions of Hamilton-Jacobi equations. *Trans. Amer. Math. Soc.*, 277(1):1–42, 1983.

[CL84] M. G. Crandall and P.-L. Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Math. Comp.*, 43(167):1–19, 1984.

[Cla75] Frank H. Clarke. Generalized gradients and applications. *Trans. Amer. Math. Soc.*, 205:247–262, 1975.

[CLSW98] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth analysis and control theory*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1998.

[CLZ97] Anthony Cassandra, Michael L. Littman, and Nevin L. Zhang. Incremental pruning: A simple, fast, exact method for partially observable markov decision processes. In *In Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 54–61. Morgan Kaufmann Publishers, 1997.

[CLZ98] Shuping Chen, Xunjing Li, and Xun Yu Zhou. Stochastic linear quadratic regulators with indefinite control weight costs. *SIAM J. Control Optim.*, 36(5):1685–1702 (electronic), 1998.

[Con90] John B. Conway. *A course in functional analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1990.

[Cop71] W. A. Coppel. *Disconjugacy*. Lecture Notes in Mathematics, Vol. 220. Springer-Verlag, Berlin, 1971.

[DD97] T. Drezner and Z. Drezner. Replacing continuous demand with discrete demand in a competitive location model. *NRL*, 44(1):81–95, 1997.

[DM11] P.M. Dower and W.M. McEneaney. A max-plus based fundamental solution for a class of infinite dimensional Riccati equations. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, pages 615 –620, dec. 2011.

[Dob56] R. Dobrushin. Central limit theorem for non-stationary Markov chains. I. *Teor. Veroyatnost. i Primenen.*, 1:72–89, 1956.

[EM82] L. Erbe and S. Mysore. Comparison theorems and nonoscillation for differential equations in a B^*-algebra. *Nonlinear Anal.*, 6(1):21–33, 1982.

[EN95] Simon P. Eveson and Roger D. Nussbaum. An elementary proof of the Birkhoff-Hopf theorem. *Math. Proc. Cambridge Philos. Soc.*, 117(1):31–55, 1995.

[ES84] L. C. Evans and P. E. Souganidis. Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations. *Indiana Univ. Math. J.*, 33(5):773–797, 1984.

[Fal87] M. Falcone. A numerical approach to the infinite horizon problem of deterministic control theory. *Appl. Math. Optim.*, 15(1):1–13, 1987. Corrigenda in *Appl. Math. Optim.*, 23:213–214, 1991.

[Far96] D. R. Farenick. Irreducible positive linear maps on operator algebras. *Proc. Amer. Math. Soc.*, 124(11):3381–3390, 1996.

[Fer00] E. Feron. Nonconvex quadratic programming, semidefinite relaxations and randomization algorithms in information and decision systems. In *System theory: modeling, analysis and control (Cambridge, MA, 1999)*, volume 518 of *Kluwer Internat. Ser. Engrg. Comput. Sci.*, pages 255–274. Kluwer Academic Publishers Group, Boston, MA, 2000.

[FF94] M. Falcone and R. Ferretti. Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations. *Numer. Math.*, 67(3):315–344, 1994.

[FHH+01] Marián Fabian, Petr Habala, Petr Hájek, Vicente Montesinos Santalucía, Jan Pelant, and Václav Zizler. *Functional analysis and infinite-dimensional geometry*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 8. Springer-Verlag, New York, 2001.

[FLS94] Maurizio Falcone, Piero Lanucara, and Alessandra Seghini. A splitting algorithm for Hamilton-Jacobi-Bellman equations. *Appl. Numer. Math.*, 15(2):207–218, 1994. Innovative methods in numerical analysis (Bressanone, 1992).

[FM00] W. H. Fleming and W. M. McEneaney. A max-plus-based algorithm for a Hamilton-Jacobi-Bellman equation of nonlinear filtering. *SIAM J. Control Optim.*, 38(3):683–710, 2000.

[Fol99] Gerald B. Folland. *Real analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York, second edition, 1999. Modern techniques and their applications, A Wiley-Interscience Publication.

[FR75] Wendell H. Fleming and Raymond W. Rishel. *Deterministic and stochastic optimal control*. Springer-Verlag, Berlin, 1975. Applications of Mathematics, No. 1.

[GG04] S. Gaubert and J. Gunawardena. The Perron-Frobenius theorem for homogeneous, monotone functions. *Trans. of AMS*, 356(12):4931–4950, 2004.

[GMQ11] Stephane Gaubert, William M. McEneaney, and Zheng Qu. Curse of dimensionality reduction in max-plus based approximation methods: Theoretical estimates and improved pruning algorithms. In *CDC-ECE*, pages 1054–1061. IEEE, 2011.

