Minimal upper bounds in the löwner order and application to invariant computation for switched systems

Nikolas Stott

▶ **To cite this version:**

Nikolas Stott. Minimal upper bounds in the löwner order and application to invariant computation for switched systems. General Mathematics [math.GM]. Université Paris Saclay (COmUE), 2017. English. NNT : 2017SACLX106 . tel-03596438

HAL Id: tel-03596438
https://theses.hal.science/tel-03596438

Submitted on 3 Mar 2022

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

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

NNT : 2017SACLX106

Majorants minimaux dans l'ordre de Löwner et application au calcul d'invariants de systèmes commutés

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'École polytechnique

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

Thèse présentée et soutenue à Palaiseau, le 23/11/2017, par

Nikolas Stott

Composition du jury :

Yacine Chitour Professeur, Centrale-Supélec	Président
Rajendra Bhatia Professeur, Ashoka University	Rapporteur
Thao Dang Professeur, Verimag	Rapporteur
Xavier Allamigeon Chargé de recherche, INRIA & Ecole polytechnique	Examinateur
Jean-Eric Pin Professeur, IRIF	Examinateur
Sriram Sankaranarayanan Professeur associé, University of Colorado	Examinateur
Stéphane Gaubert Directeur de recherche, INRIA & Ecole polytechnique	Directeur de thèse
Eric Goubault Professeur, Ecole polytechnique	Co-Directeur de thèse

Thèse de doctorat

Remerciements

Je souhaite en tout premier lieu remercier mes quatres directeurs de thèse – Stéphane Gaubert, Éric Goubault, Xavier Allamigeon et Sylvie Putot – pour leur passion de la vérification qu'ils ont su me transmettre, l'autonomie dont ils m'ont laissé jouir, ainsi que leur patience et leurs encouragements lorsque l'une de mes nouvelles pistes était sans cesse *à portée de main*. Les idées présentes dans cette thèse se situent à l'intersection de la vérification informatique formelle, de la théorie du contrôle et de l'optimisation conique et c'était un véritable plaisir de travailler sur ces sujets avec des encadrants experts de ces domaines à mes côtés. Je n'ai cessé d'apprendre de belles mathématiques à leur côté, de mon premier jour de stage jusqu'à tard dans la rédaction de ce manuscrit, presque une nouvelle dans chacune de nos réunions de travail.

Je souhaite ensuite remercier Thao Dang et Rajendra Bhatia d'avoir accepté de rapporter sur ce manuscrit, chacun expert de l'une des parties (presqueà antipodales. Un grand merci également à Yacine Chitour, Jean-Éric Pin et Sriram d'avoir fait partie de mon jury de thèse.

Un grand merci également à mes frères de thèse, Pascal, Vianney, Mateucz, Jean-Bernard et Paulin, pour toutes nos discussions et nos partages plus ou moins tropicaux aux cours de ces années communes au CMAP. Plus généralement, merci au reste de l'équipe tropicale, dont Marianne, Jessica, Corinne et Hanadi.

Merci aux assistantes du CMAP, Nasséra et Alexandra, et à Sylvain à l'informatique pour votre soutien, (trop) souvent de dernière minute (oups).

Merci aux membres de l'ANR CAFEIN, Pierre, Pierre-Loïc, Alexandre, Marc, Tim et Assalé, de m'avoir accueilli aux Angles pour une semaine de ski-info, pour votre enthousiasme, nos partages autour de Kerbal Space Program et les descentes de la même rouge à l'ombre par 10 degrés et pour nos discussions lorsque nous nous sommes recroisés en conférence.

Merci à mes collègues de bureau, qui, malgré mon absence régulière, m'ont permis d'apprécier un peu plus les bureaux et les couloirs parfois austères du CMAP: Martin, Céline, Juliette, Heythem, Fédor, Antoine & Arthur.

Merci à Valérie pour m'avoir permis d'enseigner l'openGL et le C++ pendant (au moins) 3 ans à l'École de Mines et à Hassan pour tous ces moments passés à se répartir les bugs dans

le code des élèves.

Merci à mes 3 séries de colocataires de m'avoir supporté pendant une année de thèse chacuns. Dans la joie du départ et le stress du premier article entre 2 marathons de cinéma fous et des jeux vidéos en buvant du thé à la menthe avec Benoît; le moment difficile du milieu de la deuxième année pour prouver mon "petit lemme facile" avec Maxence et Félix qui ont su agrémenter mes soirées avec Smash Bros et #Manu dans la grande maison à Gentilly; et la rédaction qui n'en finit plus avec François, dans le petit appart parisien, entre un nouveau jeu de société et un partie de Sherlock.

Un grand merci à tous mes autres amis: Pauline, Yolande, Alexandre, Guillaume(s), Blandine, Julien, Thomas, Nicolas, Hombeline, les Ellipses, Adrien, ainsi que _____ que j'ai malheureusement oublié de mentionner...

Enfin, je souhaite remercier les véritables pilliers de cette thèse: Maman, Papa, mes frères Andréas et Alexandre, Elsa notre labdrador, mes grands-parents, oncles, tantes cousin(e)s, Yiannis et Catherine et toute ma famille. Merci pour votre mélange de curiosité et de peur à chaque fois que vous me demandiez ce que je faisais en ce moment, vos encouragement et votre soutien à toute épreuve pendant ces trois années. Cette thèse vous est dédiée et je vous souhaite bon courage pour sa lecture !

Contents

Introduction (français)

Cette thèse développe plusieurs méthodes pour calculer des invariants de systèmes dynamiques. Ce travail est motivé par des problèmes issus de la vérification de programme et de la théorie du contrôle. Ces méthodes reposent sur la géométrie du cône des matrices positives semidéfinies, et plus généralement sur la géométries des cônes. Nous commençons par donner les principales applications motivant ce travail, puis nous résumons nos résultats.

Contexte and motivation

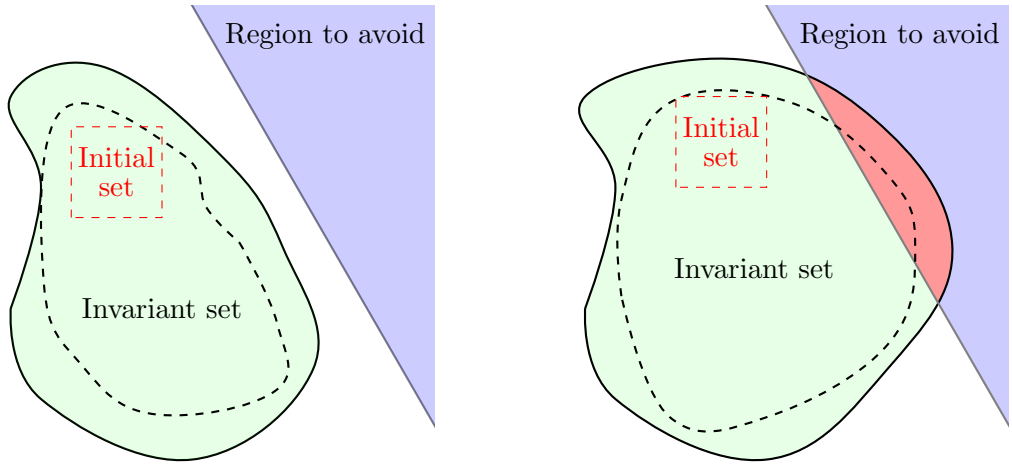
Nous appelons *système commuté* une collection de systèmes qui partagent un même ensemble de paramètres, chacun définissant un processus d'évolution dynamique, et un mécanisme global qui orchestre l'activation de ces systèmes, tel un commutateur.

Les systèmes commutés sont une sous-classe des *systèmes hybrides*. Ces derniers sont également autorisés à modifier les paramètres du système lors de la commutation, en remettant à zéro ou en inversant certaines valeurs par exemple, voir [ACH$^+$95, TD09].

On y inclut la classe des systèmes de contrôle embarqués, qui constituent des systèmes commutés dès lors qu'ils doivent adapter leur comportements, à cause d'un changement d'environnement ou d'une instruction de l'utilisateur. Le thermostat d'une salle soumise à des variations de températures externes est un exemple classique de système hybride: il active le chauffage si la température est plus faible que $21°C$ et le désactive si celle-ci excède $24°C$. Si le mécanisme de commutation est fixé à la même température ($22.5°C$ par exemple), nous obtenons un système commuté avec 2 modes de fonctionnement.

Les systèmes hybrides apparaissent naturellement dans les applications concrètes qui possèdent plusieurs modes de fonctionnement. Ils appartiennent à la classe des *systèmes cyber-physiques*, décrivant des appareils qui sont contrôlés par un logiciel embarqué et qui interagissent avec leur environnement, par des capteurs et/ou des actuateurs, communiquant souvent dans un réseau, dans un "Internet des Objets". Nous nous référons à [SWYS11, HLLL17] ainsi qu'à [AIM10, AFGM$^+$15] pour davantage de contexte.

Les systèmes embarqués critiques sont une sous-classe commune des systèmes cyber-physiques et des systèmes hybrides. Ce sont des systèmes pour lesquels il n'est pas acceptable de dévier de la spécification, par risque de pertes humaines ou matérielles considérables en cas de déviation inattendue. Ces systèmes comprennent les moyens de transport (automobile, avion, train, spatial, etc) ou les appareils médicaux (pacemaker, scanner, etc). L'inclusion de modélisations commutées ou hybrides dans des systèmes critiques a de nombreux avantages: réduction de dépenses énergétiques [NW05], détection et prévention d'erreurs de capteurs en



(a) Invariant sûr (b) Invariant non-conclusif

Figure 1: Certifier l'absence de mauvais comportement: invariants (plein), leur image par le système (pointillé) et la région interdite (bleu)

ligne [Die11]. En échange, le logiciel gagne en complexité et il est plus difficile de contrôler et/ou prévoir le comportement de ces systèmes.

Les systèmes commutés ont été étudiés dans deux domaines en réponse à ces problématiques: en vérification de programmes et en théorie du contrôle.

En vérification de programme, le but est de calculer des certificats de bon comportement, comme des invariants dans l'espace d'état. Les invariants sont des sous-ensembles de l'espace des valeurs prises par les paramètres du systèmes tels que toute évolution future démarrant dans un tel ensemble y reste, voir Figure 1. Il est alors suffisant de montrer qu'un tel invariant ait une intersection vide avec une région dictée comme interdite par la spécification pour prouver la sécurité d'une propriété. Cette approche peut produire des faux-négatifs si l'invariant n'est pas "minimal". En effet, la qualité d'un invariant est très sensible au type de système étudié et à la méthode utilisée pour l'obtenir.

Le calcul d'invariant est d'importance identique en théorie du contrôle, sous deux aspects principaux: le calcul de sur-approximations sûres de l'ensemble des valeurs atteignables par un système hybride [KV97, FLGD$^+$11] et garantir la stabilité asyptotique d'un système dynamique. Dans ce deuxième cas, les objets centraux de l'étude sont les fonctions de Lyapunov, dont les sous-niveaux produisent des ensembles invariants. Les systèmes commutés posent des difficultés en vue de la deuxième problématique, car il peut exister des couplages instables: il ne suffit pas que chaque mode possède un invariant pour que le système commuté complet en possède un.

Les systèmes commutés en vérification de programme

Les systèmes commutés sont généralement présentés sous la forme d'une boucle *while* dans laquelle plusieurs blocs *if-then-else* régissent l'attribution de nouvelles valeurs, et apparaissent classiquement dans des logiciels de contrôle-commande. Nous nous limitons à affectations affines, comme dans Program 1. Le processus de commutation peut dépendre de l'état, i.e. il est gouverné par une garde affine, ou non-déterministe, où tous les modes sont activables à

tout moment.

Program 1: Programme affine commuté avec gardes

$x \leftarrow \mathcal{I}$;
while *true* **do**
  $u \leftarrow \mathcal{U}$;
  if $(x_1 \geqslant 0)$ **then**
    $x \coloneqq A_1 \times x + B_1 \times u + c_1$;
  end
  if $(x_1 < 0)$ **then**
    $x \coloneqq A_2 \times x + B_2 \times u + c_2$;
  end
end

La vérification de programmes affines (commutés ou non) a longtemps utilisé des méthodes polyédrales pour calculer des invariants. Les analyseurs basés sur l'interprétation abstraite [CC77a] ont majoritairement utilisé des domaines polyédraux ou sub-polyédraux (boîtes, octogones, zonotopes) comme dans [Mat07, Gou13].

Les invariants quadratiques (ou ellipsoidaux) ont permis d'obtenir des invariants plus précis pour certains classes de programmes. Ils sont réputés optimaux pour l'étude de systèmes affines non-commutés et apparaissent comme les sous-niveaux de fonction de Lyapunov quadratiques. Ils ont été utilisés dans des applications de contrôle linéaire par Kurzhanski et Vályi dans [KV97] et en vérification de filtres linéaires récursifs par Feret [Fer04], ainsi que localement dans l'analyseur statique Astrée [CCF$^+$05]. D'autres applications en validation de programmes plus générales peuvent être trouvées dans [Cou05, AGG12]. Cette dernière référence développe une approche par "gabarit", basée sur l'idée de gabarit linéaire de Sankaranarayanan et al. [SSM05]. Dans cette approche, la forme de l'ellipsoide est décidée par avance par l'utilisateur. Cette approche est cependant adaptée pour des applications en contrôle: on peut utiliser par exemple comme gabarit les fonctions que les théoriciens ont utilisé pour prouver la stabilité de l'algorithme sous-jacent, comme remarqué par Feron and Alegre dans [FA08a, FA08b]. Récemment, plusieurs méthodes reposant sur la programmation semidéfinie ont été proposées pour synthétiser des invariants ellipsoidaux de systèmes linéaires non-commutés [RJGF12, Rou13, RMF13].

Deux défauts majeurs émergent de ces techniques. Tout d'abord, les domaines numériques ci-présents ne sont pas des treillis: il n'existe pas d'abstraction triviale de l'union de deux ellipsoides pour limiter la complexité de la représentation. De plus, les approches qui reposent sur la solution de problèmes d'optimisation peuvent atteindre un coût calculatoire prohibitif dans des cas de grande dimension ou mal conditionnés.

Stabilité de systèmes linéaires commutés

Il est difficile de décider de la stabilité d'un système commuté, même dans le cas linéaire, où seules les matrices A_i sont non-nulles et le processus de commutation est non-déterministe. On se ramène alors au calcul du rayon spectral joint de la famille de matrices A_i, défini comme

le plus grand taux de croissance des produit de ces matrices:

$$\rho(\mathcal{A}) \coloneqq \lim_{k\to+\infty} \max_{1\leqslant i_1,\dots,i_k\leqslant p} \|A_{i_1}\dots A_{i_k}\|^{1/k} \ .$$

Lorsque $\rho < 1$, toutes les variables du systèmes tendent vers 0 à vitesse géométrique. En revanche, lorsque $\rho > 1$, il existe une manière de choisir à chaque itération un mode qui fait diverger les variables du programme. Ceci indique la présence d'une instabilité non-désirée dans le design ou l'implémentation du programme.

Blondel et Tsitsiklis ont prouvé [BT00] qu'il n'est pas possible de calculer la valeur exacte du rayon spectral joint, ou même de décider s'il est plus petit que 1, car ces problèmes sont indécidables. Plusieurs schémas d'approximation ont donc été développé pour confirmer ou infirmer les propriétés de stabilité. Des approximations inférieures sont obtenues en calculant des produits de matrices de longueur croissante, en combinaison avec un pruning des mauvais candidats [CB11]. L'approche standard pour sur-approximer ρ est d'utiliser la théorie des normes de Barabanov: une norme v telle que $\max_i v(A_i x) = \rho v(x)$ pour tout x. Les implémentations d'une telle approche calculent une norme approximative telle que $v(A_i x) \leqslant \rho' v(x)$, où ρ' est alors un majorant du rayon spectral joint. Ainsi, si $\rho' < 1$, la norme des variables d'état décroit avec un taux géométrique et garantit ainsi la stabilité du système. Il existe plusieurs manières de construire de telles normes: Kozyakin [Koz10] a proposé une approche semi-Lagrangienne, Guglielmi et al. ont proposé un algorithm pour calculer une norme polyédrale [GZ14], i.e. dont la boule unité est un polyèdre. Parrilo et al. ont proposé une approche par sommes de carrés [PJ08]. Enfin, Ahmadi et al. construisent une norme dans [AJPR14] qui est le supremum de (racines de) fonctions quadratiques. Ces méthodes reposent sur la programmation linéaire/semidéfinie et sont lourdes en ressources informatiques.

Si la commutation est gouvernée par une garde linéaire $\sum_i a_i x_i \geqslant 0$, ces approches ne sont plus applicables. La difficulté persiste, car Blondel et Tsitsiklis ont prouvé que la stabilité reste indécidable [BT99, BT00]. Adjé et al. ont proposé une relaxation efficace et des invariants précis sous forme de fonctions quadratiques par morceaux sont obtenus par une combinaison de programmation semidefinie et d'itération sur les politiques, voir [AG15].

Systèmes commutés en contrôle optimal

McEneaney a étudié le problème de contrôle optimal hybride dans lequel un contrôle discret μ permet à l'utilisateur de commuter entre plusieurs modèles linéaires quadratiques:

$$V(x) \coloneqq \sup_{u\in\mathcal{U}} \sup_{\mu\in\mathcal{D}} \sup_{t>0} \int_0^t \frac{1}{2}\xi(s)^T D^{\mu(s)}\xi(s) - \frac{\gamma^2}{2}|u(s)|^2\, ds$$
$$\text{avec} \quad \dot{\xi}(s) = A^\sigma \xi(s) + B^\sigma u(s)\ ,\ \xi(0) = x\,.$$

La fonction valeur V prend des valeurs finies sous plusieurs hypothèses techniques, en particulier si le paramètre γ est suffisamment grand, et est un objet d'intérêt en contrôle H-infini.

La fonction valeur V est également la solution d'une équation aux dérivées partielles (EDP) d'Hamilton-Jacobi [McE07] $H(x, \nabla V) = 0$, avec H un supremum de Hamiltoniens quadratiques:

$$H(x,p) = \max_\sigma \left[(A^\sigma x)^T p + \frac{1}{2}x^T D^\sigma x + \frac{1}{2}p^T \Sigma^\sigma p\right].$$

Ainsi, le calcul de V est lié à la *programmation dynamique*.

La programmation dynamique est une des principales méthodes utilisées pour résoudre des problèmes de contrôle optimal. Elle caractérise la fonction valeur par la solution d'une équation fonctionnelle ou d'une EDP d'Hamilton-Jacobi. Elle fournit une loi de retour qui garantit une optimalité globale. Cependant, elle subit la *malédiction de la dimension*. En effet, les méthodes numériques les plus populaires (schéma monotone aux différences finies, schéma semi-Lagrangien [CD83, CL84, FF94, CFF04] ou les schémas anti-diffusifs [BZ07]) reposent sur des grilles. Ainsi le temps de calcul pour obtenir une solution approximative est exponentiel en la dimension de l'espace d'état.

La méthode de McEneaney [McE07] approxime la fonction valeur par un supremum de fonctions élémentaires comme des formes quadratiques. Elle appartient donc à la famille des "méthodes max-plus" [FM00, AGL08]. Cette méthode atténue la malédiction de la dimension, puisque le temps de calcul ne dépend plus que cubiquement de la dimension, comme le montrent les estimations de complexité de Kluberg et McEneaney [MK10] et de Qu [Qu14]. La méthode de McEneaney a été étudiée et approfondie dans une série de travaux [SGJM10, GMQ11, MD15, KM16]. L'atténuation de la malédiction de la dimension est échangée pour une malédiction de la complexité: en effet, la fonction valeur est représentée par un nombre exponentiel de formes quadratiques au cours du calcul. Plusieurs schémas de pruning ont été proposés pour réduire le nombre de formes quadratiques dans la représentation par des méthodes de programmation semidéfinies. [Qu13]. Cette approche est néanmoins coûteuse: des exemples de dimension 6 ont été résolus en 2 heures, mais 98% du temps de calcul était investi dans la procédure de pruning.

Ordre de Löwner et sélections de majorants minimaux

La programmation semidéfinite est un outil très puissant [BEFB94, BTN01] et consiste en la résolution de problèmes d'optimisation dans le cône des matrices positives semidéfinies. Elle a cependant plusieurs faiblesses: la complexité de la résolution n'est pas encore bien comprise [Ram97] et les approches de points intérieurs ne sont pas adaptées pour des problèmes dont les dimensions excèdent plusieurs milliers de variables (temps de calcul excessif ou problèmes de mémoire). De plus, ces méthodes ne donnent que des solutions approchées et sont sensibles à des instabilités numériques [RVS16].

Une motivation majeure de ce travail est d'obtenir des méthodes de calcul d'invariants de systèmes commutés alternatives qui ne reposent pas sur la programmation semidéfinie afin de diminuer grandement le temps de calcul, en échange d'une petite perte de précision. L'usage intensif de la programmation semidéfinie dans les travaux mentionnés précédemment nous motive aussi à mieux comprendre le cône des matrices positives semidéfinies et les propriétés géométriques sous-jacentes à ces problèmes.

L'ordre induit par ce cône est appelé l'ordre de Löwner et il correspond à l'ordre point par point des formes quadratiques associé à deux matrices symétriques. Un théorème célèbre de Kadison [Kad51] montre que l'espace des matrices symétriques est un *anti-treillis* lorsqu'il est équippé de cette relation d'ordre:

Theorem. *Deux matrices symétriques ont un unique majorant minimal dans l'ordre de Löwner si et seulement si elles sont comparables dans cet ordre.*

Autrement dit, l'opération maximum n'est pas définie, et elle est remplacée par une sélection de majorants minimaux, qui ne sont pas uniques dans les cas non-triviaux.

Savoir sélectionner des majorants minimaux est une étape clé dans beaucoup d'applications et est retrouvée à travers de nombreux domaines appliquées. Cet opération est utilisée en contrôle [LL06] et en information quantique [And99, MG99, DDL06], et elle est plus généralement liée au problème d'inclure un ensemble convexe dans une ellipsoide, qui a des applications en analyse d'atteignabilité de systèmes dynamiques [KV00, KV06] et en vérification de programmes [Fer04]. Des sélections spécifiques ont été utilisées en géométrie de l'information et en morphologie mathématique [Ang13, BBP$^+$07], plus spécifiquement dans le cadre d'images colorées [BK13]. Par exemple, des sélections apparaissant de considération de volume (ellipsoide de Löwner [Bal97]) ou de trace sont fréquemment utilisées. Ces sélections sont généralement obtenues par solution de programmes semidéfinis, même si des formules explicites existent dans certains cas [BK13].

Contributions

Les contributions de cette thèse sont doubles. D'un point de vue appliqué, nous développons dans la Part II plusieurs algorithmes qui calculent des invariants quadratiques, sous la forme d'intersections ou d'unions d'ellipsoids. Plus précisément, nous développons 3 schémas iteratifs, chacun adapté à l'une des classes de systèmes commutés présentés précédemment: programme affine commuté avec gardes, programme linéaire commuté non-déterministe et problème de contrôle optimal linéaire quadratique commuté. La nouveauté réside dans l'absence de résolution de programmes semidéfinis de grande taille, en résolvant soit des programmes de petite taille (dans le cas affine) ou en évitant entièrement ce type de problème (autres cas). Les invariants sont obtenus comme des points fixes d'opérateurs non-monotone par une variation de l'itération de Kleene, ou comme des vecteurs propres non-linéaires par un algorithme power modifié. Ces méthodes alternatives passent beaucoup mieux à l'échelle: en effet, nous pouvons obtenir des approximations du rayon spectral joint de matrices de dimension 500, ce qui est probablement inaccessible aux méthodes de programmation dynamique. Les analyses ci-présentes ne rentrent pas dans le cadre standard de calcul d'invariants, et demandent donc de nouvelles preuves, utilisant des arguments métriques plutôt que de monotonie.

Afin d'établir la convergence de nos algorithmes, nous prouvons dans la Part I des propriétés fondamentales des majorants minimaux dans l'ordre de Löwner et dans des cônes plus généraux. Nous donnons plusieurs caractérisations et paramétrisations, qui nous mènent à des formules algébriques explicites pour calculer de telles sélections. Nous nous concentrons alors sur l'ellipsoide de Löwner. Nous généralisons cette notion à une classe plus grande de cône, et en déduisons une nouvelle inégalité matricielle. Nous obtenons aussi des estimations métriques au sujet de l'application qui à plusieurs ellipsoides fait correspondre l'ellipsoide de Löwner de leur réunion. Ces résultats sont appliqués dans la Part II pour ajuster un paramète de perturbation qui garantit la convergence de nos algorithmes.

La Part I contient les chapitres 1,2,3,4 et traite des propriétés géométriques du cône des matrices positives: sélection de majorants minimaux et propriétés métriques et géométriques de l'ellipsoide de Löwner. La Part II contient les chapitres 5,6,7,8 et développe le cadre d'étude ellipsoidal et son implémentation pour calculer des invariants quadratiques de systèmes commutés en grande dimension. Nous décrivons à présent le contenu de chaque chapitre.

Nous donnons dans le chapitre 1 deux caractérisations de majorants minimaux dans un cône. Nous prouvons qu'ils apparaissent comme des "éléments extrêmes": ce sont exactement

les points "positivement exposés" de l'ensemble des majorants. De plus, nous leur associons des "faces de tangences", qui doivent engendrer le cône complet. Nous illustrons ce théorème avec 3 cônes classiques, dont le cône des matrices positives semidéfinies.

Nous étudions dans le chapitre 2 plus minutieusement les majorants minimaux de deux matrices symétriques dans l'ordre de Löwner. Nous montrons une version quantitative du théorème de Kadison, i.e. que l'ensemble des majorants minimaux de deux matrices A, B est de dimension pq, où (p, q) est l'inertie de la matrice $A-B$, et nous donnons une paramétrisation complète de cet ensemble. Ces résultats sont partiellement étendus au cas de $k \geqslant 3$ matrices. Nous montrons également que l'ensemble des minorants maximaux positifs semidéfinis de deux matrices positive semidéfinies ont une structure similaire, ce qui répond à une question posée par Moreland, Gudder et Ando dans le contexte d'observables quantiques [MG99, And99]. Les preuves impliquent les noyaux des différences $C - A$ et $C - B$, avec C un majorant minimal. Nous inversons le point de vue, en posant la question suivante: étant donnés deux sous-espaces de $\mathbb{R}^n$, existe-t-il des majorants minimaux de A, B pour lesquels les noyaux précédemment définis contiennent ces sous-espaces ? Nous donnons des conditions géométriques simples qui garantissent l'existence de tels majorants minimaux, et donnons une paramétrisation le cas échéant.

Nous généralisons dans le chapitre 3 la notion d'ellipsoide de Löwner à une classe plus générale de cônes. La minimisation de volume est replacée par la minimisation de la fonction caractéristique du cône sur l'ensemble des majorants d'un ensenmble fini $\mathcal{A}$. Cette fonction est la transformée de Laplace de l'indicatrice de l'intérieur du cône dual (dans le cas des matrices positives semidéfinies, elle est égale au déterminant, à une normalisation près). Nous récupérons ainsi l'ellipsoide de Löwner comme un cas particulier. Nous montrons également que cette sélection possède une propriété remarquable: c'est le seul majorant minimal qui se sélectionne lui-même via le processus présenté dans le chapitre 1. De plus, cette sélection commute avec les automorphismes du cône, donc nous la nommons "sélection invariante". Nous utilisons cette nouvelle approche pour prouver une nouvelle inégalité matricielle et nous étudions certaines propriétés supplémentaires de cette sélection.

Nous prouvons dans le chapitre 4 que la sélection invariante est nonexpansive dans la métrique invariante de Riemann, une des métriques principales définies sur l'intérieur du cône [Bha03]. Les métriques de Thompson et de Hilbert [Nus88] sont deux autres métriques importantes. Nous montrons que la constante de Lipschitz de la sélection invariante dans ces métriques croit comme $\log n$. Les preuves de ces résultats reposent sur le caractère Finsler des métriques mentionnées précédemment et sur le calcul de bornes précises de normes de multiplicateurs de Schur, basé sur un résultat de Mathias [Mat93].

Nous développons dans le chapitre 6 un domaine abstrait construit sur les ellipsoides et nous présentons deux algorithmes scalables pour calculer des invariants de systèmes commutés linéaires et affines par des unions d'ellipsoides. Ces deux algorithmes partagent une étape de "partitionement de traces" [MR05] à l'aide d'un automate, qui fait correspondre deux traces qui partagent le même suffixe. Nous reformulons alors le problème de calcul d'invariant en un problème de point fixe (dans le cas affine) ou de calcul d'un vecteur propre non-linéaire (dans le cas linéaire). Nous ajoutons un paramètre de perturbation pour garantir la convergence du schéma, de manière additive ou mutliplicative. Nous comparons notre approche avec les méthodes existantes basées sur les ellipsoides [RG13, AJPR14]. La nature disjonctive de l'approche permet de rafiner la précision de l'invariant selon les souhaits de l'utilisateur, et l'approche par point fixe permet de préserver un calcul rapide, même pour des instances de grande dimension.

Le chapitre 7 introduit la notion d' *application de Kraus tropicale*. Ces applications à valeur d'ensembles sont les analogues des applications de Kraus (application complètement positive préservant la trace) qui apparaissent en information quantique. Nous montrons que les vecteurs propres de ces applications fournissent des invarants de systèmes commutés donnés par des intersections d'ellipsoides. Nous utilisons une itération de type Krasnoselskii-Mann pour calculer ces vecteurs propres de manière scalable, et nous comparons les performances de notre méthode avec des approches alternatives. Nous prouvons que ces vecteurs propres existent sous certains conditions et nous discutons la convergence de l'itération présentée vers un vecteur propre. Nous prouvons que des variantes perturbées de l'itération (de manière additive ou multiplicative) convergent en utilisant des arguments de géométrique métrique.

Le chapite 8 décrit l'implémentation des algorithmes présentés dans les chapitres précédents dans notre outil MEGA (*Minimal Ellipsoids Geometric Analyzer*). Nous décrivons aussi les classes de programmes que l'analyseur peut gérer et illustrons l'outil sur deux exemples.

Le chapitre 5 sert d'introduction à la partie II et le lecteur trouvera dans le chapitre A une courte introduction à la programmation semidéfinie.

Le chapitre 2 est une version adaptée de l'article [Sto16], publié dans les Proceedings de l'AMS. Le chapitre 6 est une combinaison des articles de conférences [AGG$^+$15, AGG$^+$17] d'EMSOFT'15 et d'EMSOFT'17 (des versions plus longues ont été publiées dans ACM TECS). Les idées du chapitre 7 ont été annoncées sans preuves dans le preprint [GS17], et seront publiées dans les proceedings de la conférence CDC'17.

Introduction (english)

This thesis develops methods to compute invariant sets of dynamical systems, motivated by problems in program verification and in control. These methods rely on the geometry of the cone of positive semidefinite matrices and more generally on the geometry of cones. We first review the main applied motivations of this work. We then summarize our results.

Context and motivation

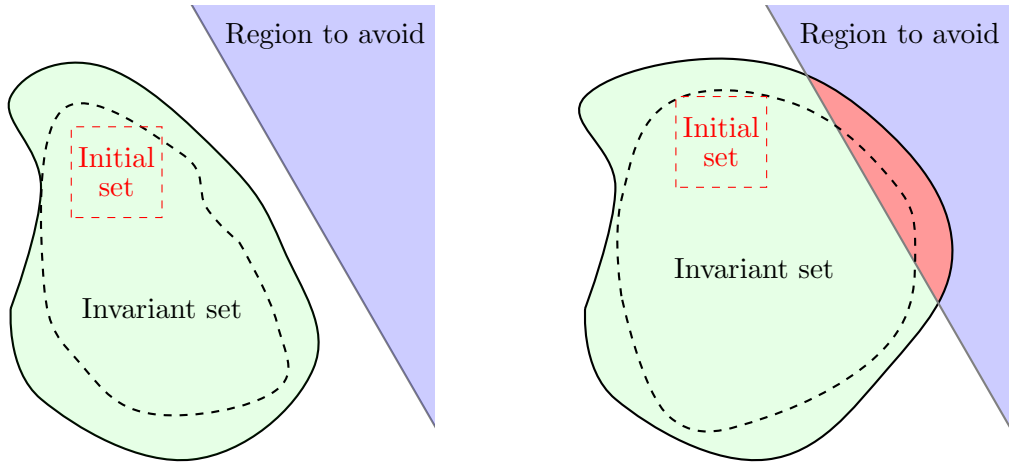
A *switched system* refers to a collection of systems sharing the same parameter pool, each defining a dynamical evolution process, and a global discrete mechanism that orchestrates which system is "active", by switching between them.

Switched systems are a subclass of *hybrid systems*. Instances of the latter class also transform the system's parameters during the (instantaneous) switching, by resetting or reversing values for instance, see [ACH+95, TD09]. This includes the class of on-board controllers, which constitute switched systems as soon as they must adapt their behavior, due to environment changes or user input for instance. A classical hybrid system is the thermostat regulating the temperature of a room subjected to outside-temperature variations, activating the heater if the temperature is below $21°C$ and deactivating it if the temperature exceeds $24°C$. If switching the state of the heater is determined by a same temperature threshold, we obtain a switched system with two modes.

Hybrid systems arise naturally from real-world systems that have several operating modes. They belong to the class of *cyber physical systems*, that describes devices that are controlled by embedded software and interact with their environment, using sensors and actuators, and that communicate within a network, in an Internet of Things. We refer to [SWYS11, HLLL17] as well as [AIM10, AFGM+15] for more background and references therein.

Critical embedded systems constitute a common subclass of cyber physical and hybrid systems and refer to systems for which failure to meet the specification is not acceptable, due to possible harmfulness or high material cost in case of an unexpected deviation. They cover for instance transportation means (car, aircraft, rail, space-vehicles, etc.) or medical devices (pacemaker, scanners, etc.). The inclusion of switched or hybrid modelizations on critical systems has had many benefits like reduced energy expenses [NW05] or on-the-fly sensor error detection and prevention [Die11]. The trade-off is an increased complexity of the software and a reduced capacity to control and predict the behavior of these systems.

Switched systems have been studied in two fields to answer these concerns: in program verification and in control.



(a) Successful invariant certificate (b) Non-conclusive invariant

Figure 2: Certifying absence of bad behavior: invariants (plain), their image by one transformation step (dashed) and bad region (blue)

In program verification, the goal is to compute certificates of good behavior, such as invariants within the state-space. Invariants are subsets of the set of values taken by the system's parameters that contain all possible future evolutions of the system starting from any point in this subset, see Figure 2. It is sufficient to show that the invariant and the forbidden region determined by the specification do not overlap to obtain the safety of the system. This method may produce false negatives if the invariant is not tight. Indeed, the quality of an invariant is highly dependent on the type of system and the method that produced it.