[GMSW86] Philip E. Gill, Walter Murray, Michael A. Saunders, and Margaret H. Wright. User's guide for NPSOL (version 4.0): A Fortran package for nonlinear programming. Technical Report SOL 86-2, Department of Operations Research, Stanford University, Stanford, CA 94305, 1986.

[GQ12a] Stephane Gaubert and Zheng Qu. The contraction rate in thompson metric of order-preserving flows on a cone with application to generalized Riccati equations. *arxiv:1206.0448*, 2012.

[GQ12b] Stephane Gaubert and Zheng Qu. Dobrushin ergodicity coefficient for markov operators on cones, and beyond. *arxiv:1302.5226*, 2012.

[GQ13] Stephane Gaubert and Zheng Qu. Markov operators on cones and noncommutative consensus. In *Proceedings of the European Control Conference ECC 2013*, pages 2693–2700. Zurich, 2013.

[Gru93a] P. M. Gruber. Asymptotic estimates for best and stepwise approximation of convex bodies. I. *Forum Math.*, 5(5):281–297, 1993.

[Gru93b] Peter M. Gruber. Asymptotic estimates for best and stepwise approximation of convex bodies. II. *Forum Math.*, 5(6):521–538, 1993.

[Gru07] P. M. Gruber. *Convex and discrete geometry*. Springer, Berlin, 2007.

[Hir89] Morris W. Hirsch. Convergent activation dynamics in continuous time networks. *Neural Networks*, 2(5):331–349, 1989.

[Hla49] E. Hlawka. Ausfüllung und überdeckung konvexer Körper durch konvexe Körper. *Monatsh. Math.*, 53:81–131, 1949.

[Hop63] E. Hopf. An inequality for positive linear integral operators. *Journal of Mathematics and Mechanics*, 12(5):683–692, 1963.

[HS99] Changqing Hu and Chi-Wang Shu. A discontinuous galerkin finite element method for hamilton-jacobi equations. *SIAM J. Sci. Comput*, 21:666–690, 1999.

[HS05] M. W. Hirsch and Hal Smith. Monotone dynamical systems. In *Handbook of differential equations: ordinary differential equations. Vol. II*, pages 239–357. Elsevier B. V., Amsterdam, 2005.

[KM97] V. N. Kolokoltsov and V. P. Maslov. *Idempotent analysis and its applications*, volume 401 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1997. Translation of *Idempotent analysis and its application in optimal control* (Russian), "Nauka" Moscow, 1994 [MR1375021 (97d:49031)], Translated by V. E. Nazaikinskii, With an appendix by Pierre Del Moral.

[KP82] Elon Kohlberg and John W. Pratt. The contraction mapping approach to the Perron-Frobenius theory: why Hilbert's metric? *Math. Oper. Res.*, 7(2):198–210, 1982.

[Kra85] Dieter Kraft. On converting optimal control problems into nonlinear programming problems. In *Computational mathematical programming (Bad Windsheim, 1984)*, volume 15 of *NATO Adv. Sci. Inst. Ser. F Comput. Systems Sci.*, pages 261–280. Springer, Berlin, 1985.

[LAK07] Asma LAKHOUA. *Méthode des éléments finis max-plus pour la résolution numérique de problèmes de commande optimale déterministe*. PhD thesis, École Polytechnique, France, 2007.

[Lew96] A. S. Lewis. Convex analysis on the Hermitian matrices. *SIAM J. Optim.*, 6(1):164–177, 1996.

[LGM10] Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Analysis of a classification-based policy iteration algorithm. In *ICML*, pages 607–614, 2010.

[Lio89] P.-L. Lions. Viscosity solutions of fully nonlinear second order equations and optimal stochastic control in infinite dimensions. II. Optimal control of Zakai's equation. In *Stochastic partial differential equations and applications, II (Trento, 1988)*, volume 1390 of *Lecture Notes in Math.*, pages 147–170. Springer, Berlin, 1989.

[LL07] Jimmie Lawson and Yongdo Lim. A Birkhoff contraction formula with applications to Riccati equations. *SIAM J. Control Optim.*, 46(3):930–951 (electronic), 2007.

[LL08] Hosoo Lee and Yongdo Lim. Invariant metrics, contractions and nonlinear matrix equations. *Nonlinearity*, 21(4):857–878, 2008.

[LMS01] G. L. Litvinov, V. P. Maslov, and G. B. Shpiz. Idempotent functional analysis: an algebraic approach. *Math. Notes*, 69(5):696–729, 2001.

[LPW09] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2009. With a chapter by James G. Propp and David B. Wilson.

[LR04] Victor Lomonosov and Peter Rosenthal. The simplest proof of Burnside's theorem on matrix algebras. *Linear Algebra Appl.*, 383:45–47, 2004.

[LS85] P.-L. Lions and P. E. Souganidis. Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaacs' equations. *SIAM J. Control Optim.*, 23(4):566–583, 1985.

[Lud99] Monika Ludwig. Asymptotic approximation of smooth convex bodies by general polytopes. *Mathematika*, 46(1):103–125, 1999.

[LV92] J.-H. Lin and J. S. Vitter. e-approximations with minimum packing constraint violation. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, STOC '92, pages 771–782, New York, NY, USA, 1992. ACM.

[LW94] Carlangelo Liverani and Maciej P. Wojtkowski. Generalization of the Hilbert metric to the space of positive definite matrices. *Pacific J. Math.*, 166(2):339–355, 1994.

[MA05] Rex A. C. Medeiros and Francisco M. De Assis. Quantum zero-error capacity. *Int. J. Quanum Inform.*, 03:135, 2005.

[Mar73] R. H. Martin, Jr. Differential equations on closed subsets of a Banach space. *Trans. Amer. Math. Soc.*, 179:399–414, 1973.

[Mas87] V. P. Maslov. *Méthodes Operatorielles*. Edition Mir, Moscou, 1987.

[Mau76] H. Maurer. Numerical solution of singular control problems using multiple shooting techniques. *J. of Optimization Theory and Applications*, 18:235–257, 1976.

[McE98] William M. McEneaney. A uniqueness result for the Isaacs equation corresponding to nonlinear H_∞ control. *Math. Control Signals Systems*, 11(4):303–334, 1998.

[McE04] W. M. McEneaney. Max-plus eigenvector methods for nonlinear H_∞ problems: Error analysis. *SIAM J. Control Optim.*, 43(2):379–412, 2004.

[McE06] W. M. McEneaney. *Max-plus methods for nonlinear control and estimation*. Systems & Control: Foundations & Applications. Birkhäuser Boston Inc., Boston, MA, 2006.

[McE07] W. M. McEneaney. A curse-of-dimensionality-free numerical method for solution of certain HJB PDEs. *SIAM J. Control Optim.*, 46(4):1239–1276, 2007.

[McE09] W. M. McEneaney. Convergence rate for a curse-of-dimensionality-free method for Hamilton-Jacobi-Bellman PDEs represented as maxima of quadratic forms. *SIAM J. Control Optim.*, 48(4):2651–2685, 2009.

[MDA05] Cao M., Spielman D., A., and Morse A., S. A lower bound on convergence of a distributed network consensus algorithm. In *Proc. of the joint 44th IEEE Conference on Decision and Control and European Control Conference*, pages 2356–2361. IEEE, 2005.

[MDG08] W. M. McEneaney, A. Deshpande, and S. Gaubert. Curse-of-complexity attenuation in the curse-of-dimensionality-free method for HJB PDEs. In *Proc. of the 2008 American Control Conference*, pages 4684–4690, Seattle, Washington, USA, June 2008.

[Mey00] Carl Meyer. *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. With 1 CD-ROM (Windows, Macintosh and UNIX) and a solutions manual (iv+171 pp.).

[MK10] William M. McEneaney and L. Jonathan Kluberg. Convergence rate for a curse-of-dimensionality-free method for a class of HJB PDEs. *SIAM J. Control Optim.*, 48(5):3052–3079, 2009/10.

[Mor05] Luc Moreau. Stability of multiagent systems with time-dependent communication links. *IEEE Trans. Automat. Control*, 50(2):169–182, 2005.

[NB03] A. Nedić and D. P. Bertsekas. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dyn. Syst.*, 13(1-2):79–110, 2003. Special issue on learning, optimization and decision making.

[NC00] Michael A. Nielsen and Isaac L. Chuang. *Quantum computation and quantum information*. Cambridge University Press, Cambridge, 2000.

[NK07] Carmeliza Navasca and Arthur J. Krener. Patchy solutions of Hamilton-Jacobi-Bellman partial differential equations. In *Modeling, estimation and control*, volume 364 of *Lecture Notes in Control and Inform. Sci.*, pages 251–270. Springer, Berlin, 2007.

[Nus88] R. D. Nussbaum. Hilbert's projective metric and iterated nonlinear maps. *Mem. Amer. Math. Soc.*, 75(391):iv+137, 1988.