The computation of invariants is equally important in control theory, with mainly two considerations: finding safe approximations of the reachable set of a hybrid system [KV97, FLGD$^+$11] and guaranteeing the asymptotic stability of a dynamical system. Fundamental objects in the analysis of the latter case are Lyapunov functions, sub-level sets of which constitute invariant sets. Switched systems are challenging with respect to the second problem, for there is the possibility of *unstable couplings*: it is not sufficient for each mode to have an invariant for the whole switched system to have an invariant.

Switched systems in program verification

Switched programs are usually presented as a while loop in which several if-then-else statements dictate how the internal state variables are assigned new values, and appear classically in control-command software. We restrict our attention to affine assignments, like in Program 2. The switching process can be state-dependent, i.e., it is governed by an affine guard, or non-deterministic, in which case any system can be activated at any time.

Program verification of switched (or unswitched) affine programs has long relied on polyhedral methods in order to compute invariants, i.e. subsets of the state space that remain invariant by every branch of the program. Analyzers based on abstract interpretation [CC77a] have mostly been using polyhedral or sub-polyhedral domains (boxes, octagons, zonotopes) as in e.g. [Mat07, Gou13].

Program 2: Switched affine program with guards

```
x ← I;
while true do
    u ← U;
    if (x_1 ⩾ 0) then
        x := A_1 × x + B_1 × u + c_1;
    end
    if (x_1 < 0) then
        x := A_2 × x + B_2 × u + c_2;
    end
end
```

Ellipsoidal (or quadratic) invariants have led to more accurate analyses for some classes of programs. They are known to be the optimal tool for studying the stability of linear (unswitched) systems as the sub-level sets of quadratic Lyapunov functions. They have been used in linear control applications by Kurzhanski and Vályi in [KV97] and in program verification of linear recursive filters by Feret [Fer04], they are also used locally in the static analyzer Astrée [CCF+05]. More general applications of ellipsoids in program validation can be found in [Cou05, AGG12]. The latter reference develops a template approach, based on the linear template original idea of Sankaranarayanan et al. [SSM05]. In template based methods, the shape of the ellipsoid has to be decided in advance by the user. Still, this is an approach which is adapted to control codes: we may use as (quadratic) templates the (quadratic) Lyapunov functions that the control theorist would have introduced to prove the stability of the underlying algorithms, as put forward by Feron and Alegre in [FA08a, FA08b]. Recently, some methods have been proposed in program validation for synthesizing invariant ellipsoids for linear (unswitched) systems e.g. [RJGF12, Rou13, RMF13], using semidefinite programming.

Two main drawbacks emerge from these techniques. First, the numerical domains dealt with here do not constitute lattices, hence there is no immediate way to abstract the union of two ellipsoids and to limit the complexity of the representation due to memory limitations. Moreover, approaches that rely on the solution of optimization problems can reach a prohibitive computational cost on instances that have high dimension or that are badly conditioned.

Stability of switched linear systems

Deciding the stability of a switched program is a difficult problem, even in the linear case, where only the matrices A_i are non-zero and the switching process is non-deterministic. It reduces to the problem of computing the joint spectral radius of the matrices A_i, defined as the largest growth rate of products of the latter matrices by

$$\rho(\mathcal{A}) \coloneqq \lim_{k\to+\infty} \max_{1\leqslant i_1,\ldots,i_k\leqslant p} \|A_{i_1}\ldots A_{i_k}\|^{1/k} \quad .$$

When $\rho < 1$, all variables of the program decay towards zero at a geometric rate. However, $\rho > 1$ means that there is a sequence of branch choices at each iteration that makes the program variables diverge. The latter indicates the presence of an undesired instability

in the design or the implementation of the program. Unfortunately it is not possible to compute exactly the joint spectral radius, or even decide easily if it is less than 1, as Blondel and Tsitsiklis have shown in [BT00] that these problems are undecidable. Instead, many approximation schemes have been proposed that provide either lower or upper bounds on ρ, in order to try to infirm or confirm stability properties. Accurate lower bounds have historically been obtained by computing products of matrices of increasing length, combined with a pruning of bad candidates [CB11]. The standard approach to over-approximate ρ is to use the theory of Barabanov norms: a norm v such that $\max_i v(A_i x) = \rho v(x)$ for all i and x. Implementations of this approach compute approximate norms such that $v(A_i x) \leqslant \rho' v(x)$, where ρ' provides an upper bound on the joint spectral radius. Hence, if $\rho' < 1$, the norm of the state variables decreases geometrically and guarantees the stability of the system. Several ways to construct such a norm have been studied: Kozyakin [Koz10] has proposed a grid-based, semi-Lagrangian approach, Guglielmi et al. have proposed in [GZ14] algorithms to produce polyhedral norms, i.e. whose unit ball is a polyhedron. Parrilo et al. have proposed a sums-of-square approach in [PJ08]. Finally, Ahmadi et al. build a norm in [AJPR14] that is the supremum of (the square root of) quadratic functions. These methods rely on linear programming or semidefinite programming and are highly computationally demanding tasks.

If the switching is governed by a linear guard of the form $\sum_i a_i x_i \geqslant 0$, these approaches are no longer valid. However, the difficulty persists, as Blondel and Tsitsiklis have shown that adding a guard does not ease the decidability of the stability [BT99, BT00]. An efficient relaxation of such systems has been studied by Adjé et al. and relies on the combination of semidefinite programming and policy iteration to compute concise invariants as piecewise quadratic sets, see [AG15].

Switched systems in optimal control

McEneaney considered hybrid optimal control problems in which a discrete control μ allows one to switch between different linear quadratic models:

$$V(x) \coloneqq \sup_{u\in\mathcal{U}} \sup_{\mu\in\mathcal{D}} \sup_{t>0} \int_0^t \frac{1}{2}\xi(s)^T D^{\mu(s)}\xi(s) - \frac{\gamma^2}{2}|u(s)|^2\, ds$$
$$\text{subject to} \quad \dot{\xi}(s) = A^\sigma \xi(s) + B^\sigma u(s)\ ,\ \xi(0) = x\,.$$

The value function V, also called *storage function*, takes finite values under some technical assumptions, including a condition that the penalizing parameter γ is large enough. This is a feature of interest in H-infinity control.

It is known [McE07] that the value function V is the solution of a Hamilton-Jacobi partial differential equation (PDE) $H(x, \nabla V) = 0$ where H is a supremum of quadratic Hamiltonians

$$H(x,p) = \max_\sigma \left[(A^\sigma x)^T p + \frac{1}{2}x^T D^\sigma x + \frac{1}{2}p^T \Sigma^\sigma p\right],$$

and thus relates to the field of *dynamic programming*.

Dynamic programming is one of the main methods to solve optimal control problems. It characterizes the value function as the solution of a functional equation or of a Hamilton-Jacobi partial differential equation. It provides a feedback law that is guaranteed to be globally optimal. However, it is subject to the "curse of dimensionality". Indeed, the main numerical methods, including monotone finite difference or semi-Lagrangian schemes [CD83,

CL84, FF94, CFF04], and the anti-diffusive schemes [BZ07], are grid-based. It follows that the time needed to obtain an approximate solution with a given accuracy is exponential in the dimension of the state space.

The method he developed [McE07] approximates the value function by a supremum of elementary functions like quadratic forms, hence it belongs to the family of "max-plus basis methods" [FM00, AGL08]. The method of [McE07] has a remarkable feature: it attenuates the curse of dimensionality to a cubic cost in the dimension, as shown by the complexity estimates of Kluberg and McEneaney [MK10] and of Qu [Qu14]. McEaneney's method [McE07] has been studied and extended in a series of works [SGJM10, GMQ11, MD15, KM16]. This reduction in dimensionality is traded for a "curse of complexity". It represents the value function as a supremum of quadratic forms that are accumulated at great rate along the computation. It is thus necessary to control the number of quadratic forms. Several pruning schemes to remove inefficient quadratic forms have been successful in dealing with this growth, making use of semidefinite programs [Qu13]. This approach is costly: examples of PDE in dimension 6 have been solved in 2 hours, but the pruning procedure takes more than 98% of the total computation time.

Löwner order and upper bound selections

Semidefinite programming is a powerful tool [BEFB94, BTN01] and consists in solving optimization problems in the cone of positive semidefinite matrices. However, it has several weaknesses: the complexity of semidefinite programs is not yet well understood [Ram97] and resolution methods based on interior points approaches are not well suited for problems whose dimension exceeds thousands of variables (excessive computation time or lack of memory issues). Moreover, these methods only return approximate solutions and they are prone to numerical instabilities [RVS16].

A core motivation of this work is to obtain alternative methods to compute invariants of switched systems that are not based on semidefinite programming in order to greatly reduce the computation time, maybe at the expense of a small loss of precision. The intensive use of semidefinite programming in the works mentioned earlier also motivates the need for a clearer understanding of the cone of positive semidefinite matrices and the underlying geometric properties in these problems.

The ordering induced by this cone is called the Löwner order and it corresponds to the point-wise ordering of the quadratic forms associated with symmetric matrices. A famous theorem by Kadison [Kad51] states that the space of symmetric matrices equipped with the Löwner order is an *anti-lattice*:

Theorem. *Two symmetric matrices have a unique minimal upper bound in the Löwner order if and only if they are comparable.*

In other words, there is no maximum operation, and it must be substituted by the selection of a minimal upper bound, which is never unique in non-trivial cases.

Finding effective selections of these minimal upper bounds is a key ingredient in many applications and appears in a number of applied fields. It is used in control [LL06] and quantum information [And99, MG99, DDL06], and it is more generally related to the problem of enclosing a convex set by an ellipsoid, which has applications in reachability analysis of dynamical systems [KV00, KV06] and program verification [Fer04]. Specific selections have been used in information geometry and mathematical morphology [Ang13, BBP$^+$07],

in particular in the setting of colored images [BK13]. For instance, selections arising from minimum volume considerations (Löwner's ellipsoid [Bal97]) are frequently used, as well as selections based on minimum-(or maximum-)trace of the associated matrix. These selections are usually computed by semidefinite programming, although some explicit formulas exist as generalizations of the scalar case [BK13].

Contributions

The contributions of this thesis are two-fold. From an applied perspective, we develop in Part II several algorithms that compute quadratic invariants, either as the union or the intersection of ellipsoids. More precisely, we develop three iterative schemes, each one adapted to the three classes of switched systems presented earlier: switched affine programs with guards, switched linear programs with non-deterministic switching and hybrid linear-quadratic optimal control problems with switches. The novelty is that we avoid the recourse to large-scale semidefinite programs, by computing either solutions to several small-size problems (when dealing with affine programs) or without solving any semidefinite programs (in the other cases). Instead, invariants are obtained in a scalable way as post-fixed points of non-monotone maps via a variation on Kleene iteration, or as eigenvectors of non-linear maps by a power-like algorithm. In this way, we can obtain approximate solutions for instances of large dimension (for instance dimensions up to 500 for the joint spectral radius problem), probably inaccessible by dynamic programming-type approaches. These analyzes fall beside the classical framework of invariant computation and thus require different proofs, exploiting metric arguments instead of monotonicity.

In order to obtain the convergence of our algorithms, we establish in Part I fundamental properties of minimal upper bound selections in the Löwner order and in more general cones. We give several characterizations and parametrizations, which lead to explicit algebraic formulas to compute such selections. We then focus our attention on the Löwner ellipsoid. We generalize this notion to a larger class of cones and deduce a new matrix inequality. We also obtain metric properties on the map sending several ellipsoids to their Löwner ellipsoid. These results are applied in Part II to fine-tune perturbation parameter that ensure the convergence of our algorithms.

Part I is comprised of Chapters 1 to 4 and deals with the geometric properties of the cone of positive semidefinite matrices: selections of minimal upper bounds and geometric/metric properties of the Löwner ellipsoid. Part II contains Chapters 5 to 8 and develops an ellipsoidal framework and its implementation to compute invariants of switched systems in a tractable manner. We next describe the content of each chapter.

In Chapter 1, we give two geometric characterizations of minimal upper bounds in a cone. This result shows that they arise as "extreme elements". On the one hand, they correspond to the set of "positively exposed" points of the set of upper bounds. On the other hand, it relates to "tangency spaces", that must span the whole space. We illustrate this theorem on three classical cones, including the cone of positive semidefinite matrices.

In Chapter 2, we study in more detail minimal upper bounds of two symmetric matrices in the Löwner order. This entails a quantitative version of Kadison's theorem, showing that the set of minimal upper bounds of two matrices A, B has dimension pq, where (p, q) is the inertia of the matrix $A - B$, and a complete parametrization of this set. These results are partially extended to the case of $p \geqslant 2$ matrices. We also show that the set of positive semidefinite

maximal lower bounds of two positive semidefinite matrices has a similar structure and we provide a parametrization of this set. This solves a question raised by Moreland, Gudder and Ando in the setting of quantum observables [MG99, And99]. The proof involves the kernels of the differences $C - A$ and $C - B$ where C is a minimal upper bound. We reverse the point of view by determining whether, given two subspaces of $\mathbb{R}^n$, there are minimal upper bounds of A, B for which the former kernels contain these respective subspaces. We provide simple geometric conditions that guarantee the existence of such minimal upper bound and provide a parametrization of all minimal upper bounds that arise in this way.

In Chapter 3, we generalize the Löwner ellipsoid to cones other than the cone of positive semidefinite matrices. Instead of minimizing the volume of an ellipsoid, we minimize the characteristic function of the cone over the upper bounds of a finite set $\mathcal{A}$. The function is nothing but the Laplace transform of the indicator function of the dual cone (in the case of positive definite matrices, it coincides with the determinant, up to renormalization). In this way, the Löwner ellipsoid is recovered as a special case. We also show that this selection has a remarkable property: it is the unique minimal upper bound of $\mathcal{A}$ that selects itself via the selection process from Chapter 1. Moreover, this selection is the only one that enjoys a useful invariance property, hence it is christened the "invariant selection". We use this new approach to show a new matrix inequality and study properties of this selection.

In Chapter 4, we show that the invariant selection is non-expansive in the Riemann metric, which is one of the main metrics on the interior of the cone [Bha03]. Other important metrics are the Thompson and Hilbert metrics [Nus88]. We also show that the Lipschitz constant of the invariant selection in these metrics grows asymptotically like $\log n$. The proofs rely on the Finsler nature of the aforementioned metrics and the computation of accurate bounds on Schur multiplier norms, building on a result by Mathias [Mat93].

In Chapter 6, we develop a numerical abstract domain based on ellipsoids and present two scalable algorithms to compute invariants of switched affine and switched linear programs as unions of ellipsoids. A common feature in these algorithms is a "trace partitioning" step [MR05] with an automaton, that identifies traces that share a common suffix. We then reduce to the solution of a fixed-point problem (affine case) or a non-linear eigenvalue problem (linear case). A perturbation parameter is introduced to ensure convergence of the scheme, either in an additive or a multiplicative way. We compare our method with existing methods [RG13, AJPR14] based on ellipsoidal invariants. The disjunctive nature of the invariant allows one to refine the precision while the fixed-point approach preserves scalability.

In Chapter 7, we introduce set-valued *tropical Kraus maps* which are analogous of ordinary Kraus maps, trace-preserving completely positive linear maps arising in quantum information. We show that non-linear eigenvectors of these maps provide invariants of switched systems as intersection of ellipsoids. We present a scalable Krasnoselkii-Mann-like iteration to compute non-linear eigenvectors and compare the performance of our method with alternative approaches. We prove that these eigenvectors exist (under some assumptions) and discuss the convergence of the iteration towards these eigenvectors. We also show that the additive and multiplicative iterations defined in Chapter 6 do converge by exploiting metric geometry techniques.

Chapter 8 presents the implementation of the algorithms developed in earlier chapters in our tool "MEGA" (for *Minimal Ellipsoids Geometric Analyzer*). We describe the class of programs that the analyzer can handle and illustrate the tool on two benchmarks.

On top of the chapters already presented, Chapter 5 is an introductory chapter for Part II and the reader will find in Appendix A a short introduction to semidefinite programming

provided for the sake of completeness.

Chapter 2 is an adapted version of the article [Sto16] published in the Proceedings of the AMS. Chapter 6 is a combination of the conference articles [AGG$^+$15, AGG$^+$17] for EMSOFT'15 (an extended version [AGS$^+$16] has been published in ACM TECS) and for EMSOFT'17. The ideas in Chapter 7 have been announced without proofs in the preprint [GS17], to be published in the CDC conference proceedings.

Part I

Minimal upper bounds in cones
The case of the Löwner order

CHAPTER 1

Characterization of minimal upper bounds in cones

We prove in this chapter several characterizations of minimal upper bounds with respect to the order relation induced by a cone in a finite dimensional real vector space. Our main result (Theorem 1.10) states that minimal upper bounds coincide with positively exposed elements of the set of upper bounds, i.e. they are the minimizers of a strictly monotone linear function on this space. It also states that minimal upper bounds must satisfy a sufficient number tangency conditions, related to a decomposition of the dual cone as the sum of tangency sub-faces. These results are illustrated on three classical cones: the class of polyhedral cones, the Lorentz cone associated with a strictly convex norm (in particular the Euclidean Lorentz cone) and the cone of positive semidefinite matrices.

1.1 Cones, duality, order relations and faces

1.1.1 Cones and dual cones

Let E denote a n-dimensional vector space equipped with the scalar product $\langle \cdot, \cdot \rangle$. A set $\mathcal{C} \subset E$ is called a *cone* if $x \in \mathcal{C}$ and $\lambda > 0$ imply $\lambda x \in \mathcal{C}$. A set $\mathcal{X} \subset E$ is *convex* if $x, y \in \mathcal{X}$ and $0 \leqslant \lambda \leqslant 1$ imply $\lambda x + (1-\lambda)y \in \mathcal{X}$. A *convex cone* is then a set $\mathcal{C}$ such that $x, y \in \mathcal{C}$ and $\lambda, \mu \geqslant 0$ imply $\lambda x + \mu y \in \mathcal{C}$. We say that the convex cone $\mathcal{C}$ is *pointed* if it does not contain any linear subspace, i.e. that $\mathcal{C} \cap -\mathcal{C} \subseteq \{0\}$. In the sequel, unless stated specifically otherwise, a cone will refer to a *closed convex pointed cone*. In particular, $0 \in \mathcal{C}$.

The set $\operatorname{span} \mathcal{C} \coloneqq \mathcal{C} - \mathcal{C} = \{x - y \colon x, y \in \mathcal{C}\}$ is the smallest subspace of E that contains $\mathcal{C}$. It is immediately seen that the cone $\mathcal{C}$ has non-empty interior if and only if $\mathcal{C} - \mathcal{C} = E$.

The *dual cone* of any subset $\mathcal{X}$ of E is denoted by $\mathcal{X}^\star$ and is defined by

$$\mathcal{X}^\star \coloneqq \{y \in E \colon \langle y, x\rangle \geqslant 0 \text{ for all } x \in \mathcal{X}\}\,.$$

The dual cone of the cone $\mathcal{C}$ is pointed if and only if the cone $\mathcal{C}$ has non-empty interior, and we have $(\mathcal{C}^\star)^\star = \mathcal{C}$. A vector c belongs to the interior of the dual cone $\mathcal{C}^\star$ if $\langle c, x\rangle > 0$ for all non-zero $x \in \mathcal{C}$.

1.1.2 Classical cones

We describe in the following five types of cones, which will serve as running examples in this chapter: polyhedral cones, the non-negative orthant $(\mathbb{R}_+)^n$ as a special case of a polyhedral cone, the (generalized) Lorentz cones $\mathcal{L}_n$, the classical Euclidean Lorentz cone Λ_n and the cone of positive semidefinite matrices $\mathcal{S}_n^+$.

Polyhedral cone A *polyhedral cone* is a set of the form

$$\mathcal{C}_P \coloneqq \{x \in \mathbb{R}^n \colon (Px)_i \geqslant 0\ ,\ \forall i \in I\},$$

where P is a $m \times n$ matrix and $I = \{1, \dots, m\}$. It is readily checked that such a set is indeed a closed convex cone. In the sequel, we denote by p_i the rows of the matrix P: $P = \left(p_1^T \cdots p_m^T\right)^T$, so that $(Px)_i \geqslant 0 \iff \langle p_i, x\rangle \geqslant 0$.

The dual cone of the polyhedral cone $\mathcal{C}_P$ is the conic hull of the vectors $\{p_i\}_i$:

$$y \in \mathcal{C}_P^\star \iff \exists \{\lambda_i\}_{1 \leqslant i \leqslant m} \subset \mathbb{R}_+ \colon y = \sum_i \lambda_i p_i\,.$$

Indeed, recall that $x \in \mathcal{C}_P$ if and only if the vector Px has non-negative coordinates, i.e. $\langle \lambda, Px\rangle$ is non-negative for all vectors λ with non-negative coordinates. It follows that $y \in \mathcal{C}_P^\star$ if and only if $y = P^T\lambda = \sum_i \lambda_i p_i$ with $\lambda_i \geqslant 0$.

A cone is *finitely generated* if it can be written as the conic hull of finitely many vectors. Finitely generated cones are polyhedral cones, as asserted by the Minkowski-Weyl theorem, so the dual cone of a polyhedral cone is again a polyhedral cone.

Theorem 1.1 (Minkowski-Weyl, see [Min97, Wey35, Zie95])**.** *A cone is finitely generated if and only if it is polyhedral.*

The polyhedral cone generated by the vectors $\{q_i\}_i$ is pointed if and only if the vectors q_i belong to a same open half-space, i.e. there is some vector $d \in \mathbb{R}^n$ such that $\langle d, q_i\rangle > 0$ for all i. It has non-empty interior as soon as the matrix whose columns are the vectors q_i has full rank.

Finally, given a vector $b \in \mathbb{R}^m$, the set $\{x \in \mathbb{R}^n \colon (Px)_i \geqslant b_i\ ,\ \forall i \in I\}$ is called a polyhedron, see [Wil93] for more background.

We show in Figure 1.1 an instance of a polyhedral cone in $\mathbb{R}^3$ with the intersection of the cone with an affine hyperplane of the form $\{x \in \mathbb{R}^3 \colon x_3 = \text{const}\}$.

Non-negative orthant $(\mathbb{R}_+)^n$ *The non-negative orthant* is the polyhedral cone that is generated with the canonical base $(e_i)_i$ of $\mathbb{R}^n$:

$$(\mathbb{R}_+)^n = \{x \in \mathbb{R}^n \colon x_i \geqslant 0\,,\ \forall i\}\,.$$

A vector $x \in \mathbb{R}^n$ belongs to the interior of this cone if and only if every coordinate is positive. This cone is self-dual: $\left((\mathbb{R}_+)^n\right)^\star = (\mathbb{R}_+)^n$.

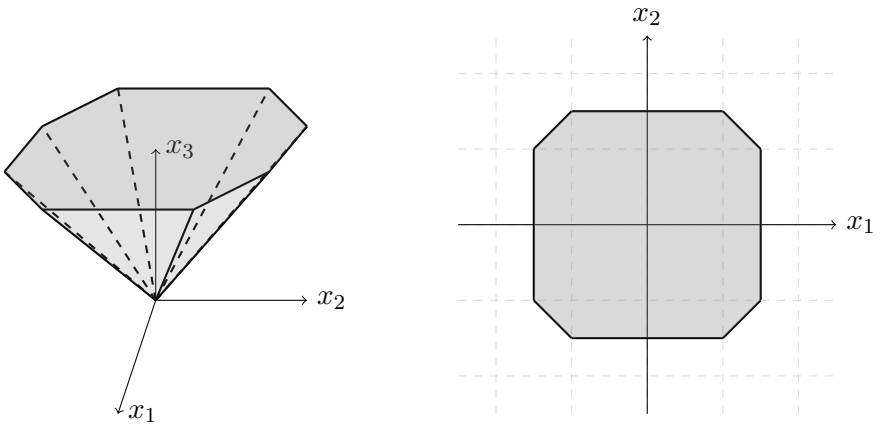


Figure 1.1: A polyhedral cone in $\mathbb{R}^3$ and its section in an $x_1 - x_2$ plane

Lorentz cone $\mathcal{L}_n$ Given any norm $\|\cdot\|$ on $\mathbb{R}^n$, the associated *(generalized) Lorentz cone* is the subset of $\mathbb{R}^{n+1}$ defined by

$$\mathcal{L}_n \coloneqq \{(t, x) \in \mathbb{R}^{n+1} \colon t \geqslant \|x\|\}\,.$$

This set is a cone since it is closed under multiplication by a positive scalar, and under addition by the triangular inequality: if $(s, x), (t, y) \in \mathcal{L}_n$, then $t + s \geqslant \|x\| + \|y\| \geqslant \|x + y\|$, thus $(s + t, x + y) \in \mathcal{L}_n$. This cone is pointed since $(t, x) \in \mathcal{L}_n \cap -\mathcal{L}_n$ if and only if $t = 0$, which in turn implies that $x = 0$. The vector (t, x) belongs to the interior of the Lorentz cone if and only if $t > \|x\|$. The boundary of the cone $\partial\mathcal{L}_n$ is hence constituted by vectors (t, x) such that $t = \|x\|$.

We are especially interested in Lorentz cones arising from strictly convex norms, meaning that the associated unit ball is strictly convex:

$$\|x\| \leqslant 1 \text{ and } \|y\| \leqslant 1 \implies \|\lambda x + (1 - \lambda)y\| < 1,\ \forall \lambda \colon 0 < \lambda < 1\,.$$

We show such a Lorentz cone associated with a strictly convex norm in Figure 1.2.

The dual cone of the Lorentz cone associated with the norm $\|\cdot\|$ is the Lorentz cone associated with the dual norm $\|\cdot\|^\star$ defined by

$$\|y\|^\star = \sup_{\|x\| \leqslant 1} \langle x, y\rangle\,.$$

In particular, the Lorentz cones associated with the p-norm and the q-norm are duals of one another if $p^{-1} + q^{-1} = 1$ $(1 \leqslant p, q \leqslant \infty)$.

The Euclidean Lorentz cone A classical example of a Lorentz cone associated with a strictly convex norm is the *Euclidean Lorentz cone*, or light cone, denoted by Λ_n, corresponding to the Euclidean norm $\|\cdot\|_2$. Then, we can also write

$$\Lambda_n = \{z \in \mathbb{R}^{n+1} \colon \langle e_1, z\rangle \geqslant 0 \quad \text{and} \quad z^T J_{1,n} z \geqslant 0\} \quad \text{with} \quad J_{1,n} \coloneqq \begin{pmatrix} 1 & \\ & -I_n \end{pmatrix}.$$

This cone is also self-dual. We point out that the Euclidean Lorentz cone corresponds to the set of vectors $z \in \mathbb{R}^{n+1}$ such that $(\sum z_i^2)^{1/2} \leqslant \langle \sqrt{2} e_1, z\rangle$.

We show the Euclidean Lorentz cone in $\mathbb{R}^3$ in Figure 1.3.

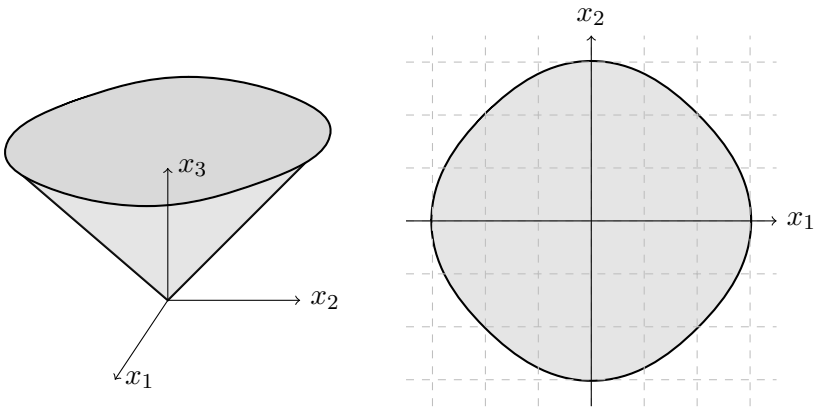


Figure 1.2: A generalized Lorentz cone associated with a strictly convex norm and its section in an $x_1 - x_2$ plane

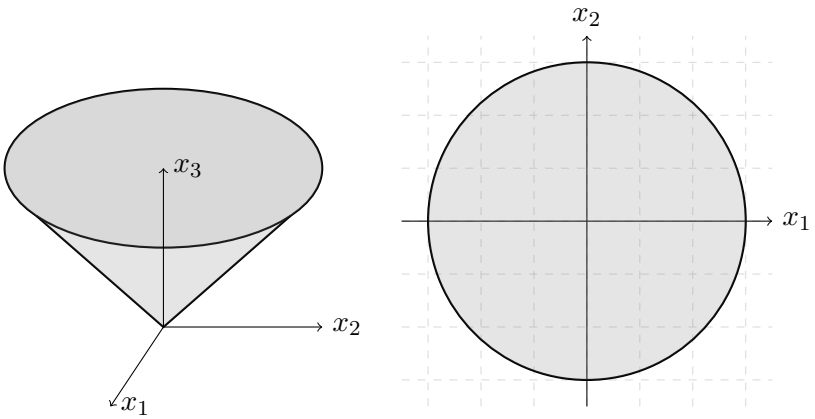


Figure 1.3: The Euclidean Lorentz cone in $\mathbb{R}^3$.

The cone of positive semidefinite matrices $\mathcal{S}_n^+$ A real $n \times n$ matrix A is *symmetric* if $A_{ij} = A_{ji}$ for all $1 \leqslant i, j \leqslant n$. The set of symmetric matrices is denoted by $\mathcal{S}_n$. A symmetric matrix A is *positive semidefinite* if $x^T Ax \geqslant 0$ for all vector $x \in \mathbb{R}^n$ (we identify the set of vectors in $\mathbb{R}^n$ with the set of $n \times 1$ real matrices). When $x^T Ax > 0$ for all non-zero vector $x \in \mathbb{R}^n$, we say that A is positive definite. The set of positive semidefinite matrices is denoted by $\mathcal{S}_n^+$ and the subset of positive definite matrices by $\mathcal{S}_n^{++}$.

It is known that the matrix A is positive semidefinite (resp. definite) if its (real) eigenvalues are non-negative (resp. positive). Moreover, the matrix A is positive semidefinite (resp. definite) if there is a matrix (resp. invertible matrix) M such that $A = MM^T$. In particular, a rank one positive semidefinite matrix is written xx^T for some $x \in \mathbb{R}^n$.

It can be readily checked that the set $\mathcal{S}_n^+$ is a closed convex pointed cone. Moreover, it is self-dual: if A is an element of the dual cone $(\mathcal{S}_n^+)^\star$, the inequality $\langle A, X \rangle \geqslant 0$ must hold in particular for all rank one matrices X, hence $x^T Ax \geqslant 0$ for all $x \in \mathbb{R}^n$ and A is positive semidefinite. The set $\mathcal{S}_n^{++}$ is also a cone and constitutes the interior of the cone $\mathcal{S}_n^+$ in the Euclidean topology.

We point out that the Euclidean Lorentz cone Λ_2 in $\mathbb{R}^3$ is isomorphic to the cone $\mathcal{S}_2^+$ by

mapping the matrix $\begin{pmatrix} t-x & y \\ y & t+x \end{pmatrix}$ to the vector (t,x,y). Indeed, the former 2×2 matrix is positive semidefinite if and only if its trace and determinant are non-negative. In this case, these conditions amount to $t \geqslant 0$ and $t^2 \geqslant x^2+y^2$, i.e. $(t,x,y) \in \Lambda_2$.

1.1.3 Order relation induced by a cone

We recall several definitions and results on order relations induced by closed convex pointed cones.

Given any (closed convex pointed) cone $\mathcal{C} \subset E$, we can equip the space E with an order relation, defined by

$$x \preccurlyeq y \iff y - x \in \mathcal{C}\,.$$

This relation is reflexive since $0 \in \mathcal{C}$. It is transitive because the cone is convex and thus closed under addition. Finally, the fact that $\mathcal{C}$ is pointed implies that the $y-x \in \mathcal{C}$ and $x-y \in \mathcal{C}$ if and only if $x=y$, thus it is antisymmetric.

In this order, an *upper bound* of a set $\mathcal{A} \subset E$ is any element $x \in E$ such that $a \preccurlyeq x$, for all $a \in \mathcal{A}$. We write $\mathcal{A} \preccurlyeq x$. A *minimal upper bound* is an element $y \in E$ such that $\mathcal{A} \preccurlyeq x \preccurlyeq y$ implies $x = y$. Similarly, a *lower bound* of $\mathcal{A}$ is any element $y \in E$ such that $y \preccurlyeq a$ for all $a \in \mathcal{A}$, and a *maximal lower bound* y satisfies $y \preccurlyeq x \preccurlyeq \mathcal{A}$ implies $x = y$ for all x. We denote by $\bigvee \mathcal{A}$ the set of minimal upper bounds of the set $\mathcal{A}$ and by $\bigwedge \mathcal{A}$ the set of maximal lower bounds of $\mathcal{A}$.

We point out that, given a vector $c \in \operatorname{int} \mathcal{C}^\star$, the map $x \mapsto \langle c, x\rangle$ is strictly monotone, meaning that $x \preccurlyeq y$ and $x \neq y$ implies $\langle c,x\rangle < \langle c,y\rangle$.

We say that the space $(E,\mathcal{C})$ is a *lattice* if every pair $\{a,b\} \subset E$ has a unique maximal lower bound and a unique minimal upper bound. In that case, the sets $\bigvee\{a,b\}$ and $\bigwedge\{a,b\}$ are reduced to a single point, respectively called the *supremum* and *infimum* of a and b. Not every cone $\mathcal{C}$ endows the space E with a lattice structure. In fact, up to an invertible linear transformation, the non-negative orthant $(\mathbb{R}_+)^n$ is the only one with this property, as shown by a theorem by Krein and Rutman:

Theorem 1.2 (Krein and Rutman (1948), see [KR48]). *The space $(E,\mathcal{C})$ is a lattice if and only if the cone $\mathcal{C}$ is simplicial, i.e. there is a basis $(q_i)_{1\leqslant i\leqslant n}$ of E such that*

$$x \in \mathcal{C} \iff \exists \lambda_i \geqslant 0 \colon x = \sum_i \lambda_i q_i\,.$$

On the other side of the spectrum, some spaces $(E,\mathcal{C})$ are "as far away from a lattice" as possible, in the sense that two elements a,b only have a unique minimal upper bound and a unique maximal lower bound when $a \preccurlyeq b$ or $b \preccurlyeq a$ (we also say that a and b are *comparable*). In this case, we say that $(E,\mathcal{C})$ is an *anti-lattice*.

We recall a famous theorem by Kadison [Kad51] that shows that the cone of positive semidefinite matrices $\mathcal{S}_n^+$ is an anti-lattice:

Theorem 1.3 (Kadison (1951), see [Kad51]). *Two symmetric matrices A,B have a unique minimal upper bound in the Löwner order if and only if $A \preccurlyeq B$ or $B \preccurlyeq A$.*

In the case $\mathcal{C} = \mathcal{S}_n^+$, the order relation $\preccurlyeq$ is called the Löwner order.

1.1.4 The faces and extreme rays of a cone

Given a convex set $\mathcal{X}$, an element $x \in \mathcal{X}$ is called an *extreme point* if for all $x_1, x_2 \in \mathcal{X}$, the equality $x = \frac{1}{2}(x_1 + x_2)$ implies $x_1 = x_2 = x$. More generally, given two convex sets $\mathcal{X}, \mathcal{Y}$ such that $\mathcal{Y} \subseteq \mathcal{X}$, the set $\mathcal{Y}$ is called an *extreme face of* $\mathcal{X}$ if for all $y \in \mathcal{Y}$ and $x_1, x_2 \in \mathcal{X}$, the equality $y = \frac{1}{2}(x_1 + x_2)$ implies $x_1, x_2 \in \mathcal{Y}$. It is readily seen that an extreme face of a cone is a cone itself. The cones $\{0\}$ and $\mathcal{C}$ are called the *trivial extreme faces* of $\mathcal{C}$.

Definition 1.1 (see [Bar81])**.** Given a subset $\mathcal{X} \subseteq \mathcal{C}$, we denote by $\mathcal{F}(\mathcal{X})$ the smallest extreme face of the cone $\mathcal{C}$ that contains the set $\cup_{x\in\mathcal{X}}\{y \in \mathcal{C} \colon 0 \leqslant y \leqslant x\}$. When the set $\mathcal{X}$ is reduced to a single element x, we simply write $\mathcal{F}(x)$.

We give in the subsequent lemma two basic properties of the map $\mathcal{F}$ and a computational description of the value $\mathcal{F}(\mathcal{X})$.

Lemma 1.4. *The map $\mathcal{F}$ is monotone and idempotent: $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{C}$ implies $\mathcal{F}(\mathcal{A}) \subseteq \mathcal{F}(\mathcal{B})$ and $\mathcal{F}\big(\mathcal{F}(\mathcal{A})\big) = \mathcal{F}(\mathcal{A})$. Moreover, we have*

$$\mathcal{F}(x) = \{\lambda y \colon 0 \leqslant y \leqslant x,\, \lambda \geqslant 0\} \quad \textit{and} \quad \mathcal{F}(\mathcal{X}) = \mathcal{F}\Big(\sum_{x\in\mathcal{X}} \mathcal{F}(x)\Big)\,.$$

Finally, the elements of the extreme face $\mathcal{F}(\mathcal{A})$ are exactly the non-negative elements of $\operatorname{span}\mathcal{F}(\mathcal{A})$*: $\mathcal{F}(\mathcal{A}) = \mathcal{C} \cap \operatorname{span}\mathcal{F}(\mathcal{A})$.*

Proof. By definition, the set $\mathcal{F}(\mathcal{B})$ contains the set $\mathcal{B}$, and thus also contains the set $\mathcal{A}$. It follows that $\mathcal{F}(\mathcal{B})$ is an extreme face that contains $\mathcal{A}$, thus $\mathcal{F}(\mathcal{A}) \subseteq \mathcal{F}(\mathcal{B})$.

The set $\mathcal{F}(\mathcal{A})$ is an extreme face, thus, by Definition 1.1, the smallest face of $\mathcal{C}$ containing $\mathcal{F}(\mathcal{A})$ is itself, i.e. $\mathcal{F}\big(\mathcal{F}(\mathcal{A})\big) = \mathcal{F}(\mathcal{A})$.

We denote by $G(x) \coloneqq \{\lambda y \colon 0 \leqslant y \leqslant x,\, \lambda \geqslant 0\}$. This set is included in $\mathcal{C}$. It is also cone since it is closed under multiplication by a positive scalar and addition since $\lambda y_1 + \mu y_2 = (\lambda+\mu)\big[\frac{\lambda}{\lambda+\mu}y_1 + \frac{\mu}{\lambda+\mu}y_2\big]$. Moreover, it trivially contains the set $\{y \in \mathcal{C} \colon 0 \leqslant y \leqslant x\}$, thus we have $\mathcal{F}(x) \subseteq G(x)$. Conversely, let $y \in \mathcal{F}(x)$. Since $\mathcal{F}(x)$ is a cone, it must contain λy for all positive λ, hence $G(x) \subseteq \mathcal{F}(x)$.

We denote by $H(\mathcal{X}) \coloneqq \mathcal{F}\big(\sum_{x\in\mathcal{X}} \mathcal{F}(x)\big)$. We have $\mathcal{X} \subseteq \sum_{x\in X} \mathcal{F}(x)$ since $x \in \mathcal{F}(x)$, thus $\mathcal{F}(\mathcal{X}) \subseteq H(\mathcal{X})$ by monotony. Conversely, following Lemma 1.4, we can write $H(\mathcal{X})$ as $\mathcal{F}(\mathcal{Y})$, with $\mathcal{Y} \coloneqq \big\{\sum_i \lambda_i y_i \colon \lambda_i \geqslant 0,\ 0 \leqslant y_i \leqslant x_i,\ x_i \in \mathcal{X}\big\}$, where all sums are taken on a finite number of elements. Since $\mathcal{F}(\mathcal{X})$ is a cone, every element $\sum_i \lambda_i y_i \in \mathcal{Y}$ must also belong to $\mathcal{F}(\mathcal{X})$. By monotony and idempotence of the map $\mathcal{F}$, we deduce that $H(\mathcal{X}) = \mathcal{F}(\mathcal{Y}) \subseteq \mathcal{F}\big(\mathcal{F}(\mathcal{X})\big) = \mathcal{F}(\mathcal{X})$.

The inclusion $\mathcal{F}(\mathcal{A}) \subseteq \mathcal{C} \cap \operatorname{span}\mathcal{F}(\mathcal{A})$ holds trivially. Conversely, let $x, y \in \mathcal{F}(\mathcal{A})$ such that $x - y \in \mathcal{C}$. We have $0 \preccurlyeq x - y \preccurlyeq x$ and $x \in \mathcal{F}(\mathcal{A})$, so by definition of an extreme face, we must have $x - y \in \mathcal{F}(\mathcal{A})$. □

Let us point out the extreme faces of the cones introduced in Section 1.1.2.

Example 1.1.

1. We have $\mathcal{F}(0) = \{0\}$ and $\mathcal{F}(x) = \mathcal{C}$ when x belongs to the interior of $\mathcal{C}$.
2. In the case of the non-negative orthant $(\mathbb{R}_+)^n$, we have $y \in \mathcal{F}(x)$ if and only if $y_i = 0$ whenever $x_i = 0$.

3. More generally, in the case of the polyhedral cone $\mathcal{C}_P$, we have $y \in \mathcal{F}(x)$ if and only if $(Py)_i = 0$ whenever $(Px)_i = 0$.

4. Let $\|\cdot\|$ denote a strictly convex norm, and $\mathcal{L}_n$ the associated Lorentz cone. Then, for $x \in \partial\mathcal{L}_n$, we have $\mathcal{F}(x) = \mathbb{R}_+ x$.

5. In the cone of positive semidefinite matrices, we have $y \in \mathcal{F}(x)$ if and only if $\operatorname{ran} y \subseteq \operatorname{ran} x$.

The smallest non-trivial extreme faces are called *extreme rays*, and consist of vectors $u \in \mathcal{C}$ such that $0 \preccurlyeq y \preccurlyeq u$ implies that $y = \lambda u$ for some non-negative λ. We denote by $\operatorname{Extr}(\mathcal{C})$ the set of extreme rays of the cone $\mathcal{C}$. Let us also point out that extreme rays are exactly the extreme points of compact sections of $\mathcal{C}$:

Lemma 1.5. *Let c in the interior of $\mathcal{C}^\star$. Then $u \in \mathcal{C}$ is an extreme ray of $\mathcal{C}$ if and only if u is an extreme point of the compact set $\{y \in \mathcal{C} \colon \langle c, u - y\rangle = 0\}$.*

We recall that a vector can be decomposed into "complementary" positive and negative parts.

Lemma 1.6. *Any vector $x \in E$ can be written as $x = x^+ - x^-$ with $x^+, x^- \succcurlyeq 0$ and $\mathcal{F}(x^+) \cap \mathcal{F}(x^-) = \{0\}$.*

Proof. Let $x \in E$. The cone $\mathcal{C}$ has non-empty interior, hence $E = \mathcal{C} - \mathcal{C}$, hence the existence of $x^+, x^- \in \mathcal{C}$ such that $x = x^+ - x^-$. Among all such decompositions, we choose one than minimizes the sum of the dimensions of the subspaces spanned by the faces $\mathcal{F}(x^+)$ and $\mathcal{F}(x^-)$ (it exists since the cone $\mathcal{C}$ is closed and finite-dimensional).

Assume that the set $F \coloneqq \mathcal{F}(x^+) \cap \mathcal{F}(x^-)$ is not reduced to 0, and let u in the relative interior of this set. We also define λ by

$$\lambda \coloneqq \sup\{\mu > 0 \colon \mu u \preccurlyeq x^+, x^-\}\,,$$

and denote by $y^\pm \coloneqq x^\pm - \lambda u \succcurlyeq 0$, so that $x = y^+ - y^-$. By symmetry, we assume that λ saturates the inequality $\lambda u \preccurlyeq x^+$. The inequality $x^+ \preccurlyeq \alpha y^+$ for $\alpha > 1$ is equivalent to $x^+ - \frac{\alpha}{\alpha-1}\lambda u \succcurlyeq 0$. Since $\frac{\alpha}{\alpha-1} > 1$, this inequality cannot be satisfied, otherwise it contradicts the definition of λ. Hence the face $\mathcal{F}(y^+)$ is a proper sub-face of $\mathcal{F}(x^+)$ and $\mathcal{F}(y^-) \subset \mathcal{F}(x^-)$, which contradicts the definition of $x^\pm$. □

Extreme rays of a cone contain a lot of information, as shown by the following lemmas. First, an element $a \in \mathcal{C}$ is the supremum of all extreme rays that it dominates. Moreover, in order to compare two elements in $\mathcal{C}$, it is sufficient to compare the extreme rays that they dominate.

Lemma 1.7. *Let $a \in \mathcal{C}$. Then a is the supremum of all $u \in \operatorname{Extr}(\mathcal{C})$ that it dominates:*

$$\{a\} = \bigvee \{u \in \operatorname{Extr}(\mathcal{C}) \colon u \preccurlyeq a\}\,.$$

Proof. For simplicity, we denote by $\mathcal{M} \coloneqq \bigvee\{u \in \operatorname{Extr}(\mathcal{C}) \colon u \preccurlyeq a\}$. By definition of a minimal upper bound, there must be $b \in \mathcal{M}$ such that $b \preccurlyeq a$.

Assume that there is $m \in \mathcal{M}$ that does not dominate a, i.e. $m - a \notin \mathcal{C}$. Then, by Lemma 1.6, we can write $m - a = x^+ - x^-$, with $x^- \neq 0$ and $\mathcal{F}(x^+) \cap \mathcal{F}(x^-) = \{0\}$.

The vector x^- can be written as the sum of non-zero distinct extreme rays $x^- = \sum_i u_i$. Let simply $u \coloneqq u_1$ and $v \coloneqq \sum_{i>1} u_i$, so that $\mathcal{F}(u) \cap \mathcal{F}(v) = \{0\}$. Hence $m + u + v = x^+ + a$. We define the map ϕ by

$$\phi_u \colon x \mapsto \sup\{\lambda > 0 \colon \lambda u \preccurlyeq x\}\,.$$

First, note that $\phi_u(a) > 0$. Then, since $\mathcal{F}(x^+) \cap \mathcal{F}(u) = \{0\}$, we must have $\phi_u(x+a) = \phi_u(a)$ and similarly $\phi_u(m+u+v) = \phi_u(m) + 1$ since $\mathcal{F}(v) \cap \mathcal{F}(u) = \{0\}$.

We deduce that the extreme ray $u_0 \coloneqq \phi_u(a)u$ satisfies $u_0 \preccurlyeq a$ holds but $u_0 \preccurlyeq m$ does not. This contradicts the fact that $m \in \mathcal{M}$, hence we must have $a \preccurlyeq m$. In particular, $b \in \mathcal{M}$, hence $b = a$. We deduce that $a \in \mathcal{M}$, so $\mathcal{M} \succcurlyeq a$ and $\mathcal{M} = \{a\}$ by minimality. □

Lemma 1.8. *Let $a, b \in \mathcal{C}$. If $u \preccurlyeq a$ implies $u \preccurlyeq b$ for all $u \in \mathrm{Extr}(\mathcal{C})$, then $a \preccurlyeq b$.*

Proof. By Lemma 1.7, we have

$$\{u \in \mathrm{Extr}(\mathcal{C}) \colon u \preccurlyeq a\} \preccurlyeq \bigvee \{u \in \mathrm{Extr}(\mathcal{C}) \colon u \preccurlyeq b\} = \{b\}$$

Hence there must be a minimal upper bound of $\{u \in \mathrm{Extr}(\mathcal{C}) \colon u \preccurlyeq a\}$ that is less than b. By Lemma 1.7, the vector a is the unique minimal upper bound of this set, hence $a \preccurlyeq b$. □

1.1.5 Faces of the dual cone

We point out a canonical way to map extreme faces of a cone $\mathcal{C}$ to extreme faces of its dual cone $\mathcal{C}^\star$.

Definition 1.2. Given an extreme face $\mathcal{F}$ of the cone $\mathcal{C}$, we denote by $\mathcal{F}^\sharp$ the set defined by

$$\mathcal{F}^\sharp \coloneqq \mathcal{C}^\star \cap \mathcal{F}^\perp\,.$$

In other words, we have

$$y \in \mathcal{F}^\sharp \iff \Big(y \in \mathcal{C}^\star \text{ and } \langle x, y\rangle = 0,\ \forall x \in \mathcal{F}\Big)\,.$$

Lemma 1.9.

1. *If $\mathcal{F}$ is an extreme face of the cone $\mathcal{C}$, then $\mathcal{F}^\sharp$ is an extreme face of the dual cone $\mathcal{C}^\star$.*
2. *We have $(\cap_k \mathcal{F}_k)^\sharp = \sum_k \mathcal{F}_k^\sharp$.*

Proof. Indeed, let $y_1, y_2 \in \mathcal{C}^\star$, $y \in \mathcal{F}^\sharp$ such that $2y = y_1 + y_2$, and $x \in \mathcal{F}$. By definition of the dual cone, we have $\langle y_i, x\rangle \geqslant 0$ for $i \in \{1,2\}$. Moreover, by definition of y, we have $\langle y_1 + y_2, x\rangle = 0$, hence $\langle y_i, x\rangle = 0$ for $i \in \{1,2\}$. This holds for all $x \in \mathcal{F}$, thus $y_1, y_2 \in \mathcal{F}^\sharp$.

By Lemma 1.4, any extreme face $\mathcal{F}$ must satisfy $\mathcal{F} = \mathcal{C} \cap \operatorname{span}\mathcal{F}$. We can then deduce from the definition the sequence of equalities:

$$(\cap_k \mathcal{F}_k)^\sharp = \mathcal{C}^\star \cap (\cap_k \mathcal{F}_k)^\perp = \mathcal{C}^\star \cap \big(\sum_k \mathcal{F}_k^\perp\big) = \sum_k \mathcal{F}_k^\sharp\,. \quad \square$$

Let us again illustrate the map $\cdot^\sharp$ on the extreme faces of the classical cones.

Example 1.2.

1. We have $\{0\}^\sharp = \mathcal{C}^\star$ and $\mathcal{F}(x)^\sharp = \{0\}$ when x belongs to the interior of $\mathcal{C}$.
2. In the case of the non-negative orthant $(\mathbb{R}_+)^n$, we have $y \in \mathcal{F}(x)^\sharp$ if and only if $y_i = 0$ whenever $x_i \neq 0$.
3. More generally, in the case of the polyhedral cone $\mathcal{C}_P$, a vector $y \in \mathcal{C}^\star$ belongs to the face $\mathcal{F}(x)^\sharp$ if and only if $y_i = 0$ whenever $(Px)_i > 0$.
4. Let $\|\cdot\|$ denote a strictly convex norm, and $\mathcal{L}_n$ the associated Lorentz cone. Then, for $x \in \partial\mathcal{L}_n$, we have $\mathcal{F}(x)^\sharp = \mathbb{R}_+\big(J_{1,n}x\big)$.
5. In the cone of positive semidefinite matrices, we have $y \in \mathcal{F}(x)^\sharp$ if and only if $\operatorname{ran} x \perp \operatorname{ran} y$, i.e. $xy = 0$

1.2 Characterization of minimal upper bounds

1.2.1 The main result

We now establish several equivalent characterizations of minimal upper bounds in orderings induced by cones.

Theorem 1.10. *Let E be a finite dimensional vector space, $\mathcal{C} \subset E$ be a cone and $\preccurlyeq$ denote the ordering induced by $\mathcal{C}$ on E. Given a compact subset $\mathcal{A} \subseteq E$ and $x \succcurlyeq \mathcal{A}$, the following assertions are equivalent:*

(i) *x is a minimal upper bound of $\mathcal{A}$ in E,*

(ii) $\cap_{a\in\mathcal{A}}\mathcal{F}(x-a) = \{0\}$,

(iii) $\sum_{a\in\mathcal{A}}\mathcal{F}(x-a)^\sharp = \mathcal{C}^\star$,

(iv) *there is $c \in \operatorname{int}\mathcal{C}^\star$ such that $\langle c, x\rangle \leqslant \langle c, y\rangle$ for all $y \succcurlyeq \mathcal{A}$.*

Moreover, when $\mathcal{A} \subseteq \mathcal{C}$, $x \succcurlyeq \mathcal{A}$ is a minimal upper bound of $\mathcal{A}$ if and only if

(v) *$x \in \mathcal{F}(\mathcal{A})$ and there is $c \in \operatorname{int}(\mathcal{C}^\star \cap V)$ such that $\langle c, x\rangle \leqslant \langle c, y\rangle$ for all $y \in V$ such that $y \succcurlyeq \mathcal{A}$, with $V = \operatorname{span}\mathcal{F}(\mathcal{A})$.*

The proof is given in Section 1.2.2. We first give a remark and describe the geometric interpretation of each assertion.

Remark 1.3. A similar result holds for maximal lower bounds of the set $\mathcal{A}$ since the map $x \mapsto -x$ is monotone. In this case, assertions (ii) and (iii) are unchanged, while assertions (iv) and (v) require instead that $\langle c, x\rangle \geqslant \langle c, y\rangle$ for all $y \preccurlyeq \mathcal{A}$.

The assertions (ii) and (iii) above are dual versions of one another and mean that, in some sense, the vector x "sticks" to the set $\mathcal{A}$ in a sufficient number of directions, and that these directions span the whole space E.

Moreover, the set $\{y \in E\colon y \succcurlyeq \mathcal{A}\}$ is convex and bounded below by $\mathcal{A}$. The map $y \mapsto \langle c, y\rangle$ is also strictly increasing when $c \in \operatorname{int}\mathcal{C}^\star$. Thus there must be $x \succcurlyeq \mathcal{A}$ which minimizes the value of this map. Assertion (iv) in Theorem 1.10 means that the minimal upper bounds of $\mathcal{A}$ are precisely the minimizers of the maps $y \mapsto \langle c, y\rangle$ for $c \in \operatorname{int}\mathcal{C}^\star$.

Assertion (v) in Theorem 1.10 can be seen as a refinement on assertion (iv) in the sense that the characterization no longer depends on the ambient space. Let us illustrate this fact. Let $\mathcal{C}_1, \mathcal{C}_2$ denote two cones in E. Then $\mathcal{C} \coloneqq \mathcal{C}_1 \times \mathcal{C}_2$ is a cone in $E \times E$, and the interior of $\mathcal{C}$ is $(\operatorname{int}\mathcal{C}_1) \times \operatorname{int}(\mathcal{C}_2)$. Now let $\mathcal{A}_1 \subset \operatorname{int}\mathcal{C}_1$, so that $\mathcal{F}(\mathcal{A}_1) = \mathcal{C}_1$. We also introduce the set $\mathcal{A} \coloneqq \mathcal{A}_1 \times \{0\} = \{(a, 0) \colon a \in \mathcal{A}_1\}$, so that $\mathcal{F}(\mathcal{A}) = \mathcal{C}_1 \times \{0\}$ and $V = E \times \{0\}$. The interior of the cone $\mathcal{C}_2$ does not appear in the data provided by the set $\mathcal{A}$. Thus, we would expect the interior of $\mathcal{C}_2$ not to appear in the characterization of minimal upper bounds of $\mathcal{A}$.

However, the latter cone does appear in assertion (iv). Indeed, it requires that a minimal upper bound $x = (x_1, x_2) \in \mathcal{C}$ minimizes the scalar product $\langle (c_1, c_2), (x_1, x_2)) \rangle$ for some $c_1 \in \operatorname{int}\mathcal{C}_1$ and $c_2 \in \operatorname{int}\mathcal{C}_2$.

Assertion (v) "corrects" this discrepancy: it is sufficient to consider candidates to be minimal upper bounds of $\mathcal{A}$ in $\mathcal{F}(\mathcal{A})$, and the vector c that selects these minimal upper bounds may be taken in the interior of the dual cone of $\mathcal{F}(\mathcal{A})$ in the vector space V. In the previous example, it is thus sufficient to consider upper bounds of $\mathcal{A}$ of the form $(x_1, 0)$ and variables c of the form $(c_1, 0)$.

1.2.2 Proof of Theorem 1.10

(i) $\iff$ (ii) Let $x \succcurlyeq \mathcal{A}$ and $u \in \cap_{a \in \mathcal{A}} \mathcal{F}(x - a)$. By definition, we have $u \succcurlyeq 0$. Since $\mathcal{A}$ is compact and E is finite-dimensional, there must be $\varepsilon > 0$ such that $\varepsilon u \preccurlyeq x - a$ for all $a \in \mathcal{A}$. Thus $a \preccurlyeq x - \varepsilon u \preccurlyeq x$ for all $a \in \mathcal{A}$. It follows that x is a minimal upper bound $\mathcal{A}$ if and only if $u = 0$.

(ii) $\iff$ (iii) This is a trivial consequence of Lemma 1.9 and the fact that $\{0\}^\sharp = \mathcal{C}^\star$.

(iii) $\implies$ (iv) The cone $\mathcal{C}^\star$ has non-empty interior, thus there must, by assertion (iii), be some vectors $a_k \in \mathcal{A}$ and $\lambda_k \in \mathcal{F}(x - a_k)^\sharp$ such that $c \coloneqq \sum_k \lambda_k \in \operatorname{int}\mathcal{C}^\star$. We compute the value of $\langle c, y \rangle$ for some $y \succcurlyeq \mathcal{A}$:

$$\langle c, y \rangle = \sum_k \langle \lambda_k, y - a_k \rangle + \sum_k \langle \lambda_k, a_k \rangle \geqslant \langle \sum_k \lambda_k, x \rangle = \langle c, x \rangle \,,$$

since, by definition of $\mathcal{F}(x - a_k)^\sharp$, we have $\langle \lambda_k, y - a_k \rangle \geqslant 0$ and $\langle \lambda_k, x - a_k \rangle = 0$.

(iv) $\implies$ (i) Assume that $x \succcurlyeq \mathcal{A}$ is not a minimal upper bound of $\mathcal{A}$, so there is a vector $z \neq x$ such that $\mathcal{A} \preccurlyeq z \preccurlyeq x$. By assertion (iv), there is a vector c in the interior of $\mathcal{C}^\star$ such that $\langle c, x \rangle \leqslant \langle c, y \rangle$ for all $y \succcurlyeq \mathcal{A}$. This holds in particular for $y = z$, so we have $\langle c, x \rangle \leqslant \langle c, z \rangle$. Moreover, since the vector c belongs to the interior of the dual cone, the map $y \mapsto \langle c, y \rangle$ is strictly monotone, thus $\langle c, z \rangle < \langle c, x \rangle$, which contradicts the previous inequality. Hence x must be a minimal upper bound.

(iv) $\implies$ (v) First we show that $x \in \mathcal{F}(\mathcal{A})$. For all $a \in \mathcal{A}$, we can write $x = (x - a) + a$, with $a, x - a \succcurlyeq 0$, thus we have $x \in \mathcal{F}(\mathcal{A}) + \mathcal{F}(x - a)$. This inclusion holds for all a. We deduce that

$$x \in \cap_{a \in \mathcal{A}} \big[\mathcal{F}(\mathcal{A}) + \mathcal{F}(x - a) \big] \subseteq \big[\cap_{a \in \mathcal{A}} \mathcal{F}(x - a) \big] + \mathcal{F}(\mathcal{A}) \,.$$

By assertion (ii), we have $\cap_{a\in\mathcal{A}}\mathcal{F}(x-a)=\{0\}$, hence $x\in\mathcal{F}(\mathcal{A})$. Next, since $\mathcal{F}(\mathcal{A})=\mathcal{C}\cap V$, we have $\mathcal{F}(\mathcal{A})^\star=\mathcal{C}^\star+V^\star=\mathcal{C}^\star+V^\perp$. We deduce that $\mathcal{C}^\star\cap V=\mathcal{F}(\mathcal{A})^\star\cap V$. The cone $\mathcal{F}(\mathcal{A})$ has non-empty interior in the vector space V, thus assertion (v) is a direct consequence of applying assertion (iv) in the space V.

(v) $\implies$ (iv) Since a minimal upper bound of $\mathcal{A}$ in $\operatorname{span}\mathcal{F}(\mathcal{A})$ is a minimal upper bound of $\mathcal{A}$ in E, this proof is the same as (iv) $\implies$ (i), except the vector space E has been replaced by the vector space V.

1.3 Application to classical cones

We now specialize Theorem 1.10 to each of the 3 types of cones that have been introduced in Section 1.1.2. We illustrate in each case the meaning of each condition, and provide some additional description of the set of minimal upper bounds in each case.

1.3.1 Polyhedral cones

In this section, let $\mathcal{C}_P$ denote the polyhedral cone associated with the full-rank matrix $P=\left(p_1^T\dots p_m^T\right)^T$ and $\preccurlyeq$ the induced ordering.

Corollary 1.11 (Minimal upper upper bounds in polyhedral cones)**.** *Given a compact subset $\mathcal{A}\subset\mathbb{R}^n$ and $x\succcurlyeq\mathcal{A}$, the following assertions are equivalent:*

1. *x is a minimal upper bound of $\mathcal{A}$ in $(E,\preccurlyeq)$,*
2. *There are linearly independent rows $\{p_i\}_{i\in I}$ of the matrix P and vectors $a_i\in\mathcal{A}$ such that $\langle p_i,x-a_i\rangle=0$ for all $i\in I$ and $\sum_{i\in I}p_i\in\operatorname{int}\mathcal{C}_P^\star$.*
3. *There is a subset $I\subset\{1,\cdots,m\}$ such that $\langle\sum_{i\in I}p_i,y-x\rangle$ is non-negative for all $y\succcurlyeq\mathcal{A}$ and $\sum_{i\in I}p_i\in\operatorname{int}\mathcal{C}_P^\star$.*

Proof. This proof is a translation of Theorem 1.10 to the special case of polyhedral cones. Let $I\subset\{1,\cdots,m\}$ such that $i\in I$ if and only if there is some $a\in\mathcal{A}$ such that $\langle p_i,x-a\rangle=0$. Let $y\in\cap_{a\in\mathcal{A}}\mathcal{F}(x-a)$. Then, for all $i\in I$, we must have $\langle p_i,y\rangle=0$. The fact that such a vector y is equal to 0 if and only if the vector $\sum_{i\in I}p_i$ is in the interior of $\mathcal{C}_P^\star$. □

In the case of polyhedral cones, we can provide some additional structural properties of the set of minimal upper bounds of a *finite set* $\mathcal{A}$. In particular, the set of minimal upper bounds is a polyhedral complex, (see [DLRS10] for more background).

Proposition 1.12. *The set of minimal upper bounds of a* finite set *$\mathcal{A}\subset E$ with respect to $\preccurlyeq$ is a compact polyhedral complex: it is given as the union of finitely many closed polyhedral cells C_I in E and $C_I\cap C_J$ is also a cell for all I,J.*

Each cell C_I corresponds to a subset $I\subset\{1,\cdots,m\}$ for which the rows $\{p_i\}_{i\in I}$ satisfy $\sum_{i\in I}p_i\in\mathcal{C}_P^\star$. It consists of the vectors x such that

$$x\succcurlyeq\mathcal{A}\text{ and }\langle p_i,x-a_i\rangle=0\,,\forall i\in I\,.$$

Moreover, the dimension of the cell C (i.e. the dimension of its affine hull) is equal to the co-dimension of $\operatorname{span}\{p_i\}_{i\in I}$ *.*

Finally, the number of such cells is bounded by $\sum_{k=2}^n\binom{m}{k}$.

Proof. The set of minimal upper bounds is closed because the cone $\mathcal{C}_P$ is closed.

By Corollary 1.11, an upper bound x of the set $\mathcal{A}$ is a minimal upper bound if and only if there is a subset $I \subset \{1, \cdots, m\}$ such that for all $i \in I$ there is a vector $a_i \in \mathcal{A}$ such that $\langle p_i, x - a_i \rangle = 0$, and such that $\sum_{i \in I} p_i \in \mathcal{C}_P^\star$.

We consider the set $S(I)$ of all minimal upper bounds y such that $\langle p_i, x - a_i \rangle = 0$ for all $i \in I$. The fact that y is an upper bound is written $\langle p_i, y - a \rangle \geqslant 0$ for all p_i and a. In combination with the previous equalities, we deduce that $S(I)$ is a polyhedron. Moreover, the vector y belongs to the intersection of the translated cone $a_1 + \mathcal{C}$ and the affine subspace defined by $\langle \sum_{i \in I} p_i, y \rangle = \sum_{i \in I} \langle p_i, a_i \rangle$ with $\sum_{i \in I} p_i \in \mathcal{C}_P^\star$. Hence, by Lemma 1.5, the polyhedron $S(I)$ is bounded. Since the cone $\mathcal{C}_P$ is pointed, no combination of inequalities of the form $\langle p_i, x - a \rangle \geqslant 0$ implies that x belongs to a lower dimensional subspace. Hence only equations of the form $\langle p_i, x - a_i \rangle = 0$ impact the dimension of the polyhedron, from which we deduce that its dimension equals the co-dimension of $\operatorname{span}\{p_i\}_{i \in I}$. This characterization also shows that the intersection of two cells is again a (possibly empty) cell.

The set of minimal upper bounds $S(I)$ is the result of choosing a subset $I \subset \{1, \cdots, m\}$ such that $\sum_{i \in I} p_i \in \mathcal{C}_P^\star$. By Theorem 1.10, it is sufficient that $\operatorname{card} I \leqslant n$. There are only finitely such subsets, hence the whole set of minimal upper bounds is bounded. We deduce at once the estimate on the number of cells by counting the number of such subsets. □

We illustrate this results on the polyhedral cone from Figure 1.1. Given two vectors in $\mathbb{R}^n$, we plot in Figure 1.4 the section of the cones emanating from these elements as well as sections from the ones emanating from minimal upper bounds. In this case, the set of minimal upper bounds is constituted of the reunion of 5 segments.

1.3.2 The (generalized) Lorentz cone

In this section, let $\|\cdot\|$ denote a strictly convex norm on $\mathbb{R}^{n+1}$, $\mathcal{L}_n$ denote the associated Lorentz cone and $\preccurlyeq$ the order induced by this cone. For convenience, we write elements of $\mathcal{L}_n$ as $\hat{x} = (t\ x)$, so that $\hat{x} \in \mathcal{L}_n \iff \|x\| \leqslant t$.

Corollary 1.13 (Minimal upper bounds in $\mathcal{L}_n$)**.** *Given a compact subset $\mathcal{A} \subset \mathbb{R}^{n+1}$ and $\hat{x} \succcurlyeq \mathcal{A}$, the following assertions are equivalent:*

1. *$\hat{x}$ is a minimal upper bound of $\mathcal{A}$,*
2. *there are $\hat{a}, \hat{b} \in \mathcal{A}$ such that $(\hat{x} - \hat{a})$ and $(\hat{x} - \hat{b})$ are not colinear and belong to the boundary of the Lorentz cone $\partial\mathcal{L}_n$,*
3. *there is $\hat{c} = (r\ c)^T \in E$ such that $\|c\| < r$ and $\langle \hat{c}, \hat{y} - \hat{x} \rangle \geqslant 0$ for all $\hat{y} \succcurlyeq \mathcal{A}$.*

Proof. Non-trivial extreme faces of the Lorentz cone associated with a strictly convex norm are 1-dimensional cones of the form $\mathcal{F}(\hat{x}) = \mathbb{R}_+ \hat{x}$ for some $\hat{x} \in \partial\mathcal{L}_n$. Thus, in order for the intersection of $\mathcal{F}(\hat{x} - \hat{a})$ over all $\hat{a} \in \mathcal{A}$ to be zero, it is both necessary and sufficient to have two elements $\hat{a}, \hat{b} \in \mathcal{L}_n$ such that $\mathbb{R}_+(\hat{x} - \hat{a}) \cap \mathbb{R}_+(\hat{x} - \hat{b})$ is zero. Finally, the vector $(r\ c)^T$ belongs to the interior of the Lorentz cone $\mathcal{L}_n$ if and only if $\|c\| < r$. □

[1] If the animation does not run due to reader compatibility issues, an online version is available at `http://www.cmap.polytechnique.fr/~stott/assets/thesis/animations/Ex_mub_poly_anim.mp4`

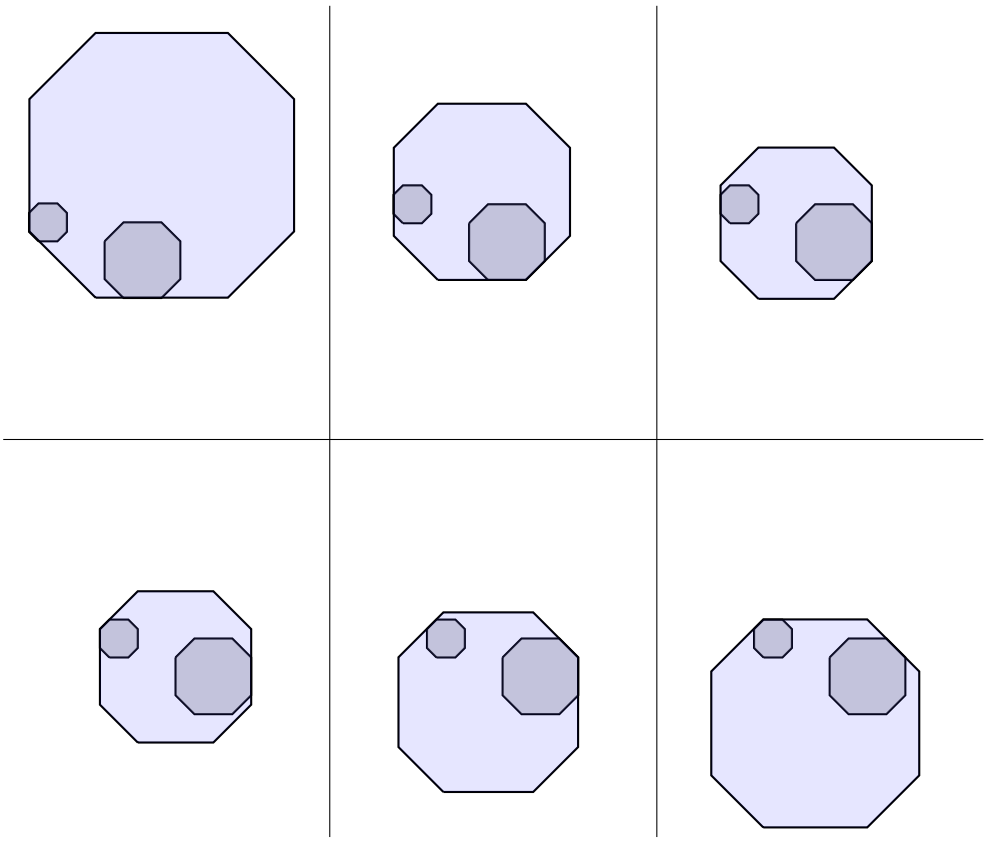


Figure 1.4: Representation of the minimal upper bounds of two elements with respect to a polyhedral cone [1]

It will be convenient to represent elements $\mathcal{A} \subset \mathbb{R}^{n+1}$ and upper bounds $\hat{x}$ of $\mathcal{A}$ by the intersections of the sets $\hat{x} - \mathcal{L}_n$ (resp. $a - \mathcal{L}_n$) with an affine hyperplane $\mathcal{H} \coloneqq \{z \colon \langle c, z\rangle = \alpha\}$ with $c \in \operatorname{int} \mathcal{L}_n$ and $\alpha < \min_{a\in\mathcal{A}} \langle c, a\rangle$. We use the notation $\mathcal{P}(\hat{x}) = (\hat{x} - \mathcal{L}_n) \cap \mathcal{H}$.