[Nus94] Roger D. Nussbaum. Finsler structures for the part metric and Hilbert's projective metric and applications to ordinary differential equations. *Differential Integral Equations*, 7(5-6):1649–1707, 1994.

[NWF78] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. I. *Math. Programming*, 14(3):265–294, 1978.

[OBSC00] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. *Spatial tessellations: concepts and applications of Voronoi diagrams*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, second edition, 2000. With a foreword by D. G. Kendall.

[OS88] Stanley Osher and James A. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *J. Comput. Phys.*, 79(1):12–49, 1988.

[OS91] Stanley Osher and Chi-Wang Shu. High-order essentially nonoscillatory schemes for Hamilton-Jacobi equations. *SIAM J. Numer. Anal.*, 28(4):907–922, 1991.

[OT09] Alex Olshevsky and John N. Tsitsiklis. Convergence speed in distributed consensus and averaging. *SIAM J. Control Optim.*, 48(1):33–55, 2009.

[PBGM62] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The mathematical theory of optimal processes*. Translated from the Russian by K. N. Trirogoff; edited by L. W. Neustadt. Interscience Publishers John Wiley & Sons, Inc. New York-London, 1962.

[Qu13a] Zheng Qu. Contraction of Riccati flows applied to the convergence analysis of a max-plus curse of dimensionality free method. *arxiv:1301.4777*, 2013.

[Qu13b] Zheng Qu. Contraction of Riccati flows applied to the convergence analysis of the max-plus curse of dimensionality free method. In *Proceedings of the European Control Conference ECC 2013*. Zurich, 2013.

[RCMZ01] Mustapha Ait Rami, Xi Chen, John B. Moore, and Xun Yu Zhou. Solvability and asymptotic behavior of generalized Riccati equations arising in indefinite stochastic LQ controls. *IEEE Trans. Automat. Control*, 46(3):428–440, 2001.

[Red72] R. M. Redheffer. The theorems of Bony and Brezis on flow-invariant sets. *Amer. Math. Monthly*, 79:740–747, 1972.

[Rei72] W. T. Reid. *Riccati differential equations*. Academic Press, New York, 1972. Mathematics in Science and Engineering, Vol. 86.

[RKW11] David Reeb, Michael J. Kastoryano, and Michael M. Wolf. Hilbert's projective metric in quantum information theory. *J. Math. Phys.*, 52(8):082201, 33, 2011.

[Roc70] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[Rog64] C. A. Rogers. *Packing and covering*. Cambridge Tracts in Mathematics and Mathematical Physics, No. 54. Cambridge University Press, New York, 1964.

[RW75] R. M. Redheffer and W. Walter. Flow-invariant sets and differential inequalities in normed spaces. *Applicable Anal.*, 5(2):149–161, 1975.

[RW98] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.

[RZ00] Mustapha Ait Rami and Xun Yu Zhou. Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls. *IEEE Trans. Automat. Control*, 45(6):1131–1143, 2000.

[Sch87] Rolf Schneider. Polyhedral approximation of smooth convex bodies. *J. Math. Anal. Appl.*, 128(2):470–474, 1987.

[Sen91] E. Seneta. Applications of ergodicity coefficients to homogeneous Markov chains. In *50 Years after Doeblin: Developments in the Theory of Markov Chains, Markov Processes and Sums of Random Variables*, Blaubeuren, Germany, 2-7 November 1991.

[Set99] J. A. Sethian. Fast marching methods. *SIAM Rev.*, 41(2):199–235, 1999.

[SGJM10] Srinivas Sridharan, Mile Gu, Matthew R. James, and William M. McEneaney. Reduced-complexity numerical method for optimal gate synthesis. *Phys. Rev. A*, 82:042319, Oct 2010.

[Sha11] Alexander Shapiro. Analysis of stochastic dual dynamic programming method. *European J. Oper. Res.*, 209(1):63–72, 2011.

[SM03] R.O. Saber and R.M. Murray. Consensus protocols for networks of dynamic agents. In *American Control Conference, 2003. Proceedings of the 2003*, volume 2, pages 951 – 956, 4-6, 2003.

[Sor96] Pierpaolo Soravia. H_∞ control of nonlinear systems: differential games and viscosity solutions. *SIAM J. Control Optim.*, 34(3):1071–1097, 1996.

[SPGWC10] Mikel Sanz, David Pérez-García, Michael M. Wolf, and Juan I. Cirac. A quantum version of wielandt's inequality. *IEEE Trans. Inf. Theor.*, 56(9):4668–4673, September 2010.