The sets $\mathcal{P}(\hat{x})$ (resp. $\mathcal{P}(\hat{a})$), also called *penumbras*, can be interpreted as the subset of $\mathcal{H}$ lit by a light source placed at $\hat{x}$ (resp. $\hat{a}$) whose light cone is $-\mathcal{L}_n$. This interpretation is a generalization of the physical case when the norm is the Euclidean norm.

In the special case where $\hat{x} = (t\ x)^T \in \mathcal{L}_n$, the penumbra $\mathcal{P}(\hat{x})$ is the $\|\cdot\|$-ball centered at x with radius t:

$$\mathcal{P}(\hat{x}) \coloneqq B(x, t) = \{y \in E \colon \|x - y\| \leqslant t\}\,.$$

The fact that $\hat{x} \in \mathcal{L}_n$ is equivalent to $0 \in \mathcal{P}(\hat{x})$. Moreover, $\hat{x} \preccurlyeq \hat{y}$ if and only if $\mathcal{P}(\hat{x}) \subseteq \mathcal{P}(\hat{y})$, because $\|y - x\| \leqslant t - s$ is equivalent to $s + \|y - x\| \leqslant t$. Moreover, recall that the vector $\hat{y}-\hat{x} = (t-s\ \ y-x)^T$ belongs is an extreme face of the Lorentz cone if and only if $t-s = \|y-x\|$, i.e. if and only if the balls $\mathcal{P}(\hat{x})$ and $\mathcal{P}(\hat{y})$ are tangent to one another. Thus $\hat{x} \in \mathbb{R}^{n+1}$ is a minimal upper bound of $\mathcal{A}$ if and only if $\mathcal{P}(\hat{x})$ is a minimal upper bound of $\cup_{\hat{a}\in\mathcal{A}} \mathcal{P}(\hat{a})$ in the inclusion order on the space of penumbras.

By Corollary 1.13, a necessary and sufficient condition for an upper bound $\hat{x}$ of $\mathcal{A}$ to be a minimal upper bound is the existence of vectors $\hat{a}$ and $\hat{b}$ such that either

1. $\hat{a} = \hat{b} = \hat{x}$ or

2. $(\hat{x} - \hat{a}), (\hat{x} - \hat{b})$ belong to distinct extreme faces.

In the first case, the vector $\hat{a} \in \mathcal{A}$ is an upper bound of $\mathcal{A}$, thus it is the only minimal upper bound of $\mathcal{A}$. In the second case, we show that there are infinitely many minimal upper bound that constitute a non-compact set. This aspect separates strictly convex Lorentz cones from polyhedral cones which are not strictly convex, since we have shown in Section 1.3.1 that the set of minimal upper bounds of a finite set with respect to a polyhedral cone is compact. This is summarized in the following proposition.

Proposition 1.14. *Let $\mathcal{L}_n$ denote the Lorentz cone associated with a strictly convex norm. The space $(E, \mathcal{L}_n)$ is an* anti-lattice*, i.e. two elements $\hat{a}, \hat{b}$ have a unique minimal upper bound if and only if they are comparable. When $\hat{a}$ and $\hat{b}$ are not comparable, the set of their minimal upper bounds is closed, unbounded and it can be identified to $\mathbb{R}^{n-1}$.*

Proof. Let us first consider the case $n = 2$. Assume that $\hat{a}, \hat{b}$ are not comparable. Then the sets $(\hat{a} + \mathcal{L}_n)$ and $(\hat{b} + \mathcal{L}_n)$ are not comparable in the inclusion order either. By Corollary 1.13, the set of minimal upper bounds of $\hat{a}, \hat{b}$ is exactly given by the intersection of the boundaries of the latter sets, i.e.

$$\bigvee\{\hat{a}, \hat{b}\} = (\hat{a} + \partial\mathcal{L}_n) \cap (\hat{b} + \partial\mathcal{L}_n)\,.$$

This intersection is a (continuous) curve in $\mathbb{R}^3$ that is unbounded in two directions, i.e. it can be identified with $\mathbb{R}$.

Now, let $n \geqslant 2$. Let $\mathcal{V}$ denote any 2-dimensional affine subspace that contains the centers of the $\|\cdot\|$-ball $\mathcal{P}(\hat{a}), \mathcal{P}(\hat{b})$. We denote by $\mathcal{D}(\hat{x})$ the disk obtained as the intersection of the $\|\cdot\|$-ball $\mathcal{P}(\hat{x})$ and the subspace $\mathcal{V}$ for any $\hat{x} = (t\ x)^T$ such that $x \in \mathcal{V}$. We then have the equivalence: $\mathcal{D}(\hat{x}) \subseteq \mathcal{D}(\hat{y})$ if and only if $\mathcal{P}(\hat{x}) \preccurlyeq \mathcal{P}(\hat{y})$. Indeed, if there were an element $z \in \mathbb{R}^n$ such that $z \in \mathcal{P}(s, x)$ and $z \notin \mathcal{P}(t, y)$, we would have

$$\|x\| + s \leqslant t < \|z\| \leqslant \|x\| + \|z - x\| \leqslant \|x\| + s\,.$$

By the first part of the proof, the set of minimal upper bounds whose penumbra's center belongs to $\mathcal{V}$ is unbounded and can be identified with $\mathbb{R}$.

The set of 2-dimensional affine subspaces in $\mathbb{R}^n$ that contains the centers of the balls $\mathcal{P}(\hat{a}), \mathcal{P}(\hat{b})$ is homeomorphic to

$$\mathcal{O}(n-1) \Big/ \big(\mathcal{O}(1) \times \mathcal{O}(n-2)\big) \cong \mathbb{R}^{n-2}$$

hence the whole space of minimal upper bounds can be identified with $\mathbb{R} \times \mathbb{R}^{n-2}$. □

Given two vectors in $\mathbb{R}^3$, we plot in Figure 1.5 the *penumbras* of these elements as well as the penumbras of some of their minimal upper bounds. In this case, the set of minimal upper bounds is parametrized by a single real variable.

[2] Again, if the animation does not run due to reader compatibility issues, an online version is available at `http://www.cmap.polytechnique.fr/~stott/assets/thesis/animations/Ex_mub_lor_anim.mp4`

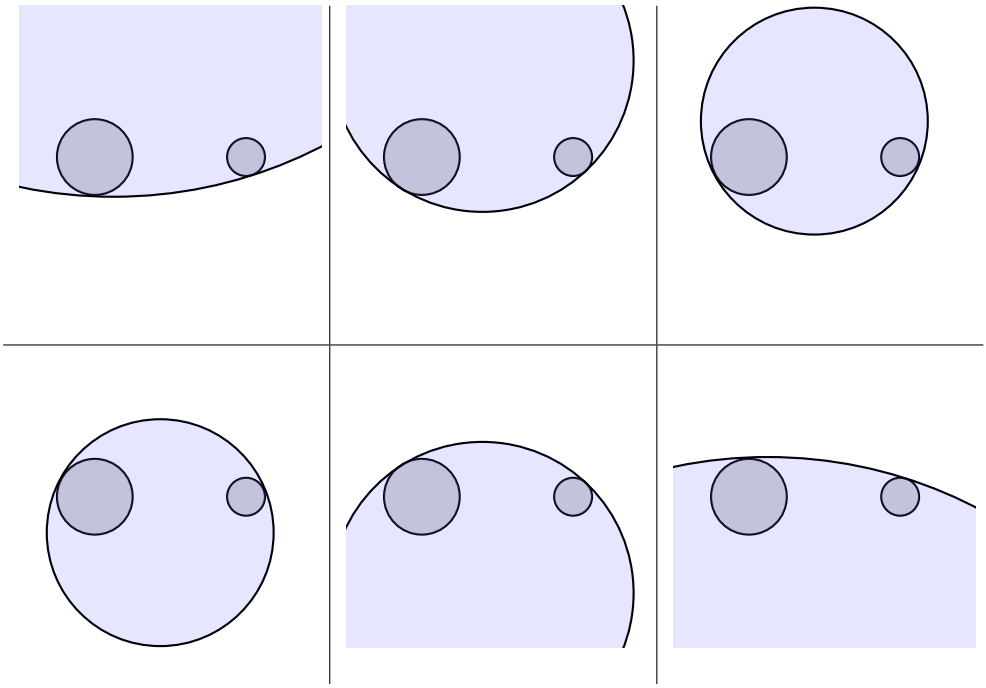


Figure 1.5: Partial representation of the minimal upper bounds of two elements with respect to the Euclidean Lorentz cone [2]

1.3.3 The cone of positive semidefinite matrices

In this section, we consider the space of symmetric matrices $\mathcal{S}_n$ endowed with the Löwner order $\preccurlyeq$ arising from the cone of positive semidefinite matrices $\mathcal{S}_n^+$. We specialize Theorem 1.10 to this case:

Theorem 1.15 (Minimal upper bounds in $\mathcal{S}_n$)**.** *Let $\mathcal{A}$ be a compact subset of $\mathcal{S}_n$ and $X \in \mathcal{S}_n$ such that $X \succcurlyeq \mathcal{A}$, where $\succcurlyeq$ denotes the Löwner order. The following assertions are equivalent:*

1. *X is a minimal upper bound of $\mathcal{A}$,*
2. *$\cap_{A\in\mathcal{A}} \operatorname{ran}(X - A) = \{0\}$,*
3. *$\sum_{A\in\mathcal{A}} \ker(X - A) = \mathbb{R}^n$,*
4. *there is a positive definite matrix C such that $\langle C, X\rangle \leqslant \langle C, Y\rangle$ for all $Y \succcurlyeq \mathcal{A}$.*

Moreover, if $\mathcal{A}$ is a compact subset of $\mathcal{S}_n^+$, the matrix $X \succcurlyeq \mathcal{A}$ is a minimal upper bound of $\mathcal{A}$ if and only if

5. *$\operatorname{ran} X = \sum_{A\in\mathcal{A}} \operatorname{ran} A$ and there is a positive semidefinite matrix C such that $\langle C, X\rangle \leqslant \langle C, Y\rangle$ for all $Y \succcurlyeq \mathcal{A}$ and $\operatorname{ran} C \supseteq \sum_{A\in\mathcal{A}} \operatorname{ran} A$.*

Proof. We have previously shown that $y \in \mathcal{F}(x)$ if and only if $\operatorname{ran} y \subseteq \operatorname{ran} x$, thus $y \in \cap_{a\in\mathcal{A}}\mathcal{F}(x-a)$ if and only $\operatorname{ran} y \subseteq \cap_{a\in\mathcal{A}} \operatorname{ran}(x-a)$. Thus we must have $\cap_{a\in\mathcal{A}} \operatorname{ran}(x-a) = \{0\}$. Taking the orthogonal complement on each side of the last equation yields $\sum_{a\in\mathcal{A}} \ker(x-a) = \mathbb{R}^n$. Moreover, since the cone $\mathcal{S}_n^+$ is self-dual, the interior of its dual cone is exactly the set of positive definite matrices. □

We study the structure of the set of minimal upper bounds in more detail in Chapter 2, Section 2.6.

CHAPTER 2

Minimal upper bounds of two symmetric matrices

This chapter is based on the article "Maximal lower bounds in the Löwner order" [Sto16].

2.1 Introduction

A classical result by Kadison shows that the space of symmetric matrices equipped with the Löwner order is an antilattice, meaning that two elements have a unique maximal lower (called the greatest lower bound) or a unique minimal upper bound (called the smallest upper bound) only in trivial cases:

Theorem 2.1 (Kadison, see [Kad51])**.** *Two symmetric matrices cannot have a greatest lower bound in the Löwner order unless they are comparable in this order.*

Equivalently, two symmetric matrices cannot have a smallest upper bound in the Löwner order unless they are comparable in this order.

We refer to the work of Kalauch, Lemmens, and van Gaans [KLvG14] for a recent approach to Kadison's theorem and generalizations in the setting of Riesz spaces. Lower bounds of symmetric matrices have also been extensively studied in the setting of *quantum observables* [And99, MG99, DDL06], where the main motivation is the uniqueness of a positive semidefinite maximal lower bound. Moreland and Gudder have solved this problem in [MG99]. Their result has been generalized to any pair of positive semidefinite bounded self-adjoint operators by Ando [And99]. His proof involved the notion of *generalized short*, which in the finite dimensional case is defined for positive semidefinite matrices X, Y by

$$[Y]X \coloneqq \max\{Z \colon 0 \preccurlyeq Z \preccurlyeq X,\ \operatorname{ran} Z \subseteq \operatorname{ran} Y\}\,.$$

Their results show that the uniqueness of a positive semidefinite maximal lower bound is decided by the comparability of such generalized shorts:

Theorem 2.2 (Moreland and Gudder, Ando, see [MG99, And99])**.** *Two positive semidefinite matrices A and B cannot have a unique positive semidefinite maximal lower bound unless the generalized shorts $[A]B$ and $[B]A$ are comparable.*

The aforementioned theorems raise the issue of characterizing the whole set of minimal upper bounds (or maximal lower bounds) of two symmetric matrices A and B. Our first main result (Theorem 2.3) shows that this set can be identified to the quotient space

$$\mathcal{O}(p,q)/(\mathcal{O}(p)\times\mathcal{O}(q))$$

where (p,q) denote the inertia of the matrix $B-A$, $\mathcal{O}(p)$ denotes the p-th orthogonal group, and $\mathcal{O}(p,q)$ is the indefinite orthogonal group arising from a quadratic form with inertia (p,q), see Definition 2.1. It follows that the set of minimal upper bounds is of dimension pq.

This result has a geometric consequence that will be dealt with in Section 2.5. In some cases (described in detail in Section 2.5), the Löwner order corresponds to the inclusion order of quadrics, up to a reversal. We deduce from Theorem 2.3 that given a quadric $\mathcal{Q}_X$ minimally enclosing two bounded quadrics $\mathcal{Q}_A, \mathcal{Q}_B$, the set of tangency points of $\mathcal{Q}_X$ with $\mathcal{Q}_A$ (resp. $\mathcal{Q}_B$) spans the kernel of $X-A$ (resp. $X-B$).

We have already shown in Section 1.3 that the matrix X is a minimal upper bound of A and B if and only if $\ker(X-A)+\ker(X-B)=\mathbb{R}^n$. It is desirable to find a parametrization of the minimal upper bounds X so as to obtain specific kernels, satisfying for instance inclusion conditions $\mathcal{U}\subseteq\ker(X-A)$ and $\mathcal{V}\subseteq\ker(X-B)$. Such conditions on the kernels arise from choosing tangency conditions on the associated quadrics. Our second main result (Theorem 2.10) leads to such a parametrization.

Finally, we explore in Section 2.6 the case of $p\geqslant 3$ matrices and point out the challenges related to the parametrization of their minimal upper bounds.

Although the present results are stated for real quadratic forms, they carry over to hermitian forms, up to immediate changes.

2.2 Notation

The *inertia* of the symmetric matrix A is the triple (p,q,r), where p (resp. q,r) is the number of positive (resp. negative, zero) eigenvalues of A, counted with multiplicities. We denote by $J_{p,q,r}$ the canonical bilinear form of inertia (p,q,r) on $\mathbb{R}^{p+q+r}$. It is defined by

$$J_{p,q,r}(x,y)=\sum_{i=1}^{p}x_iy_i-\sum_{i=p+1}^{p+q}x_iy_i\,,$$

and the corresponding matrix in the canonical basis of $\mathbb{R}^n$ is $I_p\oplus(-I_q)\oplus 0_r$, where I_n (resp. 0_n) denote the identity matrix (resp. zero matrix) of size $n\times n$. When $r=0$, we use the notation $J_{p,q}$.

Definition 2.1 (Indefinite orthogonal group $\mathcal{O}(p,q)$)**.** We denote by $\mathcal{O}(p,q)$ the indefinite orthogonal group of square matrices S such that $SJ_{p,q}S^T=J_{p,q}$. When $q=r=0$, $\mathcal{O}(p,q)$ becomes the standard orthogonal group $\mathcal{O}(p)$.

We recall that, given a positive semidefinite matrix M, the square root of the matrix M is the unique positive semidefinite matrix, denoted $M^{1/2}$, such that $M^{1/2}M^{1/2} = M$. The absolute value of a symmetric matrix M is given by $|M| = (MM^T)^{1/2}$. The set of $p \times q$ matrices is denoted by $\mathcal{M}_{p,q}$.

2.3 Parametrization of the set of minimal upper bounds of two symmetric matrices

2.3.1 Statement of the main theorem

Our first main result parametrizes the set of minimal upper bounds of two symmetric matrices with respect to the Löwner order. It implies that this set is of dimension pq and that it can be identified with $\mathcal{O}(p,q)/\big(\mathcal{O}(p) \times \mathcal{O}(q)\big)$, the quotient set of the indefinite orthogonal group $\mathcal{O}(p,q)$ by the maximal compact subgroup $\mathcal{O}(p) \times \mathcal{O}(q)$.

Theorem 2.3. *Let $A, B, X \in \mathcal{S}_n$ be such that $X \succcurlyeq A, B$, and let (p,q,r) denote the inertia of $B - A$. The following statements are equivalent:*

(i) *X is a minimal upper bound of A and B*

(ii) $\ker(X - A) + \ker(X - B) = \mathbb{R}^n$

(iii) *there is a positive definite matrix C such that*

$$X = \frac{A+B}{2} + C^{1/2}\frac{|C^{-1/2}(B-A)C^{-1/2}|}{2}C^{1/2}\,.$$

(iv) *For all $P \in \mathrm{GL}_n$ revealing the inertia of $B - A$, i.e. such that $B - A = PJ_{p,q,r}P^T$, there exists a unique $M \in \mathcal{M}_{p,q}$ such that:*

$$\begin{aligned} X &= A + PS(I_p \oplus 0_q \oplus 0_r)SP^T \quad \textit{with} \\ S &= \begin{pmatrix} (I_p + MM^T)^{1/2} & M \\ M^T & (I_q + M^TM)^{1/2} \end{pmatrix} \oplus 0_r \ . \end{aligned}$$

Each assertion brings a different perspective on the nature of minimal upper bounds and their characterization.

Assertion (ii) states that the matrix X "sticks tightly" to the matrices A, B, in the sense that the space $\mathbb{R}^n$ can be decomposed as a sum of two subspaces, such that the linear map associated with X coincides with the linear map associated with either A or B on each of these subspaces. This is developed in detail in Section 2.4.

Assertion (iii) is an explicit formulation of the characterization given by Theorem 1.15, exhibiting the minimal upper bound as a positively exposed point on the boundary of the set of upper bounds of A, B.

Assertion (iv) give an exact parametrization of the set of minimal upper bounds in terms of its $p \times q$ degrees of freedom, and shows in the process that it is a non-compact set.

Remark 2.1. Assertion (iv) can also be rewritten in terms of the matrix B:

$$X = B + PS(0_p \oplus I_q \oplus 0_r)SP^T$$

Remark 2.2. A similar theorem holds for maximal lower bounds, in which case (ii) is unchanged, while (iii) and (iv) read :

(iii) there is a positive definite matrix C such that

$$X = \frac{A+B}{2} - C^{1/2}\frac{|C^{-1/2}(B-A)C^{-1/2}|}{2}C^{1/2}\,.$$

(iv) $X = A - PS(0_p \oplus I_q \oplus 0_r)SP^T = B + PS(I_p \oplus 0_q \oplus 0_r)SP^T$

Before proving Theorem 2.3, we draw two corollaries. Theorem 2.3, Corollary 2.4 and Corollary 2.5 are proved in Section 2.3.3.

Corollary 2.4. *Let $A, B \in \mathcal{S}_n$, and let (p, q, r) denote the inertia of $B - A$. Then, the sets of minimal upper bounds and maximal lower bounds of A and B are homeomorphic to the quotient set*

$$\bigvee\{A, B\} \cong \bigwedge\{A, B\} \cong \mathcal{O}(p, q) \Big/ \big(\mathcal{O}(p) \times \mathcal{O}(q)\big) \cong \mathbb{R}^{pq}\ .$$

Corollary 2.5. *Let $A, B \in \mathcal{S}_n^+$, and let (p', q', r') denote the inertia of $B[A] - A[B]$. The rank of a positive semidefinite maximal lower bound of A, B cannot exceed $p' + q' + \dim\ker(B - A)$. Moreover, the set of* positive semidefinite *maximal lower bounds of A and B which have this rank is homeomorphic to the quotient set*

$$\mathcal{S}_n^+ \cap \bigwedge\{A, B\} \cong \mathcal{O}(p', q') \Big/ \big(\mathcal{O}(p') \times \mathcal{O}(q')\big) \cong \mathbb{R}^{p'q'}\ .$$

We note that Kadison's result can be recovered as a special case of Corollary 2.4. Indeed, the existence of greatest lower bound of two matrices A, B is equivalent to the existence of a unique maximal lower bound of these matrices, which, by Corollary 2.4, cannot happen unless $pq = 0$, meaning that $A \preccurlyeq B$ or $B \preccurlyeq A$.

The result from Moreland and Gudder can be recovered from Corollary 2.5 in the same way. If two positive semidefinite matrices A, B have a unique positive semidefinite maximal lower bound X, then the uniqueness implies that $p'q' = 0$, which means that $[B]A \preccurlyeq [A]B$ or $[A]B \preccurlyeq [B]A$.

2.3.2 Preliminary lemmas

We present two results which will be useful in the proof of Theorem 2.3.

Lemma 2.6. *Let $P, Q \in \mathcal{S}_n$. We have*

$$\ker P + \ker Q = \mathbb{R}^n \implies \ker P \cap \ker Q = \ker(P - Q)$$

Proof. The inclusion $\ker P \cap \ker Q \subseteq \ker(P - Q)$ is trivial. Let $x \in \ker(P - Q)$ and assume $x \notin \ker P$. Then $Px = Qx \neq 0$, and so $\operatorname{ran} P \cap \operatorname{ran} Q \neq \{0\}$. Taking the orthogonal complement contradicts $\ker P + \ker Q = \mathbb{R}^n$. □

Lemma 2.7 (Polar decomposition of $\mathcal{O}(p, q)$, see [Dra12, Section 6.2] and [Gal12, Proposition 4.11])**.** *For every $S \in \mathcal{O}(p, q)$, there exists a unique triple $(M, U, V) \in \mathcal{M}_{p,q} \times \mathcal{O}(p) \times \mathcal{O}(q)$ such that:*

$$S = \begin{pmatrix} \big(I_p + MM^T\big)^{1/2} & M \\ M^T & \big(I_q + M^TM\big)^{1/2} \end{pmatrix} \begin{pmatrix} U & \\ & V \end{pmatrix}\ .$$

2.3.3 Proof of Theorem 2.3, Corollary 2.4 and Corollary 2.5

We now prove Theorem 2.3. The equivalence of (i) and (ii) has already been shown in Theorem 1.15. We shall prove

$$\text{(ii)} \iff \text{(iv)} \text{ and } \text{(i)} \iff \text{(iii)}\,.$$

(ii) $\implies$ (iv)
Without loss of generality, we may assume that $P = I_n$, so that $B - A = J_{p,q,r}$. We build a basis of $\mathbb{R}^n$ respecting the decomposition

$$\mathbb{R}^n = K_A \oplus K_B \oplus \ker(A-B),$$
$$\ker(X-A) = K_A \oplus \ker(B-A)\,, \qquad \ker(X-B) = K_B \oplus \ker(B-A)\,.$$

We take a basis $\mathcal{B}_A$ of K_A, a basis $\mathcal{B}_B$ of K_B and $\mathcal{B}_{B-A}$ of $\ker(B-A)$, and our basis of $\mathbb{R}^n$ is $\big[\mathcal{B}_B\,;\,\mathcal{B}_A\,;\,\mathcal{B}_{B-A}\big]$. In this basis, the matrices of the quadratic forms $X-A$ and $X-B$ are block-diagonal matrices:

$$X - A = (0_p \oplus M_q \oplus 0_r) \text{ and } X - B = (M_p \oplus 0_q \oplus 0_r)\,,$$

where the off-diagonal blocks are zero, because the matrices M_p and M_q (respectively of size $p \times p$ and $q \times q$) are positive definite. The matrix $\Sigma = M_p^{1/2} \oplus M_q^{1/2}$ is in the indefinite orthogonal group $\mathcal{O}(p,q)$, since $\Sigma J_{p,q} \Sigma^T = J_{p,q}$. By Lemma 2.7, there is a unique tuple $(M,U,V) \in \mathcal{M}_{p,q} \times \mathcal{O}(p) \times \mathcal{O}(q)$ such that :

$$\Sigma = \begin{pmatrix} (I_p + MM^T)^{1/2} & M \\ M^T & (I_q + M^T M)^{1/2} \end{pmatrix} \begin{pmatrix} U & \\ & V \end{pmatrix}.$$

The matrix M does not depend on the choice of the matrix Σ: it is easily shown that all matrices Ξ such that $\Xi(\Xi^T) = M_p \oplus M_q$ only differ from Σ by a block-diagonal orthogonal-block matrix :

$$\Xi = \Sigma(U' \oplus V')\ ,\ U' \in \mathcal{O}(p),\ V' \in \mathcal{O}(q)\,.$$

Conversely, any block-diagonal orthogonal-block matrix of this form multiplied on the right vanishes when computing X. Indeed, if we denote

$$S = \begin{pmatrix} (I_p + MM^T)^{1/2} & M \\ M^T & (I_q + M^T M)^{1/2} \end{pmatrix} \oplus 0_r \quad \text{and} \quad W = U \oplus V \oplus 0_r\,,$$

we have $X = A + (SW)(0_p \oplus I_q \oplus 0_r)(SW)^T = A + S(0_p \oplus I_q \oplus 0_r)S$.

(iv) $\implies$ (ii)
Without loss of generality, we may again assume that $P = I_n$. After change of basis with the invertible matrix $Q \coloneqq \big(S + (0_p \oplus 0_q \oplus I_r)\big)^{-1}$, we have $Q(X-A)Q^T = (I_p \oplus 0_q \oplus 0_r)$ and $Q(X-B)Q^T = (0_p \oplus I_q \oplus 0_r)$. The sum of the kernels of those matrices is $\mathbb{R}^n$, thus this is also the case for $X-A$ and $X-B$.

(i) $\iff$ (iii)
By Remark 1.3 and Theorem 1.15, X is a minimal upper bound of A, B if and only if there is a positive definite matrix C such that $\langle C, X\rangle \leqslant \langle C, Y\rangle$ for all matrices $Y \succcurlyeq A, B$. First, note that $\langle C, Y\rangle = \langle I, C^{-1/2}YC^{-1/2}\rangle$ and $Y \succcurlyeq A$ is equivalent to $C^{-1/2}YC^{-1/2} \succcurlyeq C^{-1/2}AC^{-1/2}$, thus we may assume that $C = I_n$. Moreover, we may assume that the matrix $B - A$ is a diagonal matrix, also denoted by $D \coloneqq B - A$, and whose diagonal entries are sorted in decreasing order. Thus, we shall show that the unique matrix X which minimizes the map $Y \mapsto \langle I_n, Y\rangle = \mathrm{trace}(Y)$ over all upper bounds Y of A, B is given by

$$\frac{A+B}{2} + \frac{|B-A|}{2}\,. \tag{2.1}$$

Such a matrix X is the primal solution to the primal-dual pair of problems

$$\begin{array}{lll} \underset{X}{\text{minimize}} & \mathrm{trace}(X) & \underset{\lambda,\mu}{\text{maximize}} \quad \langle \lambda, A\rangle + \langle \mu, B\rangle\,. \\ \text{subject to} & X \succcurlyeq A, B & \text{subject to} \quad \lambda, \mu \succcurlyeq 0 \\ & & \qquad\qquad \lambda + \mu = I_n \end{array}$$

Here λ, μ are symmetric $n \times n$ matrices. The optimality conditions imply that X is the unique matrix such that

$$\begin{gathered} \lambda(X-A) = 0 \qquad \mu(X-B) = 0 \qquad \lambda + \mu = I_n \\ \lambda, \mu \succcurlyeq 0 \qquad X \succcurlyeq A, B\,, \end{gathered}$$

for some symmetric matrices λ, μ. One can then easily check that the matrix X given in Equation (2.1) satisfies these equations with $\lambda = 0_p \oplus I_{q+r}$ and $\mu = I_p \oplus 0_{q+r}$. This concludes the proof of Theorem 2.3. □

Proof of Corollary 2.4. We have shown in the proof $(ii) \implies (iv)$ of Theorem 2.3 that, given a matrix P revealing the inertia of the matrix $B - A$, we can associate with every matrix $\Sigma \in \mathcal{O}(p, q)$ a minimal upper bound X of A and B, via the continuous map Φ from $\mathcal{O}(p, q)$ to $\mathcal{S}_n$ defined by:

$$\Phi : \Sigma \mapsto A + P \begin{pmatrix} \Sigma & \\ & 0_r \end{pmatrix} \begin{pmatrix} I_p \oplus 0_q & \\ & 0_r \end{pmatrix} \begin{pmatrix} \Sigma & \\ & 0_r \end{pmatrix}^T P^T\,.$$

Moreover, we have previously shown that two matrices $\Sigma_1, \Sigma_2 \in \mathcal{O}(p, q)$ produce the same minimal upper bound X if and only if $\Sigma_1 = \Sigma_2(U \oplus V)$ for some matrices $U \in \mathcal{O}(p), V \in \mathcal{O}(q)$. By Theorem 2.3, the image of the map Φ is precisely the set of minimal upper bounds of A and B. This proves that the map Φ is a bijection from $\mathcal{O}(p, q)/\big(\mathcal{O}(p) \times \mathcal{O}(q)\big)$ to the set of minimal upper bounds of A, B. By Lemma 2.7, the quotient set can be identified to $\mathcal{M}_{p,q} \cong \mathbb{R}^{pq}$ by means of the continuous bijection S defined by:

$$S : M \mapsto \begin{pmatrix} (I_p + MM^T)^{1/2} & M \\ M^T & (I_q + M^TM)^{1/2} \end{pmatrix}.$$

It remains to show that the map $\Phi \circ S$ has a continuous inverse. We write

$$\Phi \circ S(M) = A + P\left[\begin{pmatrix} A(M) & B(M) \\ B(M)^T & M^TM \end{pmatrix} \oplus 0_r\right] P^T\,,$$

with $A(M) \coloneqq I_p + MM^T$ and $B(M) \coloneqq (I_p + MM^T)^{1/2}M$. The matrix M can be recovered continuously with $M = A(M)^{-1/2}B(M)$, since $A(M) \succcurlyeq I_p$ cannot vanish. □

Proof of Corollary 2.5. Before treating the general case, we shall prove the corollary when A, B are positive definite. Note that, in this case, we have $[A]B = B$ and $[B]A = A$.

First, the inertias of the matrices $B - A$ and $B^{-1} - A^{-1}$ are the same: the matrices A, B can be reduced simultaneously by an invertible congruence $X \mapsto PXP^T$ to diagonal matrices with positive diagonal elements a_i and b_i. The fact that $a_i - b_i > 0$ is equivalent to $b_i^{-1} - a_i^{-1} > 0$ shows that the inertias are identical. When the matrices A and B are positive definite, we have $p' + q' + \dim\ker(B - A) = n$, so matrices whose rank is equal to $p' + q' + \dim\ker(B - A)$ are invertible.

The map $X \mapsto X^{-1}$ is monotonically decreasing on the set of positive definite matrices [Bha07, Exercise 1.2.12]. Thus, it is a (continuous) bijection between the set of minimal upper bounds of A^{-1}, B^{-1} (which are positive definite) and the set of positive definite minimal upper bounds of A, B. By Corollary 2.4, the former set is homeomorphic to $\mathcal{O}(p', q')/\big(\mathcal{O}(p') \times \mathcal{O}(q')\big) \cong \mathbb{R}^{p'q'}$, so the same is true for the latter set.

Now let A, B denote positive semidefinite matrices. We may assume that $\ker(B - A) = \{0\}$, since it does not influence the structure of the set of minimal upper bounds of A, B by Theorem 2.3, so that $\ker A \cap \ker B = \{0\}$.

Let $R_{A,B}$ denote the set $\operatorname{ran} A \cap \operatorname{ran} B$. We claim that there are direct summands R_A and R_B of $R_{A,B}$ in $\operatorname{ran} A$ and $\operatorname{ran} B$ respectively so that the matrices of the quadratic forms A, B are block-diagonal in $\mathbb{R}^n = R_A \oplus R_{A,B} \oplus R_B$.

Indeed, we have $\mathbb{R}^n = \ker B \oplus R_{A,B} \oplus \ker A$. In such a decomposition, the quadratic forms A, B have matrices of the form

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix} \oplus 0_b \quad \text{and} \quad B = 0_a \oplus \begin{pmatrix} B_{22} & B_{23} \\ B_{23}^T & B_{33} \end{pmatrix}.$$

We define the matrix U mapping $w = (x, y, z) \in \ker B \oplus R_{A,B} \oplus \ker A$ to $Uw = (x - A_{11}^{-1}A_{12}y, y, z - B_{33}^{-1}B_{23}^T y)$. One can easily check that the subspaces R_A and R_B defined as the image of $\ker B$ and $\ker A$ respectively by U satisfy the desired condition. Moreover, up to a transformation $X \mapsto V^T XV$ with V block-diagonal, we may assume that $A = I_a \oplus S_A \oplus 0_b$ and $B = 0_a \oplus S_B \oplus I_b$, where S_A, S_B denote positive definite matrices such that $S_A - S_B = J_{p',q'}$. Note that the short $[B]A$ (resp. $[A]B$) is given by $0_a \oplus S_A \oplus 0_b$ (resp. $0_a \oplus S_B \oplus 0_b$).

Let X denote a positive semidefinite minimal upper bound of A, B. Using the characterization in Theorem 2.3, X is given in block form by

$$X = \begin{pmatrix} -\alpha\alpha^T - \beta\beta^T & * & * \\ * & * & * \\ * & * & -\beta^T\beta - \delta^T\delta \end{pmatrix} \text{ with } M = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in \mathcal{M}_{a+p',b+q'}.$$

The fact that X is positive semidefinite implies that the matrices α, β, δ are zero matrices, so that $X = 0_a \oplus S_X \oplus 0_b$, where the matrix S_X is given by

$$S_X = S_A - \begin{pmatrix} I_{p'} + \gamma\gamma^T & (I_{p'} + \gamma\gamma^T)^{1/2}\gamma \\ \gamma^T(I_{p'} + \gamma\gamma^T)^{1/2} & \gamma^T\gamma \end{pmatrix}.$$

By Theorem 2.3, S_X is a (positive semidefinite) minimal upper bound of the matrices S_A and S_B. This concludes the proof since S_A, S_B are positive definite. □

2.4 Minimal upper bounds selection under tangency constraints

2.4.1 Notation and preliminary lemma

We first give some notation that will be useful in the sequel. We define the linear operators π_p, π_q and π_r, mapping respectively $\mathbb{R}^n$ to $\mathbb{R}^p$, $\mathbb{R}^q$ and $\mathbb{R}^r$, that select the first p coordinates, the following q and the last r coordinates. Their matrices in the canonical basis of $\mathbb{R}^n$ are

$$\pi_p = \begin{pmatrix} I_p & 0_{pq} & 0_{pr} \end{pmatrix}, \qquad \pi_q = \begin{pmatrix} 0_{qp} & I_q & 0_{qr} \end{pmatrix}, \qquad \pi_r = \begin{pmatrix} 0_{rp} & 0_{rq} & I_r \end{pmatrix}.$$

We denote by $\|\cdot\|$ the spectral norm (largest singular value) of a matrix. We define $\mathcal{B}_{p,q}$ to be the open unit ball of $\mathcal{M}_{p,q}$ with respect to this norm:

$$\mathcal{B}_{p,q} := \left\{ M \in \mathcal{M}_{p,q} \colon \|M\| < 1 \right\}.$$

Lemma 2.8. *The map $\phi_{p,q}$ from $\mathcal{M}_{p,q}$ to $\mathcal{B}_{p,q}$ defined by :*

$$\phi_{p,q}(M) = \left(I_p + MM^T\right)^{-1/2} M$$

is a bijection, with inverse

$$\psi_{p,q}(N) = \left(I_p - NN^T\right)^{-1/2} N \ .$$

Moreover,

$$\phi_{p,q}(M) = M\left(I_q + M^T M\right)^{-1/2} \quad \textit{and} \quad \left(\phi_{p,q}(M)\right)^T = \phi_{q,p}(M^T) \ .$$

Proof. Let $M = UDV^T$ denote the singular value decomposition of M, so that U, V are orthogonal matrices, and D is a matrix consisting of a diagonal block and a zero block. Then, $\phi_{p,q}(M) = U\phi_{p,q}(D)V^T$, and a similar property holds for the map $\psi_{p,q}$. Therefore, it suffices to check that $\psi_{p,q} \circ \phi_{p,q}(M) = M$ when $M = D$, which is straightforward. By symmetry, we obtain that $\phi_{p,q} \circ \psi_{p,q}(N) = N$ holds for all N. The other properties are proved similarly. □

2.4.2 Statement of the problem and the theorem

As stated in Theorem 2.3, the kernels $\ker(X-A)$ and $\ker(X-B)$ are central to the characterization of minimal upper bounds. In the following, we investigate the problem of the selection of a minimal upper bound of two symmetric matrices where subspaces of those kernels have been predetermined. When $u^T X u > 0$, the line $\mathbb{R}u$ meets the surface $\{z \in \mathbb{R}^n \colon z^T X z = 1\}$ at two opposite points $\pm\alpha u$, with $\alpha > 0$. Moreover, if $u \in \ker(X - A)$, the surfaces $\{z \in \mathbb{R}^n \colon z^T X z = 1\}$ and $\{z \in \mathbb{R}^n \colon z^T A z = 1\}$ are tangent at those points. Indeed, the equation $A(\alpha u) = X(\alpha u)$ means that the gradient of the quadratic forms $z \mapsto z^T X z$ and $z \mapsto z^T A z$ at the point αu are colinear. Both isosurfaces contain the point αu, thus they are tangent at this point. When $u^T X u \leqslant 0$, it may be interpreted as a tangency at ∞. For this reason, constraints on the kernels are called *tangency constraints*.

Finally, following Theorem 2.3, the dimension of the kernel of $B - A$ does not influence the structure of the set of minimal upper bound of A and B. Thus, we assume that a reduction has been done and state Problem 2.9 and Theorem 2.10 accordingly.

Problem 2.9 (minimal upper bounds with tangency constraints)**.** *Let $A, B \in \mathcal{S}_n$ and let $(p, q, 0)$ denote the inertia of $B - A$. Let $\mathcal{U}, \mathcal{V}$ be subspaces of $\mathbb{R}^n$. We wish to find in $X \in \mathcal{S}_n$ such that:*

$$\begin{cases} X \text{ is a minimal upper bound of } A, B \\ \forall u \in \mathcal{U},\ Xu = Bu \\ \forall v \in \mathcal{V},\ Xv = Av \end{cases}$$

Our second main result gives conditions for Problem 2.9 to have a solution and, if these conditions are met, a parametrization of the set of solutions. It shows that the set of solutions can be identified with an affine subspace of $\mathcal{M}_{p,q}$ of dimension $(p - \dim \mathcal{U})(q - \dim \mathcal{V})$, so that the problem has a unique solution if and only if one of the subspaces has maximal dimension.

Theorem 2.10. *Problem 2.9 has a solution if and only if*

(i) $B - A$ is positive definite over $\mathcal{U}$

(ii) $B - A$ is negative definite over $\mathcal{V}$

(iii) $\mathcal{U}$ and $\mathcal{V}$ are orthogonal with respect to the indefinite scalar product $B - A$

If these conditions are met, then the set of solutions can be parametrized as in Theorem 2.3, assertion (iv)*, with*

$$M \in \phi_{p,q}^{-1}(\mathcal{B}_{p,q} \cap \mathcal{W})$$

where $\mathcal{W}$ is the affine subspace of $\mathcal{M}_{p,q}$ defined by

$$R \in \mathcal{W} \iff \begin{cases} \forall u \in \mathcal{U},\ R^T \pi_p(u) + \pi_q(u) = 0 \\ \forall v \in \mathcal{V},\ R\pi_q(v) + \pi_p(v) = 0 \end{cases} . \tag{2.2}$$

The intersection $\mathcal{B}_{p,q} \cap \mathcal{W}$ is nonempty when the conditions (i, ii, iii) are met.

Corollary 2.11. *The set of solutions of Problem 2.9 is parametrized by a subspace of $\mathcal{M}_{p,q}$ of dimension $(p - \dim \mathcal{U})(q - \dim \mathcal{V})$, so that the solution is unique if and only if*

$$\dim \mathcal{U} = p \text{ or } \dim \mathcal{V} = q\,.$$

Remark 2.3. When $\mathcal{U}$ and $\mathcal{V}$ have maximal dimension, $\mathcal{V}$ is the orthogonal complement of $\mathcal{U}$ with respect the indefinite form $B - A$. Thus Theorem 2.10 establishes a bijective correspondence between minimal upper bounds of A, B and p-dimensional subspaces over which the matrix $B - A$ is positive definite. In this way, the set of minimal upper bounds is parametrized by an open semi-algebraic subset of the Grassmannian $\mathrm{Gr}(n, p)$.

2.4.3 Preliminary lemmas

Before proving Theorem 2.10, we prove two useful results. First, Lemma 2.12 shows that when the matrix $J_{p,q}$ is negative definite over a subspace $\mathcal{V}$, then there is a contractive mapping from the last q coordinates of any vector of $\mathcal{V}$ to its first p coordinates.

Lemma 2.12. *Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$ over which $J_{p,q}$ is negative definite, with $p + q = n$. There is a matrix $R \in \mathcal{M}_{p,q}$ with $\|R\| < 1$ such that :*

$$\forall v \in \mathcal{V},\ \pi_p(v) = R\pi_q(v)$$

Proof. First, we show that the map π_q is a bijection from $\mathcal{V}$ to $\pi_q(\mathcal{V})$. Indeed, let $v, w \in \mathcal{V}$ such that $\pi_q(v) = \pi_q(w)$. We have $v - w \in \mathcal{V}$, thus $(v-w)^T J_{p,q}(v-w) \leqslant 0$. This is rewritten as $\|\pi_p(v) - \pi_p(w)\|_2 \leqslant \|\pi_q(v) - \pi_q(w)\|_2 = 0$, hence $\pi_p(v) = \pi_p(w)$, which implies $v = w$. Thus the map π_q is injective. It is also surjective by definition of $\pi_q(\mathcal{V})$. We denote its inverse by π_q^{-1}.

Let π denote the orthogonal projection from $\mathbb{R}^q$ onto $\pi_q(\mathcal{V})$. Then, we define the linear map R from $\mathbb{R}^q$ to $\mathbb{R}^p$ by $R \coloneqq \pi_p \circ \pi_q^{-1} \circ \pi$. By definition, we have $R\pi_q(v) = \pi_p(v)$ for all $v \in \mathcal{V}$. Since the map R is zero on $\pi_q(\mathcal{V})^\perp$, it is sufficient to show that it is a contraction on $\pi_q(\mathcal{V})$. The matrix $J_{p,q}$ is negative definite over $\mathcal{V}$, meaning that $\|\pi_p(v)\|_2 < \|\pi_q(v)\|_2$ when $v \in \mathcal{V}$ is nonzero, which implies that $\|R\pi_q(v)\|_2 = \|\pi_p(v)\|_2 < \|\pi_q(v)\|_2$ holds for all $v \in \mathcal{V}$. □

Then, we solve Problem 2.9 in the easiest case, when the subspaces are $\mathcal{U} = \{0\}$ and $\mathcal{V} = \mathbb{R}x$, for $x \in \mathbb{R}^n$. Since the proposition does not change if $r \neq 0$, we give its statement in the most general case.

Proposition 2.13. *Let $A, B \in \mathcal{S}_n$ and $v \in \mathbb{R}^n$. Then, there exists a minimal upper bound X of A and B such that $Av = Xv$ if and only if $v^T Av > v^T Bv$ or $Av = Bv$.*

Proof. Without loss of generality, we may assume that $B - A = J_{p,q,r}$.
$(\implies)$.
Assume that X is a minimal upper bound of A and B such that $Av = Xv$. The constraint $Av = Xv$ implies that $v^T Av = v^T Xv \leqslant v^T Bv$ holds. We shall thus show that if $v^T Av = v^T Bv$, then $Av = Bv$. Using Theorem 2.3, there is $M \in \mathcal{M}_{p,q}$ such that

$$X = A + \begin{pmatrix} I_p + MM^T & (I_p + MM^T)^{1/2} M \\ M^T (I_p + MM^T)^{1/2} & M^T M \end{pmatrix} \oplus 0_r \,.$$

The condition $Av = Xv$ is rewritten as

$$\phi_{p,q}(M)\pi_q(v) = -\pi_p(v) \,.$$

The function $\phi_{p,q}$ maps the matrix M to an element in the open ball $\mathcal{B}_{p,q}$, so $\|\phi_{p,q}(M)\| < 1$. It follows that $\|\pi_p(v)\|_2 = \|\phi_{p,q}(M)\pi_q(v)\|_2 < \|\pi_q(v)\|_2$ if $\pi_q(v) \neq 0_{q,1}$. However, the assumption $v^T(B-A)v = 0$ implies $\|\pi_p(v)\|_2 = \|\pi_q(v)\|_2$, thus $\pi_q(v) = 0_{q,1}$ and $\pi_p(v) = 0_{p,1}$. We conclude with $(B - A)v = 0_{n,1}$.
$(\impliedby)$.
If $Av = Bv$, then, by a calculation similar to the above, every minimal upper bound satisfies $Av = Xv$. If $v^T(B - A)v < 0$, then $\|\pi_p(v)\|_2 < \|\pi_q(v)\|_2$. It is easily seen that using

$$M = \phi_{p,q}^{-1}\left(\frac{-\pi_p(v)\pi_q(v)^T}{\|\pi_q(v)\|_2{}^2} \right)$$

in the characterization in Theorem 2.3 provides a solution satisfying $Av = Xv$. □

2.4.4 Proof of Theorem 2.10

First, we show that a solution to Problem 2.9 satisfies all three conditions. Given a solution X to Problem 2.9, we have for $u \in \mathcal{U}$, $u^T(B - A)u = u^T(X - A)u - u^T(X - B)u$, where

the first term is non-negative and the second is zero. Hence $B - A$ is non-negative over $\mathcal{U}$. For $v \in \mathcal{V}$, the reverse holds and $B - A$ is non-positive over $\mathcal{V}$. Moreover, if we have $x^T(B - A)x = 0$ for some $x \in \mathcal{U} \cup \mathcal{V}$, then by Proposition 2.13, we have $(B - A)x = 0$ and $x = 0$ as $B - A \in GL_n$. This shows that $B - A$ is positive definite over $\mathcal{U}$ and negative definite over $\mathcal{V}$. Finally, for $u \in \mathcal{U}$ and $v \in \mathcal{V}$, as $\mathcal{U} \subseteq \ker(B - X)$ and $\mathcal{V} \subseteq \ker(A - X)$, we have $u^T(B - A)v = u^T(X - A)v - u^T(X - B)v = 0$, so $\mathcal{U}$ and $\mathcal{V}$ are orthogonal with respect to $B - A$.

Conversely, we show that a matrix satisfying all three conditions provides a solution to Problem 2.9. We will use the characterization (iv) in Theorem 2.3 to build a solution to Problem 2.9. Without loss of generality, we may assume that we work in a basis of $\mathbb{R}^n$ revealing the inertia of $B - A = J_{p,q}$. Furthermore, we may assume that $\dim \mathcal{U} = p$ and $\dim \mathcal{V} = q$. If this is not the case, let $\mathcal{U}_0$ denote a subspace of $\big[(B - A) \cdot \mathcal{U}\big]^\perp$ over which $B - A$ is positive definite that has maximal dimension. Then let $\mathcal{V}_0$ denote a subspace of $\big[(B-A) \cdot \mathcal{V}\big]^\perp \cap \big[(B-A) \cdot (\mathcal{U} \oplus \mathcal{U}_0)\big]^\perp$ over which $B - A$ is negative definite that has maximal dimension. The subspaces $\mathcal{U} \oplus \mathcal{U}_0$ and $\mathcal{V} \oplus \mathcal{V}_0$ then satisfy the assumptions. We will prove that there is a matrix R of size $p \times q$ satisfying:

$$u \in \mathcal{U} \iff \pi_q(u) = -R^T \pi_p(u)\,, \quad v \in \mathcal{V} \iff \pi_p(v) = -R\pi_q(v)\,, \quad \|R\| < 1$$

The proof is done in two steps. First, we build a matrix R satisfying the second and third conditions using Lemma 2.12. Now, since the matrix $J_{p,q}$ is negative definite over $\mathcal{V}$, we have $\pi_q(\mathcal{V}) = \mathbb{R}^q$. Also, for all nonzero $v \in \mathcal{V}$, we have $\|\pi_q(v)\|_2^2 > \|\pi_p(v)\|_2^2 \geqslant 0$, so $\pi_q(v) \neq 0_{q,1}$. Next, we use the orthogonality condition (iii) to show that the first equivalence holds. For $u \in \mathcal{U}$ and $v \in \mathcal{V}$, we have

$$\begin{aligned} \pi_q(v)^T\big(-R^T\pi_p(u) - \pi_q(u)\big) &= \pi_p(u)^T\pi_p(v) - \pi_q(u)^T\pi_q(v) \\ &= u^T J_{p,q} v \\ &= 0\,. \end{aligned}$$

Hence $R^T\pi_p(u) - \pi_q(u) \in \mathbb{R}^q$ is orthogonal to $\pi_q(\mathcal{V}) = \mathbb{R}^q$, and is thus zero. It now suffices to take $M = \phi_{p,q}^{-1}(R)$ to build a solution to Problem 2.9 using (iv) in Theorem 2.3.

We now show that the solutions of Problem 2.9 are parametrized by matrices in the affine subspace $\mathcal{W}$. Let X be a solution of Problem 2.9. According to Theorem 2.3, we can associate with X a unique $M \in \mathcal{M}_{p,q}$. Given vectors $u \in \mathcal{U}$ and $v \in \mathcal{V}$, the constraints $Av = Xv$ and $Bu = Xu$ can be rewritten as

$$\phi_{p,q}(M)^T\pi_p(u) = -\pi_q(u) \;\text{ and }\; \phi_{p,q}(M)\pi_q(v) = -\pi_p(v)\,.$$

Moreover, we have $\|\phi_{p,q}(M)\| < 1$, so that $\phi_{p,q}(M) \in \mathcal{W} \cap \mathcal{B}_{p,q}$.

Conversely, one checks easily that any solution R of (2.2) provides a solution, as long as $R \in \mathcal{B}_{p,q}$. We have shown previously that as soon as the problem is feasible, the set $\mathcal{W} \cap \mathcal{B}_{p,q}$ is nonempty.

2.4.5 Proof of Corollary 2.11

Let $R \in \mathcal{W} \cap \mathcal{B}_{p,q}$. If $\dim \mathcal{U} \neq p$ and $\dim \mathcal{V} \neq q$, since $\dim \pi_p(\mathcal{U}) = \dim \mathcal{U}$ and $\dim \pi_q(\mathcal{V}) = \dim \mathcal{V}$, we can choose nonzero vectors $u_p \in \pi_p(\mathcal{U})^\perp$ and $v_q \in \pi_q(\mathcal{V})^\perp$. The ball $\mathcal{B}_{p,q}$ is an open set, thus for small enough positive ϵ, the matrix $R' := R + \epsilon u_p v_q{}^T$ is also in $\mathcal{B}_{p,q}$ and satisfies

the equations (2.2). The matrix R' produces a different solution than R since $R \neq R'$ and $\phi_{p,q}$ is a bijection, so that $\dim \mathcal{W} \geqslant \dim \pi_p(\mathcal{U})^\perp \times \dim \pi_p(\mathcal{V})^\perp = (p - \dim \mathcal{U})(q - \dim \mathcal{V})$.

If $R, R' \in \mathcal{W}$ are solutions of (2.2), then we have

$$\forall u \in \mathcal{U},\ (R - R')\pi_q(u) = 0 \qquad \forall v \in \mathcal{V},\ \pi_p(v)^T(R - R') = 0$$

which yields the reverse inequality $\dim \mathcal{W} \leqslant \dim \pi_p(\mathcal{U})^\perp \times \dim \pi_p(\mathcal{V})^\perp$. □

2.5 Examples

We recall the definition of quadrics, the equivalence between the inclusion of quadrics and the Löwner order and the algebraic counterpart of tangency between quadrics.

Definition 2.2. We denote by $\mathcal{Q}_A$ the quadric associated with the *symmetric* matrix A, defined by:

$$\mathcal{Q}_A = \{x \in \mathbb{R}^n \colon x^T A x \leqslant 1\}\,.$$

The set $\mathcal{Q}_A$ is convex if and only if the matrix A is positive semidefinite. The set $\mathcal{Q}_A$ is bounded if and only if the matrix A is positive definite. Moreover, it always has a nonempty interior. If the matrix A is positive semidefinite, the inclusion of the quadric $\mathcal{Q}_A$ in the quadric $\mathcal{Q}_B$ is equivalent to the positivity of the matrix $A - B$, meaning that the inclusion of quadrics and the ordering of the corresponding matrices is equivalent, up to reversal:

$$\text{if } A \succcurlyeq 0\,,\ \text{then } \mathcal{Q}_A \subseteq \mathcal{Q}_B \iff B \preccurlyeq A\,.$$

This also means that, given positive definite matrices A, B, the quadric $\mathcal{Q}_X$ associated with a maximal lower bound X of A and B in the Löwner order is a minimal upper bound for the bounded quadrics $\mathcal{Q}_A$ and $\mathcal{Q}_B$, in the inclusion order.

Remark 2.4. In the general case,

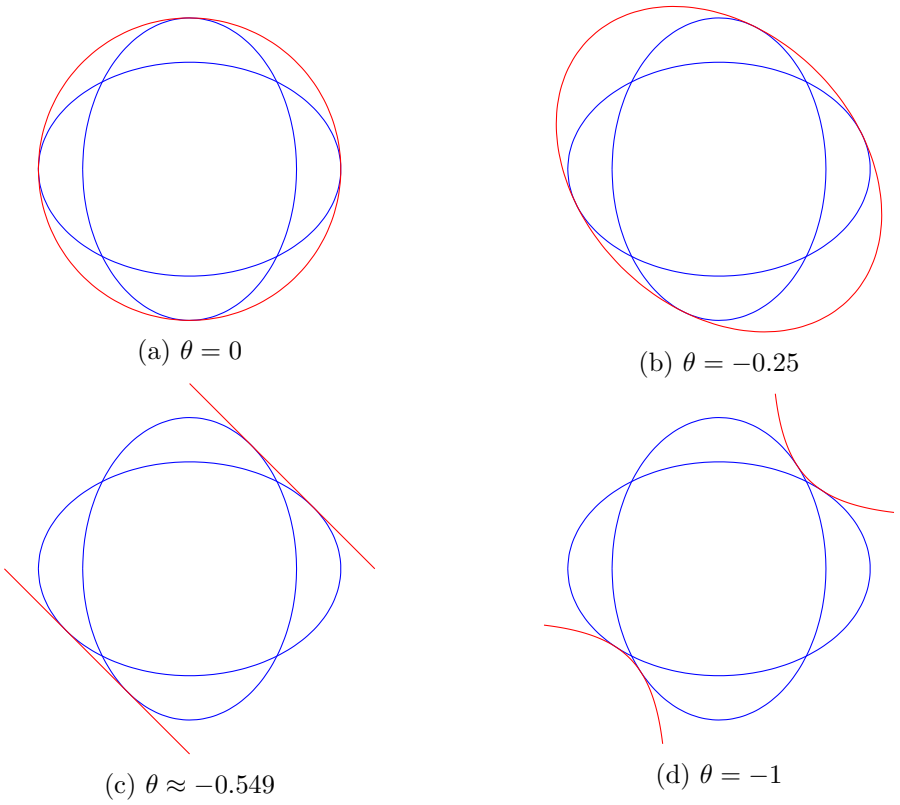
$$\mathcal{Q}_A \subseteq \mathcal{Q}_B \not\Longrightarrow B \preccurlyeq A\,,$$

as shown with $A = 2 \oplus (-2)$ and $B = 1 \oplus (-1)$. For $(x, y) \in \mathcal{Q}_A$, one clearly has $2x^2 - 2y^2 \leqslant 1 \leqslant 2$, which implies $(x, y) \in \mathcal{Q}_B$. However, we have $A - B = 1 \oplus (-1) \not\succcurlyeq 0$.

2.5.1 In dimension 2: $\mathcal{O}(1,1)/\big(\mathcal{O}(1) \times \mathcal{O}(1)\big)$

This case arises whenever two symmetric matrices A and B of order 2 are not comparable. The maps $(X \mapsto X + \lambda I_n)_{\lambda \in \mathbb{R}}$ and $(X \mapsto U^T X U)_{U \in \mathrm{GL}_n}$ are all order-preserving isomorphisms. This implies that, given such an isomorphism ϕ, the set of maximal lower bounds of $\phi(A)$ and $\phi(B)$ is exactly the image of the set of maximal lower bounds of A and B by the map ϕ. Thus one can easily show that we may assume without loss of generality that

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\,.$$



(a) $\theta = 0$ (b) $\theta = -0.25$

(c) $\theta \approx -0.549$ (d) $\theta = -1$

Figure 2.1: Minimal quadrics $\mathcal{Q}_\theta$ (in red) associated with $\mathcal{Q}_A$ and $\mathcal{Q}_B$ (in blue) for various values of θ.

We have an explicit description of the set of hyperbolic isometries $\mathcal{O}(1,1)$ [Gal12, Proposition 9.19]:

$$\mathcal{O}(1,1) = \left\{ \begin{pmatrix} \epsilon_1 \cosh\theta & \epsilon_2 \sinh\theta \\ \epsilon_1 \sinh\theta & \epsilon_2 \cosh\theta \end{pmatrix} : \theta \in \mathbb{R},\, \epsilon_1, \epsilon_2 \in \{-1, 1\} \right\}.$$

The quotient set $\mathcal{O}(1,1)/\big(\mathcal{O}(1) \times \mathcal{O}(1)\big)$ is in this case equal to the classical set of hyperbolic rotations:

$$\mathcal{O}(1,1)/\big(\mathcal{O}(1) \times \mathcal{O}(1)\big) = \left\{ \begin{pmatrix} \cosh\theta & \sinh\theta \\ \sinh\theta & \cosh\theta \end{pmatrix} : \theta \in \mathbb{R} \right\}.$$

Note that in this special case, the quotient set has a group structure. This gives us the parametrization of the maximal lower bounds X_θ of A and B:

$$X_\theta = \begin{pmatrix} 2 - \cosh^2\theta & \cosh\theta \sinh\theta \\ \cosh\theta \sinh\theta & 2 - \cosh^2\theta \end{pmatrix}.$$

The tangency subspace between $\mathcal{Q}_A$ and $\mathcal{Q}_{X_\theta}$ is $\mathbb{R}\begin{pmatrix}\sinh\theta & -\cosh\theta\end{pmatrix}^T$ and the tangency subspace between $\mathcal{Q}_B$ and $\mathcal{Q}_{X_\theta}$ is $\mathbb{R}\begin{pmatrix}\cosh\theta & -\sinh\theta\end{pmatrix}^T$. This is depicted in Figure 2.1.

2.5.2 The quotient Lorentz set: $\mathcal{O}(n,1)/\big(\mathcal{O}(n)\times\mathcal{O}(1)\big),\ n\geqslant 2$

Following Lemma 2.7, the set $\mathcal{O}(n,1)/\big(\mathcal{O}(n)\times\mathcal{O}(1)\big)$ can be identified to $\mathbb{R}^n$ via the bijection $\phi := \phi_{n,1}$ defined by

$$\phi : w \mapsto \begin{pmatrix} (I_n + ww^T)^{1/2} & w \\ w^T & \sqrt{1+w^Tw} \end{pmatrix}.$$

In this case, when $pq = n > 1$, the quotient set does not have a group structure. Let $(e_i)_{1\leqslant i\leqslant n}$ denote the canonical base of $\mathbb{R}^n$. The product $M := \phi(e_1)\phi(e_2)$ can be computed explicitly and it is not even symmetric: we have $M_{2,1} = 0$ whereas $M_{1,2} = 1$.

We shall illustrate the results of Theorem 2.10 on an example with $p = 2$ and $q = 1$, with the matrices $A = 2\oplus 2\oplus 1$ and $B = 1\oplus 1\oplus 2$, so that $B - A = J_{2,1}$. Theorem 2.3 states that the set of maximal lower bounds of A and B, denoted $\bigwedge_{A,B}$, has dimension 2 and its elements X_w are given, for $w\in\mathbb{R}^2$ by

$$X_w = A - \begin{pmatrix} I_2 + ww^T & (I_2+ww^T)^{1/2}w \\ w^T(I_2+ww^T)^{1/2} & w^Tw \end{pmatrix}.$$

For all $w\in\mathbb{R}^2$, we also have $\dim\ker(A - X_w) = 1$ and $\dim\ker(B - X_w) = 2$.

Let $v = (x\ 0\ z)^T$ denote some non-zero vector. We shall solve Problem 2.9 in the cases where $(\mathcal{U},\mathcal{V}) = (\mathbb{R}v,\{0\})$ and $(\mathcal{U},\mathcal{V}) = (\{0\},\mathbb{R}v)$.

Case 1: $\mathcal{U} = \mathbb{R}v$ and $\mathcal{V} = \{0\}$.
In this case, we have $p \neq \dim\mathcal{U}$ and $q \neq \dim\mathcal{V}$, so by Theorem 2.10 the set of solutions is not reduced to a point. The problem has a solution if and only if $x^2 > z^2$, and the solutions are parametrized by the contractive elements of the affine subspace $\mathcal{W}$ of $\mathcal{M}_{2,1}$ defined by $R\in\mathcal{W}$ if and only if $R^T(x\,0)^T + z = 0$. Denoting $r = -z/x$, so that $|r| < 1$, we have

$$\mathcal{W} = \big\{R_t := (r\ t)^T : t\in\mathbb{R}\big\}.$$

Moreover, we have $\|R_t\|^2 = r^2 + t^2$ so that, since $r^2 < 1$, the set $\mathcal{W}\cap\mathcal{B}_{p,q}$ is non-empty. Then, for $|t| < \sqrt{1-r^2}$, we recover the matrix

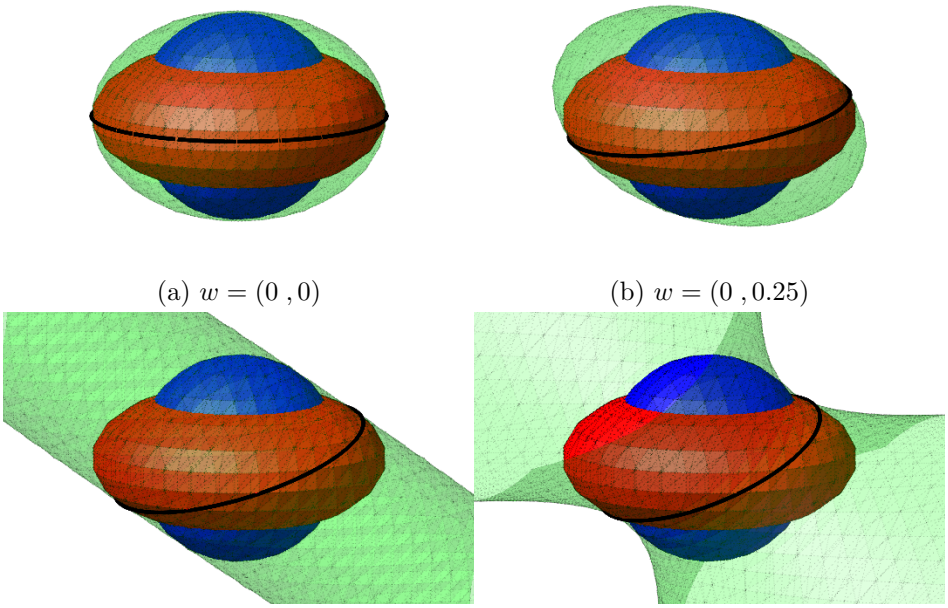
$$w = \phi_{p,q}^{-1}(R_t) = (1 - r^2 - t^2)^{-1/2}\begin{pmatrix} r \\ t \end{pmatrix}.$$

Finally, we get the parametrization of the kernels:

$$\ker(A - X_w) = \operatorname{span}\Big\{\begin{pmatrix} z & -tx & x \end{pmatrix}^T\Big\},$$
$$\ker(B - X_w) = \operatorname{span}\Big\{\begin{pmatrix} x & 0 & z \end{pmatrix}^T,\ \begin{pmatrix} txz & x^2+y^2 & -x^2t \end{pmatrix}^T\Big\}.$$

The set of solutions is parametrized by a single real parameter t as expected from Theorem 2.10.

Case 2: $\mathcal{U} = \{0\}$ and $\mathcal{V} = \mathbb{R}v$.
In this case, we have $q = \dim\mathcal{V}$, so the solution is unique. Indeed, the problem has a solution if and only if $x^2 < z^2$ and the affine subspace $\mathcal{W}$ of $\mathcal{M}_{2,1}$ is reduced to the point $R := (-x/z\,,0)$, which satisfies $\|R\| < 1$. Figure 2.2 depicts several minimal quadrics associated with the quadrics $\mathcal{Q}_A$ and $\mathcal{Q}_B$.



(a) $w = (0\ ,0)$ (b) $w = (0\ ,0.25)$

(c) $w = (0\ ,0.549)$ (d) $w = (0\ ,1)$

Figure 2.2: Minimal quadrics $\mathcal{Q}_{X_w}$ (in green) associated with $\mathcal{Q}_A$ and $\mathcal{Q}_B$ (in blue and red) for various values of w. The black line shows the tangency points between the quadrics $\mathcal{Q}_B$ and $\mathcal{Q}_{X_w}$.

2.5.3 Outer approximation of the union of quadrics

Corollary 2.14. *Let A, B be positive semidefinite matrices. Then*

$$\bigcap_{C \in \bigwedge\{A,B\}} \mathcal{Q}_C = \mathcal{Q}_A \cup \mathcal{Q}_B\,. \tag{2.3}$$

Proof. The inclusion $\mathcal{Q}_A \cup \mathcal{Q}_B \subset \mathcal{Q}_C$ holds for all maximal lower bounds C of A and B, hence $\mathcal{Q}_A \cup \mathcal{Q}_B$ is included in the intersection of all the quadrics $\mathcal{Q}_C$ for $C \in \bigwedge\{A, B\}$. It remains to show that, given some vector $x \in \mathbb{R}^n$, if the inequalities $x^T C x \leqslant 1$ hold for all $C \in \bigwedge\{A, B\}$, then $x \in \mathcal{Q}_A \cup \mathcal{Q}_B$.

Let $x \in \mathbb{R}^n$ such that $x^T C x \leqslant 1$ holds for all $C \in \bigwedge\{A, B\}$. We shall distinguish several cases.

Case 1: $x^T A x \neq x^T B x$

By symmetry, assume that $x^T A x < x^T B x$. Then, as a consequence of Proposition 2.13, there is a maximal lower bound of A and B, denoted C, satisfying $Ax = Cx$. By assumption, we have $x^T C x \leqslant 1$, which combined with the previous equality yields $x^T A x \leqslant 1$, meaning that $x \in \mathcal{Q}_A$.

Case 2.1: $x^T A x = x^T B x$ *and* $Ax = Bx$

Under these assumptions, every maximal lower bound C satisfies $Ax = Cx$ by Theorem 2.3. Given $C \in \bigwedge\{A, B\}$, it follows immediately that $x \in \mathcal{Q}_C$ implies $x \in \mathcal{Q}_A$.