[Sri12] Srinivas Sridharan. Deterministic filtering and max-plus methods for robust state estimation in multi-sensor settings. *arxiv:1211.1449*, 2012.

[SSR10] Rodolphe Sepulchre, Alain Sarlette, and Pierre Rouchon. Consensus in noncommutative spaces. In *Proceedings of the 49th IEEE Conference on Decision and Control*, pages 6596–6601, Atlanta, USA, Dec 2010.

[Str00] Steven H. Strogatz. From kuramoto to crawford: exploring the onset of synchronization in populations of coupled oscillators. *Phys. D*, pages 1–20, 2000.

[SV03] James A. Sethian and Alexander Vladimirsky. Ordered upwind methods for static Hamilton-Jacobi equations: theory and algorithms. *SIAM J. Numer. Anal.*, 41(1):325–363, 2003.

[TBA86] John N. Tsitsiklis, Dimitri P. Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Trans. Automat. Control*, 31(9):803–812, 1986.

[Tho63] A. C. Thompson. On certain contraction mappings in a partially ordered vector space. *Proc. Amer. Math. soc.*, 14:438–443, 1963.

[TVR97] John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Control*, 42(5):674–690, 1997.

[Vaz01] V. V. Vazirani. *Approximation algorithms*. Springer-Verlag, Berlin, 2001.

[VJAJ05] D. Blondel Vincent, Hendrickx Julien, M., Olshevsky Alex, and Tsitsiklis John, N. Convergence in multiagent coordination, consensus, and flocking. In *Proceedings of the joint 44th IEEE Conference on Decision and Control and European Control Conference*, pages 2996–3000. IEEE, 2005.

[YZ99] Jiongmin Yong and Xun Yu Zhou. *Stochastic controls*, volume 43 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1999. Hamiltonian systems and HJB equations.

[ZZ01] Nevin L. Zhang and Weihong Zhang. Speeding up the convergence of value iteration in partially observable markov decision processes. *Journal of Artificial Intelligence Research*, 14:2001, 2001.

Nomenclature

$\perp$ disjointness, page 69

$\mathscr{P}(\mathbf{e})$ abstract simplex, page 66

$\mathscr{X}$ Banach space, page 24

$\mathscr{X}^\star$ dual space of $\mathscr{X}$, page 24

$\Delta(x)$ diameter of the vector x, page 65

$\delta_H(\tilde{V})$ maximal backsubstitution error on the unit sphere, page 171

det determinant, page 120

$\operatorname{diam} T$ projective diameter of T, page 73

$\mathbb{E}$ real Hilbert space, page 37

$\operatorname{extr}(\cdot)$ set of extreme points, page 69

$\omega(\cdot/\cdot)$ oscillation, page 65

$\lfloor x$ the integer part of x, page 129

$\mathscr{K}_d$ the volume of the unit ball in $\mathbb{R}^d$, page 120

$\operatorname{conv}(\cdot)$ convex hull, page 67

$\|\cdot\|_H$ Hilbert's seminorm, page 65

$\|\cdot\|_T$ Thompson's norm, page 65

∂f subdifferential of f, page 191

$\mathscr{P}$ pruning operation, page 136

$\preccurlyeq$ Loewner order, page 25

$\preccurlyeq$ partial order, page 24

$\mathrm{ri}\, X$ relative interior of X, page 118

$\mathrm{Sym}(\mathbb{E})$ symmetric bounded linear operators on $\mathbb{E}$, page 37

$\mathbf{e}$ unit element, page 65

ϑ_d the minimum density of covering of $\mathbb{R}^d$ with Euclidean balls of unit radius , page 120

$A(\cdot)$ $A(S)$ is the area of S, page 179

A_γ Riemannian measure, page 122

$B_T^\star(\mathbf{e})$ unit ball in the dual space of $(\mathscr{X}, \mathbf{e}, \|\cdot\|_T)$, page 67

$D\phi_t(x)$ Fréchet derivative of $\phi(\cdot,\cdot)$ with respect to x at point (t,x), page 27

$d_H(\cdot,\cdot)$ Hilbert's projective metric, page 73

$d_T(\cdot,\cdot)$ Thompson's part metric, page 25

I_n n-dimensional identity matrix, page 53

$M_\cdot(\cdot)$ flow, page 27

$S_d(1)$ unit sphere in $\mathbb{R}$, page 171

$t_{\mathscr{U}}(\cdot,\cdot)$ the first time when a trajectory leaves from the set $\mathscr{U}$, page 27

$x \vee y$ maximum of x and y, page 111

$x \wedge y$ minimum of x and y, page 115

Index