Case 2.2: $x^T A x = x^T B x$ *and* $Ax \neq Bx$

Proposition 2.13 implies that, under these assumptions, there is no maximal lower bound C such that $x^T A x = x^T C x$. However, we shall show that for any positive ϵ, there is a maximal

lower bound C_ϵ such that

$$x^T A x \leqslant x^T C_\epsilon x + \epsilon\,. \tag{2.4}$$

Equation (2.3) is invariant under linear transformation, so that we may assume that $B - A = J_{p,q,r}$. Using the notation of Theorem 2.3 and Section 2.4, it is readily shown that a matrix $M \in \mathcal{M}_{p,q}$ produces a maximal lower bound C_ϵ satisfying Equation (2.4) if and only if $\|(I + MM^T)^{1/2}\pi_p(x) + M\pi_q(x)\|_2^2 \leqslant \epsilon$.

Let s denote some real number such that $0 \leqslant s < 1$. The assumption $x^T A x = x^T B x$ implies that $\|\pi_p(x)\|_2 = \|\pi_q(x)\|_2$. Thus, the matrix R_s given by

$$R_s = -(1-s)\pi_p(x)\pi_q(x)^T / \|\pi_q(x)\|_2^2$$

is in the open ball $\mathcal{B}_{p,q}$. Let M_s denote the matrix $M_s = \phi_{p,q}^{-1}(R_s)$. It is then easily shown that

$$\begin{aligned}(I + M_s M_s^T)^{1/2}\pi_p(x) + M_s\pi_q(x) &= \eta(s)\pi_p(x)\\ \text{with}\quad \eta(s) &= s^{-1/2}\big[(1+2s-s^2)^{1/2} - 1 + s\big]\,,\end{aligned}$$

where η satisfies $\eta(s) = 2\sqrt{s} + o(\sqrt{s})$ for s small enough. As a consequence, for s small enough, using the characterization (iv) in Theorem 2.3, the matrix M_s provides a maximal lower bound C_ϵ satisfying (2.4). □

Remark 2.5. If the approximating quadrics $\{\mathcal{Q}_C \colon C \in \bigwedge\{A, B\}\}$ are required to be convex, meaning that the matrices C are required to be positive semidefinite, then the resulting outer approximation is a convex set, and instead approximates the convex hull of $\mathcal{Q}_A \cup \mathcal{Q}_B$. In this case again, it can be shown that the outer approximation is exact.

Proposition 2.15. *Let A, B be positive semidefinite matrices. Then*

$$\bigcap_{C \in \mathcal{S}_n^+ \cap \bigwedge\{A,B\}} \mathcal{Q}_C = \operatorname{conv}\left(\mathcal{Q}_A \cup \mathcal{Q}_B\right).$$

where $\operatorname{conv}(K)$ *denotes the convex hull of the set K.*

Let K denote the set $\mathcal{Q}_A \cup \mathcal{Q}_B$. The convex hull $\operatorname{conv}(K)$ of the set K is given by the intersection of all half-spaces $\mathcal{H}_{c,T} = \{x \in \mathbb{R}^n \colon c^T x \leqslant T\}$ that contain K. Let $\mathcal{H}_{c,T}$ denote such a half-space. The set K is centrally symmetric, meaning that for all $x \in K$, we have $-x \in K$. This implies that the half-space $\mathcal{H}_{-c,-T}$ also contains K:

$$\forall x \in K,\ -T \leqslant c^T x \leqslant T\,. \tag{2.5}$$

The quantity T is positive given that the set K has nonempty interior. Equation (2.5) describes a quadric $\mathcal{Q}_R$, where $R = T^{-2}\,cc^T$, that contains $\mathcal{Q}_A$ and $\mathcal{Q}_B$. Finally, by definition, there is a maximal lower bound C of A and B such that $\mathcal{Q}_A \cup \mathcal{Q}_B \subseteq \mathcal{Q}_S \subseteq \mathcal{Q}_R$. To summarize, we have shown that we can associate with each half-space containing the set K a quadric $\mathcal{Q}_C$ where C is a maximal lower bound of A and B. Hence, the convex hull of K contains the intersection of convex quadrics associated to maximal lower bounds of A and B:

$$\operatorname{conv}(K) = \bigcap_{\{c,T\colon K \subset \mathcal{H}_{c,T}\}} \mathcal{H}_{c,T} \supseteq \bigcap_{C \in \mathcal{S}_n^+ \cap \bigwedge\{A,B\}} \mathcal{Q}_C\,.$$

The latter set is an intersection of convex sets, thus it is convex. Moreover, it contains K, since $K \subset \mathcal{Q}_C$ for all $C \in \mathcal{S}_n^+ \cap \bigwedge\{A, B\}$. As a consequence, it contains the convex hull of K.

2.6 Partial extension to minimal upper bounds of p matrices

We provide in this section a description of the set of minimal upper bounds of finitely many symmetric matrices in the Löwner order. This result builds on the characterization proved in Theorem 1.15 and the second parametrization given in Section 2.4.

2.6.1 Generalizing the parametrization

Contrary to the case of 2 matrices, the set of minimal upper bounds is not given by a quotient of groups in the general case. However, the description using tangency subspaces remains valid, thus it provides a good entry point for understanding the underlying geometry.

In the 2 matrix case, we have shown that the set of minimal upper bounds can be identified to an open semi-algebraic subset of the Grassmannian manifold. In the more general case, this identification remains in spirit, since the set of minimal upper bounds is given by the reunion of finitely many "parts" $\mathcal{C}_q^{\mathrm{sdp}}$ which are indexed by the vector of tangency-subspace dimensions $q = (q_1, \ldots, q_p)$. Contrary to the former case, these subspaces may not be in direct sum, even if we assume that the difference of any two matrices is not singular. This phenomenon is best represented by "the atom" in Figure 2.3.

The sets $\mathcal{C}_q^{\mathrm{sdp}}$ have a canonical hierarchy. Some correspond to tangency subspaces which are in direct sum. These sets can be identified to open semi-algebraic subsets of a flag on the Grassmannian manifold, and constitute the base of the hierarchy. The other elements are obtained as the intersection of the closures of several base sets.

Theorem 2.16. *Let $\mathcal{A} = \{A_1, \ldots, A_p\}$ denote a finite subset of $\mathcal{S}_n$. The set of minimal upper bounds of $\mathcal{A}$ has the decomposition*

$$\bigvee \mathcal{A} = \biguplus_{q \in \mathcal{Q}} \mathcal{C}_q^{sdp} \quad \text{with} \quad \mathcal{Q} \subset \{q \in \mathbb{N}^p \colon \sum_k q_k \geqslant n\}\,.$$

The set $\mathcal{C}_q^{sdp}$ is the subset of $\bigvee \mathcal{A}$ characterized by

$$X \in \mathcal{C}_q^{sdp} \iff \forall k\,, \dim \ker(X - A_k) = q_k\,.$$

It can be identified to $\mathbb{R}^{d(q)}$ with

$$d(q) \coloneqq \sum_k q_k(n - q_k) - \sum_{k<l} q_k q_l\,.$$

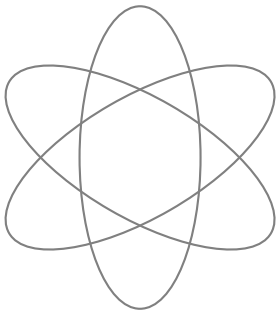
When $\sum_k q_k = n$, we have $d(q) = \sum_{k<l} q_k q_l$.

Moreover, the collection of sets $(\mathcal{C}_q^{sdp})_{q \in \mathcal{Q}}$ follows the hierarchy property

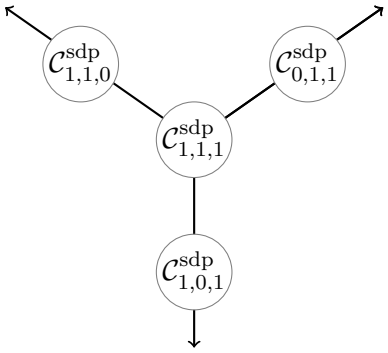
$$\overline{\mathcal{C}_p^{sdp}} \cap \overline{\mathcal{C}_q^{sdp}} \subseteq \bigcup_{r \geqslant p,q} \mathcal{C}_r^{sdp}\,,$$

where $r \geqslant p, q$ means that $r_k \geqslant \max(p_k, q_k)$ for all k and $\overline{\mathcal{X}}$ denotes the closure of $\mathcal{X}$ in the Euclidean topology.

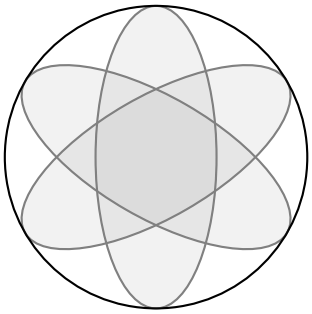
Before proving this theorem in Section 2.6.2, we point out some possible refinements. We believe in particular that the following holds.



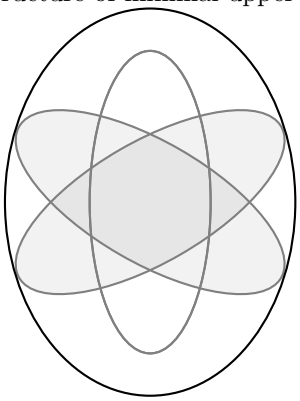
(a) Three mutually uncomparable quadrics



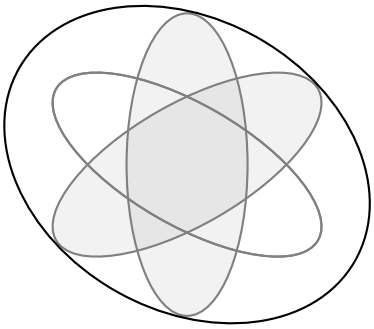
(b) Structure of minimal upper bounds



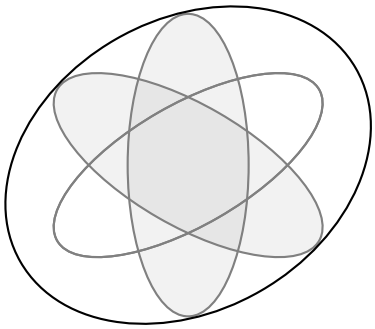
(c) The triple point $\mathcal{C}^{\mathrm{sdp}}_{111}$



(d) An element of the branch $\mathcal{C}^{\mathrm{sdp}}_{110}$



(e) An element of the branch $\mathcal{C}^{\mathrm{sdp}}_{101}$



(f) An element of the branch $\mathcal{C}^{\mathrm{sdp}}_{011}$

Figure 2.3: Minimal upper bounds of 3 quadrics - the “atom”

Conjecture 2.17. *In the previous theorem, equality holds in the hierarchy*

$$\overline{\mathcal{C}_p^{sdp}} \cap \overline{\mathcal{C}_q^{sdp}} = \bigcup_{r \geqslant p,q} \mathcal{C}_r^{sdp}.$$

In contrast with the case of 2 matrices, this description is hardly computationally accessible because obtaining the set of non-empty cells $\mathcal{C}_q^{\mathrm{sdp}}$ is difficult.

Proposition 2.18. *Deciding whether there is $q \in \mathcal{Q}$ such that $q_1 > 0$ is NP-hard.*

Proof. This decision problem is equivalent to finding a nonzero vector x such that $x^T(Q_i - Q_1)x < 0$ for all $i > 1$. This is equivalent to the fact that the optimal value v^* of the following quadratically constrained quadratic program is negative:

$$v^* \coloneqq \min_{v \in \mathbb{R}, \|x\|=1} \{v \colon x^T(Q_i - Q_1)x \leqslant v\,,\ \forall i > 1\}\,.$$

Solving this problem is NP-hard by [MK87]. □

Corollary 2.19. *Deciding the non-emptiness of a set $\mathcal{C}_q^{sdp}$ is NP-hard.*

Proof. At least one coordinate of q must be positive. Checking the positivity of this coordinate is NP-hard by Proposition 2.18. □

2.6.2 Proof of Theorem 2.16

The proof of Theorem 2.16 is split in several parts. First, we prove that the decomposition of $\bigvee \mathcal{A}$ into the reunion of several $\mathcal{C}_q^{\mathrm{sdp}}$ sets. Then, we identify the dimension of the cell $\mathcal{C}_q^{\mathrm{sdp}}$ and derive the homeomorphism to $\mathbb{R}^{d(q)}$ when $\sum_k q_k = n$. Finally, we show the hierarchy property.

Decomposition By Theorem 1.15, any element of $\mathcal{C}_q^{\mathrm{sdp}}$ for $q \in \mathcal{Q}$ is a minimal upper bound of $\mathcal{A}$, hence $\bigvee \mathcal{A} \supseteq \cup_{q \in \mathcal{Q}} \mathcal{C}_q^{\mathrm{sdp}}$. Moreover, by the same corollary, any minimal upper bound must belong to some set $\mathcal{C}_q^{\mathrm{sdp}}$, thus equality holds.

Dimension The key argument is a characterization via tangency subspaces. Let $q \in \mathcal{Q}$ and $X \in \mathcal{C}_q^{\mathrm{sdp}}$. For all k, let $\mathcal{V}_k \coloneqq \ker(X - A_k)$ and denote by W_k a subspace of $\mathcal{V}_k$ such that $\mathbb{R}^n = \oplus_k \mathcal{W}_k$. Without loss of generality, we assume that $\mathcal{W}_1$ is spanned by the first $\dim \mathcal{W}_1$ vectors of the canonical base of $\mathbb{R}^n$, $\mathcal{W}_2$ is spanned by the next $\dim \mathcal{W}_2$ canonical base vectors, and so on. Moreover, let π_k denote the orthogonal projection onto $\mathcal{W}_k$. Then the matrix X is recovered as

$$X = \sum_{i,j} \pi_i^T A_i \pi_j \tag{2.6}$$

Indeed, let us temporarily denote by Y the right hand side of Equation (2.6). Then, by orthogonality of the projectors, we have $(Y - A_k)\pi_k = \sum_i \pi_i^T (A_i - A_k)\pi_k$; and writing $A_i - A_k = (X - A_k) - (X - A_i)$ shows that $\pi_i^T(A_i - A_k)\pi_k = 0$, and we deduce that $(X - Y)\pi_k = 0$ for all k. The sum of the subspaces $\mathcal{W}_k$ is $\mathbb{R}^n$, hence $X = Y$. One can check that Equation (2.6) does not depend on the choice of the subspaces $\mathcal{W}_k$.

We introduce the subset $\mathcal{S}$ of $(\mathrm{GL}_n)^p$ defined by

$$(S_1,\dots,S_p)\in\mathcal{S} \iff \left\{\begin{array}{ll} x^T & = x^T S_i \\ y & = S_i y \\ 0 & = (S_i x)^T (A_i - A_j)(S_j y) \end{array}\right. \quad \text{for all } x\in\mathcal{V}_i\,, y\in\mathcal{V}_j\,, i\neq j\,.$$

The previous reasoning shows that $(I,\dots,I)\in\mathcal{S}$. We count the degrees of freedom, with $q_k = \dim \mathcal{V}_k$. The first and second conditions imply that S_k has only $q_k(n-q_k)$ degrees of freedom. Moreover, the third condition imposes $q_i q_j$ additional equations for all $i<j$. We deduce that the set $\mathcal{S}$ can be identified to a real vector space that has dimension

$$d(q) = \sum_i q_i(n-q_i) - \sum_{i<j} q_i q_j\,.$$

When $\sum_i q_i = n$, this reduces to

$$d(q) = \big(\sum_i q_i\big)^2 - \sum_i q_i^2 - \sum_{i<j} q_i q_j = \sum_{i<j} q_i q_j\,.$$

Hierarchy Let $X\in\overline{\mathcal{C}_p^{\mathrm{sdp}}}\cap\overline{\mathcal{C}_q^{\mathrm{sdp}}}$, meaning that there are sequences $Y_k\in\overline{\mathcal{C}_p^{\mathrm{sdp}}}$ and $Z_k\in\overline{\mathcal{C}_q^{\mathrm{sdp}}}$ converging to X.

By definition of Y_k, the kernel of $Y_k - A_i$ has dimension p_i. In other words, the sequence of vectors containing the $n-p_i+1$ minors of the latter matrix is identically equal to 0. By continuity of the determinant, this also holds for the limit X, hence $\dim\ker(X-A_i)\geqslant p_i$. By symmetry, the same is true for Z, hence $\dim\ker(X-A_i)\geqslant \max(p_i,q_i)$. This also shows by Theorem 1.15 that X is a minimal upper bound of $\mathcal{A}$. Hence $\overline{\mathcal{C}_p^{\mathrm{sdp}}}\cap\overline{\mathcal{C}_q^{\mathrm{sdp}}}\subseteq\bigcup_{r\geqslant p,q}\mathcal{C}_r^{\mathrm{sdp}}$.

CHAPTER 3

A canonical invariant minimal upper bound selection

Given a compact convex set in $\mathbb{R}^n$ with non-empty interior, there is a unique ellipsoid that contains this set and that has minimum volume. This ellipsoid is called the Löwner ellipsoid. It is a fundamental tool in many fields, such as convex geometry, statistics and control theory, see [DLL57, Bal97, Gru11]. An ellipsoid $\mathcal{E}$ can be identified with a positive semidefinite matrix A by $x \in \mathcal{E} \iff xx^T \preccurlyeq A$. The notions of inclusion and volume of ellipsoids are translated into the Löwner order and the determinant of their associated matrices. In particular, the problem of computing the Löwner ellipsoid of a union of ellipsoids can be written as a semidefinite program.

We extend in this chapter the definition of the Löwner ellipsoid to all proper convex cones Ω by replacing the volume/determinant considerations by the *characteristic function* φ associated with the cone Ω.

Our main result (Theorem 3.4) states that there is a unique vector in the cone Ω which minimizes the value of the characteristic function over the set of upper bounds of a finite collection of vectors. Moreover, it states that it is the only minimal upper bound that selects itself as an exposed point (see Theorem 1.10) via a canonical bijection between the cone and its dual, provided the cone satisfies some geometric conditions. Finally, we prove that this minimal upper bound selection has desirable invariance properties with respect to order-preserving transformations.

Our result holds in particular for the non-exceptional symmetric cones. When applied to the Euclidean Lorentz cone, we obtain a seemingly new minimal upper bound selection and exhibit its properties. We use this new point of view in the case of positive semidefinite matrices to prove a matrix inequality showing that the "Löwner ellipsoid of p matrices" is dominated by the sum of these matrices in the Löwner order.

The results in Sections 3.6.2 and 3.6.4 have appeared in [AGG+15]. Further results are newer.

3.1 Notations and definitions

3.1.1 Automorphisms of convex cones

Let Ω denote a *open convex pointed cone* in a vector space E so that $E = \Omega - \Omega$. The space E is equipped with a scalar product $\langle \cdot, \cdot \rangle$. The closed cone $\overline{\Omega}$ defines an order relation on the vector space E, denoted by $\preccurlyeq$. We also denote by Ω^* the *open dual cone* of the open convex cone Ω, defined by

$$y \in \Omega^* \iff \langle y, x \rangle > 0\,,\ \forall x \in \overline{\Omega} \setminus \{0\}\,.$$

The cone Ω^* is the interior of the dual cone of $\overline{\Omega}$. We have the identity $(\Omega^*)^* = \Omega$. Moreover, we denote by $\mathrm{Aut}(\Omega)$ the set of automorphisms of E that stabilize the open cone Ω:

$$\mathrm{Aut}(\Omega) \coloneqq \{g \in \mathrm{GL}(E) \colon g(\Omega) = \Omega\}\,.$$

Since the space E is finite dimensional, every such g is continuous and thus $g \in \mathrm{Aut}(\Omega)$ if and only if $g(\overline{\Omega}) = \overline{\Omega}$. We refer to the set $\mathrm{Aut}(\Omega)$ as the set of automorphisms of Ω. As in [FK94], for $x \in \Omega$ and $g \in \mathrm{Aut}(\Omega)$, we use the short notation

$$g \cdot x \coloneqq g(x)\,.$$

We say that the cone Ω is *homogeneous* if the group $\mathrm{Aut}(\Omega)$ acts transitively on Ω, that is for all $x, y \in \Omega$, there is $g \in \mathrm{Aut}(\Omega)$ such that $y = g \cdot x$.

The adjoint of the automorphism g, denoted by g^*, is defined by

$$\langle x, g \cdot y \rangle = \langle g^* \cdot x, y \rangle\,,\ \forall x, y \in E\,.$$

The next lemma justifies the common notation of the adjoint and the open dual cone.

Lemma 3.1 (see [FK94, Proposition I.1.7])**.** *For any open convex pointed cone Ω, we have*

$$g \in \mathrm{Aut}(\Omega) \iff g^* \in \mathrm{Aut}(\Omega^*)\,.$$

3.1.1.a Automorphisms of the Euclidean Lorentz cone The group of automorphisms of the Euclidean Lorentz cone is generated by rotations around its "central axis" and the Lorentz transformations. The first kind of linear transformation is given by

$$\begin{pmatrix} t \\ x \end{pmatrix} \mapsto \begin{pmatrix} t \\ Ux \end{pmatrix} \text{ with } U \in \mathcal{O}(\mathbb{R}^n)\,.$$

The Lorentz transformations are given by the map parametrized by the angle θ:

$$\begin{pmatrix} t \\ x_1 \\ x_2 \\ \vdots \end{pmatrix} \mapsto \begin{pmatrix} t\cosh\theta + x_1 \sinh\theta \\ t\sinh\theta + x_1 \cosh\theta \\ x_2 \\ \vdots \end{pmatrix}\,.$$

The Euclidean Lorentz cone is homogeneous. We refer to [FK94, Chapter 1.2] for the proof.

3.1.1.b Automorphisms of the cone of positive definite matrices The group of automorphisms of the cone of positive definite matrices $\mathcal{S}_n^{++}$ is given by the maps

$$\Gamma_M \coloneqq X \mapsto MXM^T \text{ with } M \in \mathrm{GL}_n \,.$$

This group acts transitively on $\mathcal{S}_n^{++}$: let A, B denote two positive definite matrices, then the matrix $M \coloneqq B^{1/2}A^{-1/2}$ satisfies $MAM^T = B$.

Remark 3.1. We point out that, even though the map $X \mapsto X^{-1}$ is a bijection from $\mathcal{S}_n^{++}$ into itself, it is not considered an automorphism in the sense of [FK94], since it is not an automorphism of $E = \mathcal{S}_n^{+}$.

3.1.2 Characteristic function

Following the definition in [FK94, Chapter 1, Section 3], the *characteristic function* φ of the open cone Ω is the map defined for $x \in \Omega$ by

$$\varphi(x) \coloneqq \int_{\Omega^*} e^{-\langle y,x\rangle}\, dy \,.$$

Faraut and Korányi show in [FK94, Proposition I.3.1 and Proposition I.3.3] that the characteristic function φ is analytic over Ω, satisfies $\varphi(gx) = (\det g)^{-1}\varphi(x)$ and that the map $x \mapsto \log\varphi(x)$ is strictly convex. Moreover, given a point $x \in \Omega$, the point $\sigma_0(x) \in \Omega^*$ is defined by

$$\sigma_0(x) \coloneqq -\nabla \log\varphi(x) \,.$$

Proposition 3.2 (see [FK94], Proposition I.1.4)**.** *Let Ω denote a homogeneous cone. Then the map $\sigma_0(\cdot)$ is a bijection between Ω and Ω^*. Moreover, it satisfies*

1. *$\varphi(x)\varphi(\sigma_0(x))$ is constant,*
2. *$\sigma_0(g \cdot x) = \theta(g) \cdot \sigma_0(x)$ with $\theta(g) = (g^*)^{-1}$,*
3. *$\langle \sigma_0(x), x\rangle = n$ for all $x \in \Omega$,*
4. *$\sigma_0(\sigma_0(x)) = x$ for all $x \in \Omega$,*
5. *there is a unique $e_0 \in \Omega$ such that $\sigma_0(e_0) = e_0$.*

When the cone Ω is homogeneous, every vector $x \in \Omega$ is the unique fixed point of an automorphism in $\mathrm{Aut}(\Omega)$. Indeed, if the cone Ω is homogeneous, there is $h \in \mathrm{Aut}(\Omega)$ such that $x = h \cdot e_0$. We can then define the automorphism

$$\sigma_x \colon z \mapsto h \cdot \sigma_0(h^{-1} \cdot z)$$

and it is readily checked that x is indeed a fixed point. The fact that σ_0 has a unique fixed point ensures that it is also the case for σ_x.

Remark 3.2 (Lorentz Cone). As shown in [FK94, Chapter 1.4], the characteristic function is

$$\varphi\big[(t\ x)^T\big] = (t^2 - \|x\|^2)^{-n/2}$$

and the involution $\sigma_0(\cdot)$ is given by

$$\sigma_0(\,t\ x\,)^T = \frac{n}{t^2 - \|x\|^2}(\,t\ -x\,)^T \,.$$

Its fixed point is the vector $(\,\sqrt{n}\ 0\,)^T$.

Remark 3.3 (Cone of positive definite matrices). As shown in [FK94, Chapter 1.4], the characteristic function is

$$\varphi(X) = (\det X)^{-(n+1)/2}$$

and the involution $\sigma_0(\cdot)$ is given by

$$\sigma_0(X) = \frac{n+1}{2} X^{-1}$$

and its fixed point is the vector $\sqrt{\frac{n+1}{2}} I_n$.

3.1.3 Selections of minimal upper bounds

Let $\mathcal{D}$ denote the set $\mathcal{D} \coloneqq \wp(\overline{\Omega}) \times \Omega$. The value $\Phi(\mathcal{A}, c)$ is defined for $\mathcal{A} \subset \overline{\Omega}$ and $c \in \Omega$ by

$$\Phi(\mathcal{A}, c) \coloneqq \arg\min_{x \succcurlyeq \mathcal{A}} \langle \sigma(c), x \rangle \,. \tag{3.1}$$

This defines a map Φ from $\mathcal{D}$ to $\overline{\Omega}$. By Theorem 1.10, this map takes non-empty values when the set $\mathcal{A}$ is finite and it is a multivalued selection of minimal upper bounds. Indeed, given a finite subset $\mathcal{A} \subset \overline{\Omega}$ and $c \in \Omega$, every value $y \in \Phi(\mathcal{A}, c)$ is a minimal upper bound of $\mathcal{A}$. In other words, we have $\Phi(\mathcal{A}, c) \subset \bigvee \mathcal{A}$. By Theorem 1.10, every minimal upper bound arises in this way, so that the stronger equality holds:

$$\bigvee \mathcal{A} = \bigcup_{c \in \Omega} \Phi(\mathcal{A}, c) \,.$$

When $x \in \Phi(\mathcal{A}, c)$, we say that x is a minimal upper bound of $\mathcal{A}$ that is *selected by* c.

The map Φ can be extended [1] to the set $\mathcal{D}'$ defined by

$$\mathcal{D}' \coloneqq \bigcup_{\text{face } F \text{ of } \overline{\Omega}} \wp(F) \times \operatorname{rel\,int} F \,.$$

where $\operatorname{rel\,int} F$ denotes the *relative interior* of the face F, i.e. the interior of the cone F with respect to the induced topology in the vector space spanned by F.

Indeed, we have shown in Theorem 1.10 that minimal upper bounds of a set $\mathcal{A}$ belong to the extreme face $\mathcal{F}(\mathcal{A})$ and can be selected by elements of its open dual cone $\mathcal{F}(\mathcal{A})^* \cap V$, with $V = \operatorname{span} \mathcal{F}(\mathcal{A})$. The extension is defined as follows. The cone $\mathcal{F}(\mathcal{A})$ has a characteristic function $\varphi_{\mathcal{F}(\mathcal{A})}$ which can be used to defined a bijection $\sigma_{\mathcal{F}(\mathcal{A})}$ from the interior of $\mathcal{F}(\mathcal{A})$ onto its open dual cone in V given by $\mathcal{F}(\mathcal{A})^* \cap V$. We extend the map Φ on $\wp\big(\mathcal{F}(\mathcal{A})\big) \times \big(\operatorname{int} \mathcal{F}(A)\big)$ by replacing the involution σ by $\sigma_{\mathcal{F}(\mathcal{A})}$ in Equation (3.1).

For the sake of readability, we write $\Phi_{\mathcal{A}}(\cdot)$ to mean the partial map $x \mapsto \Phi(\mathcal{A}, x)$.

3.2 The main results

We are interested in vectors x such that $x \in \Phi(\mathcal{A}, x)$, i.e. minimal upper bound x that select themselves through the involution σ. Vectors x which satisfy $x \in \Phi_{\mathcal{A}}(x)$ are then called *fixed points* of the map $\Phi_{\mathcal{A}}$.

[1] We have $\mathcal{D} \subseteq \mathcal{D}'$ since $\overline{\Omega}$ is a face of itself and $\operatorname{rel\,int} \overline{\Omega} = \Omega$.

3.2.1 Two technical assumptions

Our main results characterize the fixed points of the map $\Phi_{\mathcal{A}}$ under two technical assumptions on the cone Ω.

Assumption 3.1. *The cone Ω is* homogeneous, *i.e. for all $x, y \in \Omega$, there is $g \in \mathrm{Aut}(\Omega)$ such that $y = g \cdot x$.*

Assumption 3.2. *There is $e \in \Omega$ such that $\langle x, x\rangle = \langle e, x\rangle^2$ for all $x \in \mathrm{Extr}(\overline{\Omega})$. We denote the automorphism $\sigma_e(\cdot)$ by $\sigma(\cdot)$ for short.*

Since automorphisms $g \in \mathrm{Aut}(\Omega)$ are monotone, the image of an extreme ray is also an extreme ray. Thus a consequence of Assumption 3.1 is the invariance and homogeneity of the set of extreme rays by $\mathrm{Aut}(\Omega)$.

Assumption 3.2 implies that the set of extreme rays is a subset of the boundary of the Lorentz cone $\{x \in E\colon \langle x, x\rangle^{1/2} \leqslant \langle e, x\rangle\}$.

Assumption 3.1 and Assumption 3.2 also imply that the vector e acts as an axis of rotation for the cone Ω, whose group of rotations is exactly $\mathrm{Aut}(\Omega) \cap \mathcal{O}(E)$. In particular,the set of extreme rays must remain invariant under the action of the group $\mathrm{Aut}(\Omega) \cap \mathcal{O}(E)$, where $\mathcal{O}(E)$ is the orthogonal group of the space E. When the cone $\overline{\Omega}$ has only finitely many extreme rays $\mathbb{R}x$, it implies for instance in dimension $n = 3$ that sections of the cone taken orthogonally to the vector e are regular polygons. When the cone $\overline{\Omega}$ has infinitely many extreme rays $\mathbb{R}x$, sections of the cone are spheres.

Proposition 3.3. *Let Ω denote a cone satisfying Assumptions 3.1 and 3.2. Then the vector e is a* nonlinear eigenvector *of the involution $\sigma_0(\cdot)$: there is a positive λ such that $\sigma_0(e) = \lambda e$. Moreover, the vector e is a common fixed point of every automorphism in* $\mathrm{Aut}(\Omega) \cap \mathcal{O}(E)$.

Proof. Let $\mathcal{H}$ denote the affine hyperplane defined by $x \in \mathcal{H} \iff \langle e, x\rangle = 1$. By Assumptions 3.1 and 3.2, the group $\mathrm{Aut}(\Omega) \cap \mathcal{O}(E)$ acts transitively on the set $\mathrm{Extr}(\overline{\Omega}) \cap \mathcal{H}$. Hence, given two extreme rays $x, y \in \mathcal{H}$, there is $g \in \mathrm{Aut}(\Omega) \cap \mathcal{O}(E)$ such that $y = g \cdot x$ and we have

$$\langle g^* \cdot e, x\rangle = \langle e, g \cdot x\rangle = \|y\| = \|x\| = \langle e, x\rangle\,.$$

Hence $\langle g^* \cdot e - e, x\rangle = 0$ holds for all $g \in \mathrm{Aut}(\Omega) \cap \mathcal{O}(E)$ and $x \in \mathrm{Extr}(\overline{\Omega})$. The cone $\overline{\Omega}$ has nonempty interior, thus it must have at least n linearly independent extreme rays (this can be deduced from the Krein-Millman theorem). We obtain that $g^* \cdot e = e$ for all orthogonal automorphisms g. Such an automorphism g satisfies $gg^* = \mathrm{id}_E$, hence $g \cdot e = e$.

Moreover, the characteristic function φ is invariant by an automorphism $g \in \mathrm{Aut}(\Omega) \cap \mathcal{O}(E)$: $\varphi(g \cdot x) = \varphi(x)$ for all $x \in \Omega$. Given $u \in \mathrm{Extr}(\overline{\Omega}) \cap \mathcal{H}$, we have $\varphi(e + tg \cdot (u - e)) = \varphi(e + t(u - e))$ for all $g \in \mathrm{Aut}(\Omega) \cap \mathcal{O}(E)$. By homogeneity, the gradient of the map $\log \varphi$ must satisfy

$$\langle \nabla \log \varphi(e), u - v\rangle = 0\,, \text{ for all } u, v \in \mathrm{Extr}\,\overline{\Omega} \cap \mathcal{H}\,.$$

Again, the set of extreme rays of $\overline{\Omega}$ in $\mathcal{H}$ must span the subspace $\mathcal{H}$, thus $\nabla \log \varphi$ is orthogonal to $\mathcal{H}$: there is $\lambda > 0$ such that $\sigma_0(e) = \lambda e$ by definition of σ_0. □

Remark 3.4. Proposition 3.3 implies that the results stated for the map σ_0 in Proposition 3.2 also hold for the involution σ_e.

Remark 3.5. By Lemma 1.5, the set of extreme rays of $\overline{\Omega}$ are the non-negative multiplies of the set of extreme points of $\{x \in \overline{\Omega} \colon \langle e, x\rangle = 1\}$. This set is homogeneous under the action of $\mathrm{Aut}(\Omega) \cap \mathcal{O}(E)$ which is a compact group, thus the former set is compact too. We deduce that the set of extreme half-lines $\mathbb{R}x$ also form a compact set.

We have already shown in Section 3.1.1 that both the open Euclidean Lorentz cone $\mathrm{int}\,\mathcal{L}_n^2$ and the cone of positive definite matrices $\mathcal{S}_n^{++}$ are homogeneous, so Assumption 3.1 holds for these cones. We point out that Assumption 3.2 also holds for these cones.

Remark 3.6 (Lorentz Cone). Recall that an extreme ray of the Lorentz cone is of the form $u = \begin{pmatrix} \|x\| & x \end{pmatrix}^T$. We deduce that

$$\langle u, u\rangle = \left\langle \begin{pmatrix} \|x\| \\ x \end{pmatrix}, \begin{pmatrix} \|x\| \\ x \end{pmatrix} \right\rangle = 2\|x\|^2 = \left\langle \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix}, \begin{pmatrix} \|x\| \\ x \end{pmatrix} \right\rangle^2 = \langle e, u\rangle^2 ,$$

with $e \coloneqq \begin{pmatrix} \sqrt{2} & 0 \end{pmatrix}^T$. Finally, we have

$$\sigma\Big[\begin{pmatrix} t \\ x \end{pmatrix} \Big] \coloneqq \frac{2}{t^2 - \|x\|^2} \begin{pmatrix} t \\ -x \end{pmatrix} \quad \text{and} \quad \sigma(e) = e .$$

Remark 3.7 (Cone of positive definite matrices). Recall that an extreme ray of the cone of positive semidefinite matrices is of the form $u = xx^T$. We deduce that

$$\langle u, u\rangle = \mathrm{trace}(xx^T xx^T) = (x^T x)^2 = \mathrm{trace}(xx^T)^2 = \langle I_n, xx^T\rangle^2 = \langle e, u\rangle^2$$

with $e = I_n$. The involution $\sigma(\cdot)$ then take the form

$$\sigma(X) \coloneqq X^{-1} .$$

3.2.2 Statement of the theorems

The first theorem shows that a vector $x \in \Omega$ is a fixed point of the map $\Phi_{\mathcal{A}}$ if and only its image by the involution σ is an optimal solution to a semi-infinite conic programming problem.

Theorem 3.4. *Let Ω be a cone satisfying Assumptions 3.1 and 3.2. Given a finite set $\mathcal{A} \subset \overline{\Omega}$ such that $\mathcal{F}(\mathcal{A}) = \overline{\Omega}$ and a vector $x \in \Omega$, the following assertions are equivalent:*

(i) $x \in \Phi(\mathcal{A}, x)$,

(ii) $\sigma(x) \in \arg\min \big\{ \log\varphi(c) \colon c \in \Omega^* ,\ \langle c, u\rangle \leqslant 1 ,\ \forall u \in \mathcal{U} \big\}$

where $\mathcal{U} \coloneqq \cup_{a\in\mathcal{A}} \{u \in \mathrm{Extr}(\overline{\Omega}) \colon 0 \preccurlyeq u \preccurlyeq a\}$.

The map φ acts as a measurement of "how far inside" the cone Ω a given vector x is, since it tends to infinity when x tends to the boundary of the cone (see [FK94, Proposition I.3.2]). The smaller the value of $\varphi(x)$, the more the vector x is "inside the cone Ω". Theorem 3.4 then states that a minimal upper bound x for which the vector $\sigma(x)$ is "most inside the cone Ω" among all minimal upper bounds of $\mathcal{A}$ is a fixed point of $\Phi_{\mathcal{A}}$.

Moreover, the map $\log\varphi$ is strictly convex. Hence the conic optimization problem in Assertion (ii) above has a unique solution, which implies the following theorem.

Theorem 3.5. *Let Ω be a cone satisfying Assumptions 3.1 and 3.2. Given a finite set $\mathcal{A} \subset \overline{\Omega}$ such that $\mathcal{F}(\mathcal{A}) = \overline{\Omega}$, the map $x \mapsto \Phi(\mathcal{A}, x)$ has a unique fixed point in Ω.*

By Theorem 1.10, minimal upper bounds of a set $\mathcal{A}$ belong to the extreme face $\mathcal{F}(\mathcal{A})$ and the map $\Phi_{\mathcal{A}}$ sends $\mathcal{F}(\mathcal{A})$ into $\mathcal{F}(\mathcal{A})$. We deduce from Theorem 3.5 the following corollary.

Corollary 3.6. *Let Ω be a cone satisfying Assumptions 3.1 and 3.2. Given a finite set $\mathcal{A} \subset \overline{\Omega}$, the map $x \mapsto \Phi(\mathcal{A}, x)$ has a unique fixed point in $\mathcal{F}(\mathcal{A})$.*

3.2.3 Discussion and conjectures

Theorem 3.4 gives the reader the path that has been followed in order to obtain uniqueness of the fixed point of the map $\Phi_{\mathcal{A}}$ which is central to the analysis in Sections 3.4 to 3.6. We give a precise characterization of such fixed points as well as a computational method to compute it, via a semi-infinite conic problem.

A standard method for proving the uniqueness of the fixed point is to show that the considered map is a strict contraction with respect to some metric. Despite several attempts, this approach has failed us, mostly due to the absence of an *explicit* formula for the minimal upper bound selected by a given vector c. In the case of positive semidefinite matrices, we have an explicit formula for the minimal upper bound x in terms of the matrices $a_k \in \mathcal{A}$ and the selecting vector $c \in \Omega^*$, but it also requires the data of the Lagrange multipliers λ_k associated with the "tangency constraints". It is

$$x = c^{-1} \sum_k \lambda_k a_k \quad \text{with} \quad \sum_k \lambda_k = c\,.$$

In the case of 2 positive semidefinite matrices, we can use Theorem 2.3 to show that the map $\Phi_{\mathcal{A}}$ is indeed a contraction with contraction rate $\frac{1}{2}$ in Thompson's metric. Moreover, we have observed the same contraction rate on tens of thousands of numerical simulations, for sets ranging from 2 to 20 matrices and all dimensions lower than 20. We have done tests of similar scale for the Euclidean Lorentz cone, with the same positive result. This motivates us to formulate the following conjecture:

Conjecture 3.7. *Let Ω be a cone satisfying Assumptions 3.1 and 3.2. Then the map $\Phi_{\mathcal{A}}$ is a contraction in Thompson's metric with contraction rate $\frac{1}{2}$:*

$$d_T\big[\Phi_{\mathcal{A}}(c), \Phi_{\mathcal{A}}(d)\big] \leqslant \frac{1}{2} d_T(c, d)\,,$$

where Thompson's metric d_T is defined by

$$d_T(x, y) = \inf\{\lambda > 0 \colon \lambda^{-1} y \preccurlyeq x \preccurlyeq \lambda y\}\,.$$

Finally, the results of Section 3.2.2 remain true when the open cone Ω is the image of a cone satisfying Assumptions 3.1 and 3.2 by an invertible linear map L. Indeed, it suffices to replace the scalar product $\langle \cdot, \cdot \rangle$ by $(x, y) \mapsto \langle L^{-1}x, L^{-1}y \rangle$ to preserve the validity of the two assumptions. Moreover, the automorphism group of $L \cdot \Omega$ is simply given by $\{LgL^{-1} \colon g \in \mathrm{Aut}(\Omega)\}$ and the cone $L \cdot \Omega$ is homogeneous if and only if the cone Ω is homogeneous.

Hence Theorems 3.4 and 3.5 and Corollary 3.6 remain valid if Assumption 3.2 is replaced by the weaker assumption:

Assumption 4.2'. *There is $e \in \Omega$ and an invertible linear map L such that $\langle Lx, Lx \rangle = \langle e, Lx \rangle^2$ for all $x \in \mathrm{Extr}(\overline{\Omega})$.*

3.3 Proof of Theorem 3.4

3.3.1 Two conic optimization problems

We introduce two conic optimization problems, corresponding to Assertions (i) and (ii), and derive their dual counterpart.

Problem corresponding to Assertion (i) For $x \in \Omega$, the primal conic optimization problem ($\mathcal{P}_{\text{fix}}(x)$) associated with Assertion (i) is:

$$\begin{aligned} &\text{minimize } \langle \sigma(x), y\rangle \\ &\text{subject to } \quad y \succcurlyeq \mathcal{A}\,. \end{aligned} \tag{$\mathcal{P}_{\text{fix}}(x)$}$$

The Lagrangian, denoted by L_{fix} is

$$L_{\text{fix}}(y,\lambda) = \langle \sigma(x), y\rangle - \sum_{a\in\mathcal{A}} \langle \lambda_a, y-a\rangle\,,$$

with $\lambda_a \in \overline{\Omega^*}$. The dual counterpart ($\mathcal{D}_{\text{fix}}(x)$) is hence classically [BTN01] given by

$$\begin{aligned} &\text{maximize } \sum_{a\in\mathcal{A}} \langle \lambda_a, a\rangle \\ &\text{subject to } \quad \sum_{a\in\mathcal{A}} \lambda_a \preccurlyeq \sigma(x)\ ,\ \lambda_a \in \overline{\Omega^*}\ . \end{aligned} \tag{$\mathcal{D}_{\text{fix}}(x)$}$$

Both problems are clearly strictly feasible, thus by [Bar] the duality gap is zero and there is a primal dual optimal solution (y,λ). By complementary slackness, such an optimal solution is characterized by $\langle \lambda_a, y-a\rangle = 0$ for all $a \in \mathcal{A}$.

Problem corresponding to Assertion (ii) The set $\mathcal{U} := \cup_{a\in\mathcal{A}}\{u \in \text{Extr}(\overline{\Omega})\colon 0 \preccurlyeq u \preccurlyeq a\}$ is bounded and closed by Remark 3.5 hence it is compact. Moreover, by the assumption $\mathcal{F}(\mathcal{A}) = \overline{\Omega}$, we have $\mathcal{F}(\mathcal{U}) = \overline{\Omega}$, hence $c \in \Omega$ if and only if $\langle c, u\rangle > 0$ for all $u \in \mathcal{U}$.

Let L_φ denote the Lagrangian

$$L_\varphi(c,\nu) = \log\varphi(c) - \int_{\mathcal{U}} (1 - \langle c, u\rangle)\, d\nu(u)\,,$$

where $c \in \Omega^*$ and ν is a non-negative measure on $\mathcal{U}$.

The function L_φ is strictly convex in c and Ω^* is convex. Moreover, given any measure ν on $\mathcal{U}$ with positive weight and $c \in \Omega^*$, the value $L_\varphi(tc,\nu)$ tends to $+\infty$ when t tends to 0 or $+\infty$, or as c tends to the boundary of the cone [FK94]. Finally, given $c \in \Omega^*$ whose norm is large enough, there is $u \in \mathcal{U}$ such that $\langle c, u\rangle > 1$. Then the value $L_\varphi(c, t\delta_u)$ tends to $-\infty$ when t tends to $+\infty$. By Proposition 2.2, p.173 in [ET99], the function L_φ has a saddle point and we have

$$\max_\nu \inf_c L_\varphi(c,\nu) = \inf_c \max_\nu L_\varphi(c,\nu)\,. \tag{3.2}$$

Maximizing the quantity $L_\varphi(c,\nu)$ in the variable ν is equivalent to minimizing $\int_{\mathcal{U}}(1 - \langle c, u\rangle)\, d\nu(u)$. This value is finite if and only if $\langle c, u\rangle \leqslant 1$ for all $u \in \mathcal{U}$. Hence the left-hand

side of Equation (3.2) yields the primal conic optimization problem ($\mathcal{P}_\varphi(\mathcal{A})$) associated with Assertion (ii), up to a logarithm:

$$\begin{aligned} &\text{minimize } \log\varphi(c) && (\mathcal{P}_\varphi(\mathcal{A}))\\ &\text{subject to } \quad c \in \Omega^* \\ &\langle c, u\rangle \leqslant 1\ ,\ \forall u \in \mathcal{U}\,, \end{aligned}$$

We minimize the quantity $L_\varphi(c,\nu)$ in the variable c by differentiating in c. We obtain $\sigma_0(c) = \int_{\mathcal{U}} u\, d\nu(u)$, so by Proposition 3.2

$$L_\varphi\big(c(\nu), \nu_1, \nu_2\big) = \log\varphi \circ \sigma_0\Big[\int_{\mathcal{U}} u\, d\nu(u)\Big] + n - \int_{\mathcal{U}} d\nu(u)\,.$$

Moreover, the measure ν must have positive finite weight, otherwise the dual cost function is $-\infty$. We write $\nu = w\,\nu_0$ with $w = \nu(\mathcal{U})$, so ν_0 is a probability measure. We can rewrite the dual objective function, up to an additive constant, as $\log(w) - w - \log\varphi\Big[\int_{\mathcal{U}} u\, d\nu_0(u)\Big]$. The variables ν_0 and w are independent and maximizing in w yields $w = 1$, thus $\nu = \nu_0$ is a probability measure. The right-hand side of Equation (3.2) yields the dual conic problem ($\mathcal{D}_\varphi(\mathcal{A})$)

$$\begin{aligned} &\text{maximize } \ -\log\varphi\Big[\int_{\mathcal{U}} u\, d\nu(u)\Big] && (\mathcal{D}_\varphi(\mathcal{A}))\\ &\text{subject to } \quad \nu \text{ probability measure on } \mathcal{U}\,. \end{aligned}$$

By complementary slackness, a point (c,ν) is an optimal solution of and only if $\sigma(c) = \int_{\mathcal{U}} u\, d\nu(u)$ and

$$\operatorname{supp}\nu \subseteq \{u \in \mathcal{U} \colon \langle c, u\rangle = 1\}\,.$$

We show that x is an optimal primal solution of Problem ($\mathcal{P}_{\text{fix}}(x)$) if and only if $\sigma(x)$ is an optimal primal solution of Problem ($\mathcal{P}_\varphi(\mathcal{A})$).

Let us first show a technical lemma:

Lemma 3.8. *Let Ω be a cone satisfying Assumptions 3.1 and 3.2. For all $u \in \operatorname{Extr}(\overline{\Omega})$ and $b \in \Omega$, the inequality $\langle\sigma(b), u\rangle \leqslant 1$ holds if and only if $u \preccurlyeq b$.*

Moreover, if $u \preccurlyeq a \preccurlyeq b$ with $a \in \overline{\Omega}$ and $b = g\cdot e$ with $g \in \operatorname{Aut}(\Omega)$, then

$$\langle\sigma(b), u\rangle = 1 \implies \langle (gg^*)^{-1}\cdot u, b - a\rangle = 0\,.$$

Proof. We assume that $\langle\sigma(b), u\rangle \leqslant 1$. Let $v \in \operatorname{Extr}(\overline{\Omega})$. We have the sequence of inequalities

$$\begin{aligned} \langle v, u\rangle = \langle g^*\cdot v, g^{-1}\cdot u\rangle &\leqslant \|g^*\cdot v\|\|g^{-1}\cdot u\| = \langle e, g^*\cdot v\rangle\langle e, g^{-1}\cdot u\rangle \\ &= \langle g\cdot e, v\rangle\langle\theta(g)\cdot e, u\rangle = \langle g\cdot e, v\rangle\langle\sigma(b), u\rangle \leqslant \langle g\cdot e, v\rangle\,. \end{aligned}$$

This holds for all $v \in \operatorname{Extr}(\overline{\Omega})$, thus $u \preccurlyeq g\cdot e = b$.

Conversely, we have the sequence of inequalities:

$$\begin{aligned} \langle\sigma(b), u\rangle^2 = \langle\theta(g)\cdot e, u\rangle^2 = \langle e, g^{-1}\cdot u\rangle^2 &= \|g^{-1}\cdot u\|^2 \\ = \langle u, (gg^*)^{-1}\cdot u\rangle \leqslant \langle a, (gg^*)^{-1}\cdot u\rangle &\leqslant \langle b, (gg^*)^{-1}\cdot u\rangle \\ &= \langle g\cdot e, (gg^*)^{-1}\cdot u\rangle = \langle\sigma(b), u\rangle\,. \end{aligned}$$

We deduce that $\langle\sigma(b), u\rangle \leqslant 1$. If $\langle\sigma(b), u\rangle = 1$, all inequalities must be equalities, which implies in turn that $\langle (gg^*)^{-1}\cdot u, b - a\rangle = 0$. □

3.3.2 Assertion (ii) $\implies$ Assertion (i)

Let $(\sigma(x), \nu)$ denote an optimal solution of the pair of primal-dual optimizations problems $(\mathcal{P}_\varphi(\mathcal{A}))$ and $(\mathcal{D}_\varphi(\mathcal{A}))$, with $x \in \Omega$, i.e., ν is a probability measure,

$$\operatorname{supp}\nu \subseteq \mathcal{V} \coloneqq \mathcal{U} \cap \{u \in \operatorname{Extr}(\overline{\Omega})\colon \langle \sigma(x), u\rangle = 1\} \quad \text{and} \quad \int_{\mathcal{U}} u\, d\nu(u) = x\,.$$

First, we show that $x \succcurlyeq \mathcal{A}$. By definition of $\sigma(x)$ as a primal feasible point, we have

$$u \in \mathcal{U} \implies \langle \sigma(x), u\rangle \leqslant 1\,.$$

Thus, Lemmas 1.8 and 3.8 imply that $a \preccurlyeq x$ for all $a \in \mathcal{A}$.

Since the cone Ω is homogeneous, let $g \in \operatorname{Aut}(\Omega)$ such that $x = g \cdot e$. By Lemma 3.8, we partition the set $\mathcal{V}$ into finitely many measurable sets $\mathcal{V}(a)$:

$$\mathcal{V} = \cup_{a\in\mathcal{A}}\mathcal{V}(a)\,, \quad a \neq b \implies \mathcal{V}(a) \cap \mathcal{V}(b) = \emptyset\,, \quad u \in \mathcal{V}(a) \implies \langle (gg^*)^{-1} \cdot u, x - a\rangle = 0\,.$$

We define the collection $\{\mu_a\}_{a\in\mathcal{A}}$ by

$$\mu_a \coloneqq (gg^*)^{-1} \cdot \int_{\mathcal{V}(a)} u\, d\nu(u)\,.$$

The collection μ is a feasible point for $(\mathcal{D}_{\text{fix}}(x))$ since

$$\sum_{a\in\mathcal{A}} \mu_a = (gg^*)^{-1} \cdot \int_{\mathcal{V}} u\, d\nu(u) = (gg^*)^{-1} \cdot x = \sigma(x)\,.$$

By definition of the set $\mathcal{V}(a)$, we must have

$$\sum_{a\in\mathcal{A}} \langle \mu_a, x - a\rangle = \sum_{a\in\mathcal{A}} \int_{\mathcal{V}(a)} \langle (gg^*)^{-1} \cdot u, x - a\rangle\, d\nu(u) = 0\,.$$

Hence, the pair (x, μ) is a primal-dual optimal solution of Problems $(\mathcal{P}_{\text{fix}}(x))$ and $(\mathcal{D}_{\text{fix}}(x))$.

3.3.3 Assertion (i) $\implies$ Assertion (ii)

Conversely, let x denote a fixed point of the map $\Phi_{\mathcal{A}}$, so that $x \succcurlyeq \mathcal{A}$ and there is a collection of vectors $\lambda_a \in \overline{\Omega^*}$ such that $\sum_{a\in\mathcal{A}} \lambda_a = \sigma(x)$ and $\sum_{a\in\mathcal{A}} \langle \lambda_a, x - a\rangle = 0$.

Let $a \in \mathcal{A}$ and $u \in \operatorname{Extr}(\overline{\Omega})$ such that $u \preccurlyeq a$. Since $a \preccurlyeq \sigma(x)$, we can apply Lemma 3.8 and deduce that $\langle \sigma(x), u\rangle \leqslant 1$. Thus $\sigma(x)$ is a feasible point of Problem $(\mathcal{P}_\varphi(\mathcal{A}))$.

Given $a \in \mathcal{A}$, the vector $(gg^*) \cdot \lambda_a$ belongs to the closed cone $\overline{\Omega}$, thus it can be written as the sum of finitely many extreme rays:

$$(gg^*) \cdot \lambda_a = \sum_k \alpha_k(a) u_k(a)\,,$$

. We introduce the measure ν' defined by

$$\nu' \coloneqq \sum_{a\in\mathcal{A}} \sum_k \alpha_k(a) \langle \sigma(x), u_k(a)\rangle\, \delta\Big[\frac{1}{\langle \sigma(x), u_k(a)\rangle} u_k(a)\Big]\,,$$

where $\delta[u]$ denote the Dirac measure at the point u. By construction, this measure satisfies

$$\int_{\mathcal{U}} u \, d\nu'(u) = \sum_{a\in\mathcal{A}} \sum_{k} \alpha_k(a) u_k(a) = (gg^*) \cdot \sum_{a\in\mathcal{A}} \lambda_a = (gg^*) \cdot \sigma(x) = x \,.$$

In other words, we have $\int_{\mathcal{U}} u \, d\nu'(u) = \sigma(c)$ with $c \coloneqq \sigma(x)$. Moreover, the support of ν' is included in the set $\mathcal{U} \cap \{u \in \operatorname{Extr} \overline{\Omega} \colon \langle \sigma(x), u\rangle = 1\}$. Hence the pair $(\sigma(x), \nu')$ is a primal-dual optimal solution of Problems $(\mathcal{P}_\varphi(\mathcal{A}))$ and $(\mathcal{D}_\varphi(\mathcal{A}))$. This concludes the proof of Theorem 3.4.

3.4 Invariant minimal upper bound selection

In this section, we assume that the cone Ω satisfies Assumptions 3.1 and 3.2.

3.4.1 Commutation and uniqueness

As a consequence of Theorem 3.5, we can define a new minimal upper bound selection of the set $\mathcal{A}$, denoted by $\sqcup\mathcal{A}$, and defined by

$$\sqcup\mathcal{A} \coloneqq \text{ the unique fixed point of the map } \Phi_{\mathcal{A}} \,.$$

We say that this selection is "invariant" because it commutes with the action of an automorphism of the cone Ω:

Proposition 3.9. *Let $\mathcal{A}$ denote a finite subset of $\overline{\Omega}$ and g denote an automorphism of the cone Ω that leave the extreme face $\mathcal{F}(\mathcal{A})$ invariant. Then we have*

$$\sqcup\big(g \cdot \mathcal{A}\big) = g \cdot \big(\sqcup \mathcal{A}\big) \,.$$

Proof of Proposition 3.9. By Corollary 3.6, we can assume without loss of generality that $\overline{\Omega} = \mathcal{F}(\mathcal{A})$. Let $x = \sqcup\mathcal{A}$ and $y = \Phi(g\cdot\mathcal{A}, g\cdot x)$. Then by definition of the map Φ and Theorem 3.5, the vector y must be the unique minimizer of

$$\text{minimize } \langle \sigma(g\cdot x), z\rangle \quad \text{subject to } z \succcurlyeq g\cdot\mathcal{A} \,.$$

Writing $z' \coloneqq g^{-1}\cdot z$, the vector $y' \coloneqq g^{-1}\cdot y$ must be the unique minimizer of

$$\text{minimize } \langle \sigma(x), z'\rangle \quad \text{subject to } z' \succcurlyeq \mathcal{A} \,.$$

The unique minimizer of this quantity is x. Hence $y' = x$, and $g\cdot x = \Phi(\mathcal{A}, g\cdot x)$. □

We warn the reader that, in the general case, this is not the only selection that commutes with the action of automorphisms of the cone Ω. Indeed, for finite sets $\mathcal{A} \subset \Omega$ and positive scalars λ_a we can define a selection process T_λ by:

$$T_\lambda(\mathcal{A}) \coloneqq \Phi\big(\mathcal{A}, \sum_{a\in\mathcal{A}} \lambda_a a\big) \,.$$

The same proof as above shows that the map T_λ commutes with all automorphisms of Ω. Moreover, it is easy to find two vectors λ, μ such that the maps T_λ and T_μ are different. For instance, one may check that this is the case when $\Omega = \mathcal{S}_n^{++}$ and $\operatorname{card}\mathcal{A} \geqslant 3$. However, it is

the *only* selection that commutes with the automorphisms of Ω in the case where $\Omega = \mathcal{S}_n^{++}$ and $\operatorname{card} \mathcal{A} = 2$ by Theorem 3.13. The latter theorem is the reason why the map $\sqcup$ is called an *invariant selection of minimal upper bounds*: the selection does not depend on the choice of a basis, whether it is orthogonal or not. We also call this selection the *invariant join*, by borrowing the terminology from lattice theory.

We conjecture that this property also holds for the Euclidean Lorentz cone, and more generally for all cones that satisfy Assumptions 3.1 and 3.2.

Conjecture 3.10. *Let Ω denote an open cone that satisfies Assumptions 3.1 and 3.2. Then the restriction of $\sqcup$ to $\Omega \times \Omega$ is the only minimal upper bound selection that commutes with the action of* $\operatorname{Aut}(\Omega)$.

3.5 Application: the Euclidean Lorentz cone

In this section, we restrict our analysis to the Euclidean Lorentz cone Λ_n, and we will drop the term "Euclidean" for easy reading.

There is one specific minimal upper bound selection in the Lorentz cone that is considered in the literature for practical applications. It is given by the "minimal penumbra", as described in [BK13].

Coming back to the geometric interpretation of Lorentz cones given in Section 1.3.2, an element $\hat{x} \coloneqq (t\ x)$ of the Lorentz cone can be identified to a (Euclidean) ball $B(x, t)$ centered at x with radius t. This ball must contain the point 0 since $(t\ x) \in \Lambda_n$ if and only if $\|0 - x\| \leqslant t$.

A selection of minimal upper bounds that would arise naturally in an over-approximation context is the "minimal radius selection", meaning that we would like to select the minimal upper bound of a union of balls which has the smallest radius. The radius of the ball associated with the vector $\hat{x}$ is obtained by taking the scalar product of $\hat{x}$ with the vector $(1\ 0)$, which belongs to the interior of the Lorentz cone. Hence, by Corollary 1.13, we obtain this minimal upper bound by solving a second-order conic program. Given a finite set $\mathcal{A} = \{\hat{a}_i\}_{i \in I} \subset \Lambda_n$, the "minimal radius selection" $\hat{x}$ is the unique optimal solution of

$$\begin{aligned} &\text{minimize } \langle (1\ 0), \hat{x} \rangle \\ &\text{subject to } \ \hat{x} - \hat{a}_i \in \Lambda_n \ , \forall i \in I \end{aligned}$$

We illustrate this selection on Figure 3.1.

This selection is useful in a synchronization context, where the cone Λ_3 is the *light cone* from special relativity. We illustrate our interpretation on an example. Let $\mathcal{A} = \{\binom{t_i}{x_i}\}_{i \in I} \in \Lambda_n$ which represents a collection of observers, each located at x_i and emitting a light signal at time t_i. The smallest radius minimal upper bound $\binom{t}{x}$ gives the time t and position x of the object which is detected by all observers at the earliest time possible. Dually, the *maximal lower bound with largest radius* gives the latest time t and the position x of a lamp which would light up the combined field of view of all the observers in $\mathcal{A}$.

In most cases, this selection does not yield a fixed point of the map $\Phi_{\mathcal{A}}$. Recall that the group of automorphisms of the Lorentz cone is generated by rotations around the main axis

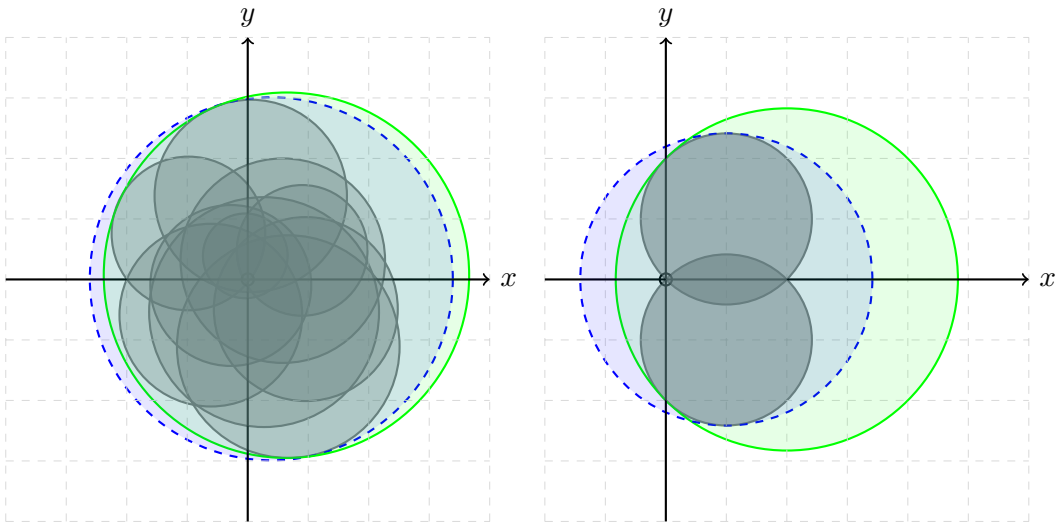


Figure 3.1: Computation of the "minimum radius" selection (blue and dashed) and invariant selection (green and plain)

$(t\ x)^T \mapsto (t\ Ux)^T$ with $U \in \mathcal{O}(n)$ and the Lorentz transformations with parameter θ:

$$\begin{pmatrix} t \\ x_1 \\ x_2 \\ \vdots \end{pmatrix} \mapsto \begin{pmatrix} t\cosh\theta + x_1 \sinh\theta \\ t\sinh\theta + x_1 \cosh\theta \\ x_2 \\ \vdots \end{pmatrix}.$$

The minimal radius selection does commute with the former rotations, which is a very natural property to require, since we expect the selection to be independent of any relativistic reference frame. However, it does not commute with the Lorentz transformations.

The latter property is satisfied by the selection $\sqcup$, which does not aim to minimize the radius t of the enclosing ball, but instead minimizes the quantity $\varphi(\hat{x}) = (t^2 - \|x\|^2)^{-1}$, and thus maximizes the quantity $t^2 - \|x\|^2$.

Moreover, it does by Proposition 3.9 commute with the Lorentz transformations. We show on Figure 3.1 computations of the *invariant selection*. Although there are many instances in which the latter selection is "close" to the minimum volume one, we show in Figure 3.1 an instance where these selections take very different values.

In the later instances, the invariant selection yields a coarse approximation of the union of the balls. This is of course a side-effect of the representation method, i.e. the balls containing 0, rather than a blame of the intrinsic quality of the selection, which depends on the application. If the volume of the ball is the criteria, the first selection is clearly optimal. However, if Lorentz transformations play a role in the application, it may appear that the second selection is optimal.

3.6 Application: the cone of positive semidefinite matrices

3.6.1 Positive semidefinite matrices and ellipsoids

Like the case of the Lorentz cone, we can provide a geometric interpretation of elements in the cone $\mathcal{S}_n^+$. Given a matrix $A \in \mathcal{S}_n^+$, we define the set of vectors $\mathcal{E}_A$ by

$$x \in \mathcal{E}_A \iff xx^T \preccurlyeq A\,.$$

This is an ellipsoid, i.e. the deformation of a euclidean ball by a linear transformation, as shown by the following lemma.

Lemma 3.11. *The image of the unit ball $\mathcal{B}(0,1)$ under the linear map $x \mapsto Mx$ is equal to the set $\mathcal{E}_A$, where $A = MM^T$.*

Proof. We first remark that z belongs to the ball $\mathcal{B}(0,1)$ if and only if $zz^T \preccurlyeq I$. Indeed, by definition of the Löwner order, the latter property amounts to $(z^Ty)^2 \leqslant y^Ty$ for all $y \in \mathbb{R}^n$. Now, assuming that $z^Tz \leqslant 1$, we get by the Cauchy-Schwarz inequality:

$$(z^Ty)^2 \leqslant (z^Tz)(y^Ty) \leqslant y^Ty\,.$$

Reciprocally, if $(z^Ty)^2 \leqslant y^Ty$ for all $y \in \mathbb{R}^n$, then taking $y = z$ provides $(z^Tz)^2 \leqslant z^Tz$. If $z \neq 0$, we obtain that $z \in \mathcal{B}(0,1)$, and this is still true if $z = 0$. This proves the expected equivalence.

Now, let $\mathcal{E}$ be the ellipsoid given by the image of $\mathcal{B}(0,1)$ under the map $x \mapsto Mx$. Let $x \in \mathcal{E}$, and $z \in \mathcal{B}(0,1)$ satisfying $x = Mz$. As $zz^T \preccurlyeq I$, we have:

$$xx^T = M(zz^T)M^T \preccurlyeq MM^T = A\,.$$

This shows $\mathcal{E} \subseteq \mathcal{E}_A$. Reciprocally, if $x \in \mathcal{E}_A$, it can be shown than x belongs to the range of A, meaning that there exists y such that $x = Ay$, as the range of the matrix M is equal to the range of the matrix A. We introduce $z := M^Ty$, so that $x = Mz$. Since $xx^T \preccurlyeq A$, we know that $(y^Tx)^2 \leqslant y^TAy = y^Tx$, hence $y^Tx \leqslant 1$. Now observe that $y^Tx = y^TMM^Ty = z^Tz$. We deduce that $z \in \mathcal{B}(0,1)$. □

Remark 3.8. By Lemma 3.8, we know that $xx^T \preccurlyeq A$ is equivalent to $x^TA^{-1}x \leqslant 1$ when A is positive definite. In this sense, the definition of ellipsoids we give here is dual to the definition of a quadric given in Section 2.5. There, we were interested in unbounded ellipsoids and non-convex overapproximating quadrics, but that representation did not allow for "flat" ellipsoids, i.e. they had to be full dimensional. The present setting allows for "flat" ellipsoids $\mathcal{E}_A$ whenever the matrix A does not have full rank.

The present representation is compatible with the Löwner order $\preccurlyeq$, since $\mathcal{E}_A \subseteq \mathcal{E}_B$ is equivalent to $A \preccurlyeq B$ whenever, $A, B \in \mathcal{S}_n^+$. Moreover, the volume of the ellipsoid $\mathcal{E}_A$ is proportional to $\sqrt{\det A}$. By [FK94, Chapter 1.3], the characteristic function φ of the cone $\mathcal{S}_n^{++}$ is given by $\log \varphi(X) = -\frac{m+1}{2} \log\det X + \log \varphi(I_n)$. Hence, by Proposition 3.2 and Theorem 3.4, the invariant join is obtained by selecting the ellipsoid $\mathcal{E}_X$ that contains $\cup_{A\in\mathcal{A}}\mathcal{E}_A$ which has the smallest volume. We recover in this way the definition of the *Löwner ellipsoid*:

Theorem 3.12. *Given a finite set $\mathcal{A} \subset \mathcal{S}_n^+$ such that $\sum_{A\in\mathcal{A}} A$ is positive definite, the ellipsoid $\mathcal{E}_{\sqcup\mathcal{A}}$ coincides with the Löwner ellipsoid of the set $\cup_{A\in\mathcal{A}}\mathcal{E}_A$:*

$$\mathcal{E}_{\sqcup\mathcal{A}} = \text{ the unique ellipsoid containing } \cup_{A\in\mathcal{A}} \mathcal{E}_A \text{ that has smallest volume}\,.$$

This correspondence gives another proof of the invariance of this selection, since the volume of the ellipsoids $\mathcal{E}_{MAM^T}$ and $\mathcal{E}_A$ only differ by the constant factor $|\det M|$.

3.6.2 The unique invariant selection

A third argument for the canonicity of the selection $\sqcup$, besides it begin the only fixed point of $\Phi_{\mathcal{A}}$ and the fact that it commutes with invertible congruences, is the fact that it is *the only selection* that commutes with invertible congruences when the set $\mathcal{A}$ contains two elements.

Theorem 3.13. *The map $\sqcup$ restricted to $\mathcal{S}_n^{++} \times \mathcal{S}_n^{++}$ is the only selection of a minimal upper bound on that set that commutes with the action of invertible congruences.*

Proof. Let $A, B \in \mathcal{S}_n^{++}$, P an invertible matrix P and a diagonal matrix D such that $A = PP^T$ and $B = PDP^T$. We recall how these matrices are obtained: let UDU^T be the eigenfactorization of the matrix $A^{-1/2}BA^{-1/2}$; then $P \coloneqq A^{1/2}U$ is an invertible matrix that satisfies the desired equalities.

Let $\bigtriangledown$ denote a selection of minimal upper bounds that commutes with congruences. In particular, it must commute with the map $X \mapsto PXP^T$, thus it is only necessary to consider the value of $I \bigtriangledown D$.

Furthermore, let S_i denote the symmetry with respect to the hyperplane $e_i^\perp$, given by $S_i(x) = x - 2\langle e_i, x\rangle e_i$. In the canonical basis of $\mathbb{R}^n$, its matrix, denoted again by S_i, is diagonal, with a 1 in each entry, except the i-th diagonal entry which contains a -1. This map is invertible, thus $\bigtriangledown$ must commute with the congruence $X \mapsto S_i X S_i^T$. Since the matrices I, D are diagonal, this amounts to $(I \bigtriangledown D)_{ij} = 0$ whenever $i \neq j$, hence $I \bigtriangledown D$ is a diagonal matrix. By minimality, we must have $(I \bigtriangledown D)_{ii} = \max(1, D_{ii})$. Hence there is only one selection that commutes with invertible congruences. □

3.6.3 Invariant join of shorted matrices

3.6.3.a A nontrivial shorted matrices equality We recall that the *short* of a positive semidefinite matrix A with respect to the subspace $\mathcal{V}$ is given by

$$\operatorname{short}(A, \mathcal{V}) \coloneqq \max\{X \succcurlyeq 0 \colon X \preccurlyeq A \text{ and } \operatorname{ran} X \subseteq \mathcal{V}\}\,.$$

It is jointly monotone [And99], meaning that $A \preccurlyeq B$ and $\mathcal{U} \subset \mathcal{V}$ imply that $\operatorname{short}(A,\mathcal{U}) \preccurlyeq \operatorname{short}(B, \mathcal{V})$. Chapter 2 introduced the generalized short that satisfies $[B]A = \operatorname{short}(A, \operatorname{ran} B)$.

We also recall that by Lemma 1.7, a positive semidefinite matrix A is equal to the supremum of all rank 1 matrices that it dominates, i.e. $\bigvee\{xx^T \colon xx^T \preccurlyeq \mathcal{A}\} = \{A\}$. A similar equality holds for the short of a matrix.

Lemma 3.14. *Given a positive semidefinite matrix A and a subspace $\mathcal{V}$, we have*

$$\bigvee\{xx^T \colon xx^T \preccurlyeq A\,,\ \ x \in \mathcal{V}\} = \{\operatorname{short}(A, \mathcal{V})\}\,.$$

Proof. For simplicity, we denote by $\mathcal{M}$ the set $\mathcal{M} \coloneqq \bigvee\{xx^T \colon xx^T \preccurlyeq A\,,\ \ x \in \mathcal{V}\}$ and S the matrix $S \coloneqq \operatorname{short}(A, \mathcal{V})$. By definition of the short operator, there must be $M \in \mathcal{M}$ such that $M \preccurlyeq S$.

Let $M \in \mathcal{M}$ and assume that $M \succcurlyeq S$ does not hold, so that there must be a nonzero vector x such that $xx^T \preccurlyeq S$ holds but $xx^T \preccurlyeq M$ does not, by a similar argument to the proof

of Lemma 1.7. By definition of S, the vector x satisfies $x \in \mathcal{V}$ and $xx^T \preccurlyeq A$, which contradicts the fact that $M \in \mathcal{M}$. Hence $M \succcurlyeq S$.

We deduce that $S \in \mathcal{M}$ and $\mathcal{M} \succcurlyeq S$, thus $\mathcal{M} = \{S\}$ by definition of minimal upper bounds. □

A consequence of Lemma 3.14 is that, given a positive semidefinite matrix A and a positive definite matrix B such that $A \preccurlyeq B$, the shorts $\operatorname{short}(A, \mathcal{V})$ and $\operatorname{short}(B, \mathcal{V})$ coincide if and only if $xx^T \preccurlyeq B \implies xx^T \preccurlyeq A$ for all $x \in \mathcal{V}$. By Lemma 3.8, the fact that $xx^T \preccurlyeq B$ is equivalent to $x^T B^{-1} x \leqslant 1$, so that the latter equivalence holds in particular for all vectors $x \in \mathcal{V}$ such that $x^T B^{-1} x = 1$. We deduce that

$$1 = (B^{-1}x)^T \, xx^T \, (B^{-1}x) \preccurlyeq (B^{-1}x)^T \, A \, (B^{-1}x) \preccurlyeq x^T B^{-1} x = 1 \,,$$

so that $B^{-1}x \in \ker(B - A)$ since $A \preccurlyeq B$. This proves the following lemma:

Lemma 3.15. *Given two symmetric matrices A, B such that $A \succcurlyeq 0$ and $B \succ 0$ and a subspace $\mathcal{V}$, the shorts of the matrices A and B by the subspace $\mathcal{V}$ coincide if and only if $\mathcal{V} \subseteq B \cdot \ker(B - A)$.*

3.6.3.b The non-commutative analogue of a classical inequality By Theorem 1.15, the matrix X is a minimal upper bound of $\mathcal{A} = \{A_k\}_{1 \leqslant k \leqslant p}$ if and only if $X \succcurlyeq \mathcal{A}$ and $\sum_k \ker(X - A_k) = \mathbb{R}^n$.

We provide another remarkable property of the invariant join. In the general case, the data of the subspaces $\ker(X - A_k)$ and the set $\mathcal{A}$ fully determines the minimal upper bound X by Theorem 1.15. It turns out that in the case of the *invariant join*, it is only necessary to have access to the subspaces $\ker(X - A_k)$ and the values of the matrices A_k shorted by the subspace $A_k \cdot \ker(X - A_k)$, which coincides with the tangency subspace of the ellipsoids $\mathcal{E}_{A_k}$ and $\mathcal{E}_X$.

Theorem 3.16. *Let $\mathcal{A}$ denote a finite subset of $\mathcal{S}_n^+$ such that $\sum_k A_k$ is positive definite. Let also $\mathcal{V}_k \coloneqq A_k \cdot \ker(\sqcup\mathcal{A} - A_k)$. Then*

$$\sqcup\mathcal{A} = \sqcup_k \operatorname{short}(A_k, \mathcal{V}_k) \,.$$

Proof. First, we denote the set of shorts by $\mathcal{B} \coloneqq \{\operatorname{short}(A_k, \mathcal{V}_k)\}_k$. We point out that the subspace $\mathcal{V}_k$ can also be written as $(\sqcup\mathcal{A}) \cdot \ker(\sqcup\mathcal{A} - A_k)$. Note that $\sum_k A_k \succ 0$ implies that $\mathcal{F}(\mathcal{A}) = \mathcal{S}_n^+$, hence $\sqcup\mathcal{A}$ is positive definite. We specialize the notation introduce for general cones in Section 3.3 to the semidefinite cone. Let $\mathcal{U}(\mathcal{A})$ denote the set $\cup_k \{x \in \mathbb{R}^n \colon xx^T \preccurlyeq A_k\}$, and $(\mathcal{P}_\varphi(\mathcal{A}))$, $(\mathcal{D}_\varphi(\mathcal{A}))$ denote the pair of primal-dual optimization problems corresponding to minimizing $-\log\det C$ over $\mathcal{U}(\mathcal{A})$:

$$\begin{array}{ll} \text{minimize} \;\; -\log\det C & (\mathcal{P}_\varphi(\mathcal{A})) \\ \text{subject to} \;\; C \succ 0 & \\ x^T C x \leqslant 1 \,, \; \forall x \in \mathcal{U}(\mathcal{A}) \,, & \end{array} \qquad \begin{array}{ll} \text{maximize} \;\; -\log\det \Big[\displaystyle\int_{\mathcal{U}(\mathcal{A})} xx^T \, d\nu(x) \Big] & (\mathcal{D}_\varphi(\mathcal{A})) \\ \text{subject to} \;\; \nu \text{ non-negative measure on } \mathcal{U}(\mathcal{A}) \,. & \end{array}$$

Recall that a primal dual solution (C, ν) is an optimal solution of and only if

$$C^{-1} = \int_{\mathcal{U}(\mathcal{A})} xx^T \, d\nu(x) \quad \text{and} \quad \operatorname{supp}\nu \subseteq \mathcal{U}(\mathcal{A}) \cap \{x \in \mathbb{R}^n \colon x^T C^{-1} x = 1\} \,,$$

and that $C^{-1} = \sqcup\mathcal{A}$. We define $\mathcal{U}(\mathcal{B})$, $\big(\mathcal{P}_\varphi(\mathcal{B})\big)$ and $\big(\mathcal{D}_\varphi(\mathcal{B})\big)$ in the same way.

First, the inclusion $\mathcal{U}(\mathcal{B}) \subseteq \mathcal{U}(\mathcal{A})$ holds, since by definition of the short operator, we have $\operatorname{short}(A_k, \mathcal{V}_k) \preccurlyeq A_k$. Thus, the feasible set of $(\mathcal{P}_\varphi(\mathcal{A}))$ is contained in the one of $\big(\mathcal{P}_\varphi(\mathcal{B})\big)$. In particular, the matrix C satisfies $y^T C y \leqslant 1$ for all $y \in \mathcal{U}(\mathcal{B})$.

By Lemma 3.8, given $x \in \mathcal{U}(\mathcal{A})$ such that $xx^T \preccurlyeq A_k$ and $x^T C x = 1$, we must have $Cx \in \ker(C^{-1} - A_k)$, i.e. $x \in \mathcal{V}_k$. By the choice of x, we have $xx^T \preccurlyeq A_k$. Hence, by definition of the short operator, we must have $xx^T \preccurlyeq \operatorname{short}(A_k, \mathcal{V}_k)$. Hence $x \in \mathcal{U}(\mathcal{B})$.

We deduce that

$$\mathcal{U}(\mathcal{A}) \cap \{x \in \mathbb{R}^n \colon x^T C x = 1\} = \mathcal{U}(\mathcal{B}) \cap \{x \in \mathbb{R}^n \colon x^T C x = 1\}\,. \tag{3.3}$$

It then follows that the optimal dual solution of $(\mathcal{D}_\varphi(\mathcal{A}))$ is also optimal for $\big(\mathcal{D}_\varphi(\mathcal{B})\big)$, which shows that $\sqcup\mathcal{B} = \sqcup\mathcal{A}$. □

A consequence of Theorem 3.16 is that the value of the invariant join depends only on the values of the linear maps associated with the matrices $\mathcal{A}$ with respect to their "tangency space": the information contained elsewhere does not influence the latter value. The following theorem shows that, when these tangency subspace are in direct sum, then the invariant join is in fact obtained as the sum of the shorted matrices. This is a witness of the fact that the shorted matrices contain no more information than is needed to compute the invariant join. The result of Theorem 3.16 also allows us to prove the analogue of the classical inequality on non-negative reals $\max_k x_k \leqslant \sum_k x_k$.

Theorem 3.17. *Let $\mathcal{A}$ denote a finite subset of $\mathcal{S}_n^+$ such that $\sum_k A_k$ is positive definite. Let $\mathcal{V}_k \coloneqq A_k \cdot \ker(\sqcup\mathcal{A} - A_k)$. Then*

$$\sqcup\mathcal{A} \preccurlyeq \sum_k \operatorname{short}(A_k, \mathcal{V}_k)\,,$$

with equality if $\mathbb{R}^n = \bigoplus_k \mathcal{V}_k$. As a consequence, we have

$$\sqcup\mathcal{A} \preccurlyeq \sum_k A_k\,.$$

Proof. We use the notation of the previous proof. We introduce the set $\mathcal{S}_k = \mathcal{V}_k \cap \{x \in \mathbb{R}^n \mid x^T C x \ = 1\}$. We have shown in Section 3.3 that there is a probability measure ν supported on $\mathcal{U}(\mathcal{A}) = \cup_k \mathcal{S}_k$ such that

$$\sqcup\mathcal{A} = \int_{\mathcal{U}(\mathcal{A})} xx^T d\nu\,.$$

Since $\mathcal{S}_k \subset \mathcal{U}(\mathcal{A})$, the matrix $\int_{\mathcal{S}_k} xx^T d\nu$ is smaller than $\sqcup\mathcal{A}$ in the Löwner order. Moreover, its range in included in $\mathcal{V}_k$, thus it is less than $\operatorname{short}(\sqcup\mathcal{A}, \mathcal{V}_k)$. Since $\mathcal{U}(\mathcal{A}) = \cup_k \mathcal{S}_k$, we have

$$\sqcup\mathcal{A} \preccurlyeq \sum_k \operatorname{short}(\sqcup\mathcal{A}, \mathcal{V}_k)\,,$$

and equality holds if $R^n = \oplus_k \mathcal{V}_k$, since then $\mathcal{U}(\mathcal{A})$ is the disjoint union of the sets $\mathcal{S}_k$. Finally, the fact that $\operatorname{short}(\sqcup\mathcal{A}, \mathcal{V}_k) = \operatorname{short}(A_k, \mathcal{V}_k)$ is a consequence of Lemma 3.15. □

Remark 3.9. We point out that these inequalities may not hold for other minimal upper bound selections. Indeed, it we consider the trace-minimizing selection on the matrices $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, we obtain $A \sqcup B = \left(\begin{smallmatrix} 1.44 & 0.72 \\ 0.72 & 1.17 \end{smallmatrix}\right)$ for which the spectrum of $(A+B)-(A \sqcup B)$ is $\{0.64, -0.24\}$ and is not non-negative.

Remark 3.10. Also note that the arithmetic mean is dominated by the invariant join. Indeed, we have $A_k \preccurlyeq \sqcup\mathcal{A}$ for all k, so $p^{-1}\sum_k A_k \preccurlyeq \sqcup\mathcal{A}$. By the same reasoning, the invariant join dominates any mean of the matrices in $\mathcal{A}$, see [KA79, PT05, Bha07] for background on matrix means.

3.6.4 Several properties of the invariant selection

We list in the following two positive results on the binary invariant join $(X, Y) \mapsto X \sqcup Y$ and several negative ones, that justify the sometimes convoluted proofs that are needed in the sequel.

First, given a map $f\colon \mathbb{R} \to \mathbb{R}$ and a symmetric matrix X, the value $f(X)$ is defined by

$$f(X) = U \operatorname{diag}\big[f(d_i)\big]U^T,$$

where $X = U \operatorname{diag}(d_i)U^T$, see [Bha97] for more background.

Proposition 3.18. *Let f denote any function preserving the set of positive reals and $M \in \mathcal{M}_n$ any matrix such that choosing*

$$MM^T = X^{1/2} f\big[X^{-1/2} Y X^{-1/2}\big] X^{1/2}$$

Then the binary invariant join $X \sqcup Y$ can be computed by

$$X \sqcup Y = \frac{X+Y}{2} + \frac{1}{2} M\big|M^{-1}(X-Y)M^{-T}\big|M^T. \tag{3.4}$$

In particular,

1. *$f \equiv 0$ (resp. $f \equiv 1$) yields $C = X$ (resp. $C = Y$),*
2. *$f \equiv x \mapsto x^{1/2}$ yields $C =$ the Riemannian barycenter (Karcher mean) of X, Y,*
3. *$f \equiv x \mapsto \max(1, x)$ yields $C = X \sqcup Y$.*

Proof. A classical result in linear algebra states that given two positive definite matrices X, Y, there is an invertible matrix P and a positive diagonal matrix D such that $X = PDP^T$ and $Y = PP^T$. Moreover, there is a (not necessarily unique) orthogonal matrix U such that $X^{1/2} = PU$. We deduce that $X^{-1/2}YX^{-1/2} = U^T DU$. By Theorem 3.13, we have $D \sqcup I = D \vee I$ where the maximum is taken entry- wise on the diagonal. Applying the congruence by P yields $X \sqcup Y = P(D \vee I)P^T$.

Let $M \in \mathcal{M}_n$ such that $MM^T = X^{1/2} f\big(X^{-1/2} Y X^{-1/2}\big) X^{1/2} = Pf(D)P^T$. Hence there is an orthogonal matrix V such that $M = Pf(D)^{1/2}V$. One can check that $M\big|M^{-1}(X - Y)M^{-T}\big|M^T = P\big|D - I\big|P^T$. □

Remark 3.11. In particular, M can be chosen to be a triangular matrix by the Cholesky decomposition.

As a direct consequence of the expression above, we obtain the continuity of the binary invariant join. It does not extend continuously to the closed cone.

Proposition 3.19. *The binary invariant join is continuous on the open cone $(\mathcal{S}_n^{++})^2$ but it is not continuous on the closed cone $(\mathcal{S}_n^{+})^2$.*

Proof. The continuity on the interior of the cone follows from Equation (3.4) and the continuity of the sum, inverse, square root and modulus on the set $\mathcal{S}_n^{++}$. We now give a counter-example showing that continuity does not hold on the closed cone. Let $X(s) \coloneqq \begin{pmatrix} \cos^2(s) & \cos(s)\sin(s) \\ \cos(s)\sin(s) & \sin^2(s) \end{pmatrix}$ and $Y(s) \coloneqq \begin{pmatrix} \cos^2(s) & -\cos(s)\sin(s) \\ -\cos(s)\sin(s) & \sin^2(s) \end{pmatrix}$. These matrices have rank 1 and their images are distinct when $s \notin \frac{\pi}{2}\mathbb{Z}$, in particular when $|s| < 1$ and $s \neq 0$. By Theorem 3.16, the invariant join of $X(s)$ and $Y(s)$ is their sum: $X(s) \sqcup Y(s) = \begin{pmatrix} 2\cos^2(s) & \\ & 2\sin^2(s) \end{pmatrix}$. Moreover, $X(0) = Y(0)$, so when s tends to 0, we have $\lim_{s\to 0} X(s) \sqcup Y(s) = \left(\begin{smallmatrix} 2 & \\ & 0 \end{smallmatrix}\right)$ but $X(0) \sqcup Y(0) = \left(\begin{smallmatrix} 1 & \\ & 0 \end{smallmatrix}\right)$. □

Proposition 3.20 (Theorem 1 in [GNS+13])**.** *No binary selection of minimal upper bounds in $\mathcal{S}_n^+$ is order preserving or associative. This holds in particular for the binary invariant join.*

Finally, we point out two properties that the maximum operator satisfies on the reals but do not extend to the case of symmetric matrices.

Proposition 3.21. *The binary invariant is not convex: the inequality*

$$(tX_1 + (1-t)X_2) \sqcup Y \preccurlyeq t(X_1 \sqcup Y) + (1-t)(X_2 \sqcup Y)\,.$$

is not satisfied in general.

Proof. We choose $X_1 = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ $X_2 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and $Y = I_2$. With $t = 0.5$, the spectrum of the difference between the right-hand side and the left-hand side is $\{0.23, -0.036\}$. □

Proposition 3.22. *The invariant join does not commute with the addition of a constant: in general, we have*

$$(X + M) \sqcup (Y + M) \neq (X \sqcup Y) + M\,,$$

for $X, Y, M \in \mathcal{S}_n^+$. Moreover, these quantities are not comparable.

Proof. We choose $X = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ $Y = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and $M = I_2$. The spectrum of the difference between the right-hand side and the left-hand side is $\{-1, 0.06\}$. □

CHAPTER 4

Lipschitz bounds on the invariant join

4.1 Introduction

We recall that the minimal upper bound selection in the cone $\mathcal{S}_n^+$ corresponding to the Löwner ellipsoid is denoted by $\sqcup$ and called the *invariant join*. We study in this chapter Lipschitz properties of the invariant join with respect to three classical metrics on the cone of positive definite matrices, namely the Thompson metric, the Hilbert (semi-)metric and the (invariant) Riemann metric. Our first main result (Theorem 4.1) shows that this operator is non-expansive with respect to the Riemann metric. Our second and third main result (Theorems 4.2 and 4.3) show that the invariant join is "possibly expansive" and that its Lipschitz constant behaves asymptotically like $\log n$.

The Thompson, Hilbert and Riemann metrics are respectively denoted by d_T, d_H and d_R and defined by

$$\begin{aligned} d_T(X,Y) &= \|\log \mathrm{Sp}(X^{-1}Y)\|_\infty\,, \\ d_H(X,Y) &= \omega\big(\log \mathrm{Sp}(X^{-1}Y)\big)\,, \\ d_R(X,Y) &= \|\log \mathrm{Sp}(X^{-1}Y)\|_2\,, \end{aligned}$$

where $\|\cdot\|_\infty$, $\omega(\cdot)$ and $\|\cdot\|_2$ respectively denote the max-norm, the diameter (or oscillation) and the Euclidean 2-norm on the space $\mathbb{R}^n$:

$$\|x\|_\infty = \max_i |x_i|\,, \quad \omega(x) = \max_i x_i - \min_i x_i\,, \quad \|x\|_2 = \big(\sum_i x_i^2\big)^{1/2}\,.$$

The *Lipschitz constant of the map f with respect to the metric $m \in \{T, H, R\}$* is the

quantity denoted $\operatorname{Lip}_m f$ and defined by

$$\operatorname{Lip}_m f = \sup_{X,Y \succ 0} \frac{d_m\big(f(X), f(Y)\big)}{d_m(X,Y)} .$$

In the present setting, the invariant join $\sqcup$ maps a pair of $n \times n$ positive definite matrices (X_1, X_2) to a single $n \times n$ positive definite matrix, given by $X_1 \sqcup X_2$. Thus it may be seen as a map sending the block-diagonal $(2n) \times (2n)$ symmetric matrix $X_1 \oplus X_2 \coloneqq \operatorname{diag}(X_1, X_2)$ to the $n \times n$ symmetric matrix $X_1 \sqcup X_2$. We are interested in computing the Lipschitz constant of that map, and it is given by

$$\operatorname{Lip}_m \sqcup \coloneqq \sup_{X_1,X_2,Y_1,Y_2} \frac{d_m(X_1 \sqcup X_2, Y_1 \sqcup Y_2)}{d_m(X_1 \oplus X_2, Y_1 \oplus Y_2)} .$$

4.1.1 The main results

Our main result shows that the Lipschitz constant of the invariant join is finite for all these metrics, that it is constant equal to 1 for the Riemann metric, and that it depends in a logarithmic way in the dimension for the Thompson and Hilbert metrics:

Theorem 4.1. *The invariant join is non-expansive in the Riemann metric:*

$$\operatorname{Lip}_R \sqcup = 1 .$$

Theorem 4.2. *The Lipschitz constant of the invariant join in the Thompson metric on $\mathcal{S}_n^+$ satisfies $\operatorname{Lip}_T \sqcup > 1$ when $n \geqslant 2$ and*

$$\frac{1}{\pi} - \frac{0.92}{\log n} \leqslant \frac{\operatorname{Lip}_T \sqcup}{\log n} \leqslant \frac{4}{\pi} + \frac{2}{\log n} + o(1) .$$

Theorem 4.3. *The Lipschitz constant of the invariant join in the Hilbert metric satisfies:*

$$\operatorname{Lip}_H \sqcup = \operatorname{Lip}_T \sqcup .$$

4.1.2 Proof outline and conjecture

Before giving the full detail of the proof, we outline the main steps that are involved. First, in a common step for all theorems, we recall in Section 4.2.1 properties arising from the Finsler nature of the considered metrics to reduce the computation of the Lipschitz constant of a smooth map to the computation of a local Lipschitz constant, by replacing each metric by its local norm counterpart. The supremum of the local Lipschitz constants along a given geodesic hence gives an upper bound on the (global) Lipschitz constant.

The aforementioned results do not apply directly to the invariant join, because it is not smooth everywhere. After a reduction step to the diagonal case in Section 4.2.2 and the computation of an explicit formula for the differential of the invariant join in Section 4.2.3, we show in Section 4.2.4 that the set of non-differentiability points "behaves well", that is one of the following cases is true:

1. there are either finitely many points on a geodesic where the invariant join is not smooth,

2. there is a small perturbation of the geodesic's end-points such that the new geodesic joining them has finitely many non-differentiability points.

Then the proofs of the theorems diverge. First, we prove Theorem 4.1 in Section 4.3. The proof of Theorem 4.2 in Section 4.4 reduces to bounding a *Schur multiplier norm* with respect to the spectral norm. This is a difficult computation, which is why we only obtain lower and upper bounds. Using results from [Mat93], we get an upper bound in $O(\log n)$ in Section 4.4.1, and we give in Section 4.4.2 an example of a point where the local Lipschitz constant scales as $\log(n)$. The proof of Theorem 4.3 in Section 4.5 uses a particular relationship between the Thompson and Hilbert metrics to deduce equality of the Lipschitz constants.

The trivial bound $\mathrm{Lip}_m \sqcup \geqslant 1$ holds since there is an injection from $\mathcal{S}_n^+$ to $\mathcal{S}_{n+1}^+$ given by $M \mapsto M \oplus 1$, showing that the Lipschitz constant $\mathrm{Lip}_m \sqcup$ is non-decreasing with the dimension. In particular, the operator $\sqcup$ coincides with the maximum operator on $\mathbb{R}_+$, hence $\mathrm{Lip}_m \sqcup \geqslant 1$. Moreover, $\mathrm{Lip}_T \sqcup > 1$ when $n = 2$ as shown by the following example:

$$X_1 = \begin{pmatrix} 2.63 & 0.33 \\ 0.33 & 1.82 \end{pmatrix} \quad Y_1 = \begin{pmatrix} 3.70 & -0.82 \\ -0.82 & 1.68 \end{pmatrix} \quad X_2 = Y_2 = \begin{pmatrix} 0.40 & 0.77 \\ 0.77 & 1.51 \end{pmatrix}.$$

We have $d_T(X_1 \sqcup X_2, Y_1 \sqcup Y_2) = 0.63$ while $d_T(X_1 \oplus X_2, Y_1 \oplus X_2) = 0.62$.

Finally, following rigorous testing on a wide range of dimension and set sizes, we conjecture that the invariant join is non-expansive in general.

Conjecture 4.4. *The invariant join is non-expansive in the Riemann metric: given two collections of positive definite matrices $\mathcal{A} = (A_k)_{1 \leqslant k \leqslant p}$ and $\mathcal{B} = (B_k)_{1 \leqslant k \leqslant p}$, we have*

$$d_R(\sqcup\mathcal{A}, \sqcup\mathcal{B}) \leqslant d_R\Big(\bigoplus_{1 \leqslant k \leqslant p} A_k, \bigoplus_{1 \leqslant k \leqslant p} B_k\Big).$$

4.2 Common step in the proofs

It will be convenient to write expressions of the form AXA^T in a concise way in the sequel. To do this, we recall the *congruence operator*, denoted by Γ and defined by

$$\Gamma_A X \coloneqq AXA^T.$$

For $X \in \mathcal{S}_n^+$ and $A, B \in \mathrm{GL}_n$, we have

$$\Gamma_A \Gamma_B X = \Gamma_{AB} X.$$

In particular, we deduce that the inverse of the congruence Γ_A is the congruence $\Gamma_A^{-1} = \Gamma_{A^{-1}}$, and that, when A is positive semidefinite, the congruence Γ_A has a square root, given by $\Gamma_A^{1/2} = \Gamma_{A^{1/2}}$.

4.2.1 Infinitesimal approach

Before proving Theorems 4.1 to 4.3, we describe a common step in all proofs and introduce the local Lipschitz constant.

The Thompson, Hilbert and Riemann metrics can be defined alternatively. Indeed, for $m \in \{T, H, R\}$, we have

$$d_m(X,Y) = \inf_{\gamma} \int_0^1 \nu_m\Big(\Gamma_{\gamma(s)}^{-1/2}\,\dot\gamma(s)\Big)\,ds \tag{4.1}$$

where γ is a curve joining X to Y. In the latter equation, $\nu_T(\cdot)$ is the spectral norm, $\nu_H(\cdot)$ is the Hilbert semi-norm and $\nu_R(\cdot)$ is the Frobenius norm:

$$\nu_T(X) \coloneqq \|X\| = \inf\{\alpha > 0\colon\ -\alpha^2 I \preccurlyeq XX^T \preccurlyeq \alpha^2 I\}\,, \tag{4.2}$$

$$\nu_H(X) \coloneqq \|X\|_H = \inf\{\beta - \alpha\colon \alpha^2 I \preccurlyeq XX^T \preccurlyeq \beta^2 I\}\,, \tag{4.3}$$

$$\nu_R(X) \coloneqq \|X\|_F = \big[\operatorname{trace}(XX^T)\big]^{1/2} = \Big(\sum_{ij} X_{ij}^2\Big)^{1/2}. \tag{4.4}$$

Hence, the Thompson, Hilbert and Riemann metrics are Finsler metrics, meaning that, locally, they resemble norms. More precisely, given any positive definite matrix Z, the metric d_m is locally given by a deformed version of the norm ν_m depending only on Z:

$$d_m(X,Y) \sim \nu_m\big(\Gamma_Z^{-1/2}(X-Y)\big) \quad \text{when } X, Y \to Z\,.$$

In all cases, the infimum in Equation (4.1) is obtained with the curve $\gamma\colon [0,1] \mapsto \mathcal{S}_n^{++}$ defined by

$$\gamma(s) \coloneqq \Gamma_X^{1/2}\Big[\big[\Gamma_X^{-1/2} Y\big]^s\Big] = X^{1/2}\Big[X^{-1/2} Y X^{-1/2}\Big]^s X^{1/2}\,, \tag{4.5}$$

which is thus a geodesic line joining X and Y with respect to all three metrics. Moreover, it is the unique geodesic joining X to Y in the Thompson and Riemann metrics, see [Nus88].

The *local Lipschitz constant of f at X with respect to the metric $m \in \{T, H, R\}$* is the quantity denoted $\operatorname{Lip}_Z^m f$ and defined by

$$\operatorname{Lip}_m^X f = \limsup_{\varepsilon \to 0} \sup_{Z\colon\, d_m(X,Z) \leqslant \varepsilon} \frac{d_m\big(f(X), f(Z)\big)}{d_m(X,Z)}\,.$$

Thus, if f is denotes a map that is differentiable at X, the local Lipschitz constant is given by

$$\operatorname{Lip}_m^X f = \sup_{H \in \mathcal{S}_n} \frac{\nu_m\Big(\Gamma_{f(X)}^{-1/2}\ (df_X \cdot H)\Big)}{\nu_m\Big(\Gamma_X^{-1/2}\ H\Big)}\,. \tag{4.6}$$

Moreover, when the map f is differentiable everywhere, its Lipschitz constant is equal to the supremum of its local Lipschitz constants:

$$\operatorname{Lip}_m f = \sup_X \operatorname{Lip}_m^X f\,. \tag{4.7}$$

We show in Section 4.2.4 that Equation (4.7) also holds for the invariant join, despite it not being differentiable everywhere.

4.2.2 Reduction to the co-diagonal case

We have shown previously that the invariant join commutes with the action of the linear group, meaning that if P is an invertible matrix, then, for all positive definite X, Y, we have

$$(\Gamma_P X) \sqcup (\Gamma_P Y) = \Gamma_P (X \sqcup Y) . \tag{4.8}$$

It follows that, if the invariant join is differentiable at some point $X \oplus Y$, its differential at this point satisfies

$$d \sqcup_{(\Gamma_P X)\oplus(\Gamma_P Y)} \cdot \big[(\Gamma_P H) \oplus (\Gamma_P K)\big] = \Gamma_P \Big[d \sqcup_{X\oplus Y} \cdot (H \oplus K) \Big] . \tag{4.9}$$

Since the matrices X, Y are positive definite, there is[1] an invertible matrix P and a diagonal matrix D, with positive diagonal entries, such that $\Gamma_P X = D$ and $\Gamma_P Y = D^{-1}$. We deduce from Equations (4.8) and (4.9) that

$$\begin{aligned} \Gamma^{-1/2}_{X\sqcup Y} \left[d \sqcup_{X\oplus Y} \cdot (H \oplus K) \right] &= \Gamma^{-1/2}_{D\sqcup D^{-1}} \left[d \sqcup_{D\oplus D^{-1}} \cdot (H' \oplus K') \right] , \\ \Gamma^{-1/2}_{X\oplus Y} \left[H \oplus K \right] &= \Gamma^{-1/2}_{D\oplus D^{-1}} (H' \oplus K') , \end{aligned}$$

where we have used the shortcut notation $H' = \Gamma_P^{-1} H$. Thus we have

$$\mathrm{Lip}_m^{X\oplus Y} \sqcup = \mathrm{Lip}_m^{D\oplus D^{-1}} \sqcup ,$$

meaning that computing the local Lipschitz constant of the invariant join at a differentiability point $X \oplus Y$ is equivalent to the computation of the local Lipschitz constant at a point where both matrices X, Y are diagonal matrices, and inverses of one another.

4.2.3 Differential of the invariant join

In the sequel, we denote by D a diagonal matrix with positive diagonal entries. We use the shorthand D_i to mean the diagonal entry D_{ii}.

We recall that the invariant join of two positive definite matrices X, Y is given by

$$X \sqcup Y = \frac{X+Y}{2} + \frac{1}{2} X^{1/2} \big| X^{-1/2} Y X^{-1/2} - I \big| X^{1/2} ,$$

where $|X|$ denotes the matrix $(XX^T)^{1/2}$.

We also recall that the *Löwner matrix* associated with a differentiable map f and a symmetric matrix X with eigenvalues $x_i \in \mathbb{R}$ is the matrix L^f_X defined by

$$[L^f_X]_{ij} \coloneqq \begin{cases} (x_j - x_i)^{-1} \big(f(x_j) - f(x_i) \big) & \text{if } i \neq j \\ f'(x_i) & \text{otherwise} . \end{cases}$$

It follows from [Mat93, Bha97] that the differential of the invariant join is given by a *Hadamard product* (or Schur product). We recall that the Hadamard product of two matrices A, B is given by $\big(A \circ B\big)_{ij} = A_{ij} B_{ij}$.

[1] See Section 3.6.2 for more detail.

Lemma 4.5. *The invariant join is differentiable at the point $D \oplus D^{-1}$ if the matrix $D - I$ is invertible and its differential is given by*

$$d \sqcup_{D\oplus D^{-1}} \cdot (H \oplus K) = L^{\sqcup}_{D^2-I} \circ H + L^{\sqcup}_{D^{-2}-I} \circ K \,, \tag{4.10}$$

where the matrix $L^{\sqcup}_X$ is the Löwner matrix *associated to the maximum operation at the diagonal matrix X, which is given by*

$$(L^{\sqcup}_X)_{ij} = \frac{1}{2} + \frac{1}{2}\frac{|X_{jj}| - |X_{ii}|}{X_{jj} - X_{ii}} = \frac{1}{2} + \frac{1}{2}\frac{X_{ii} + X_{jj}}{|X_{ii}| + |X_{jj}|} \,. \tag{4.11}$$

Remark 4.1. We point out that the map $H \mapsto L \circ H$ with $L \in \{L^{\sqcup}_{D^2-I}, L^{\sqcup}_{D^{-2}-I}\}$ commutes with the congruences $\Gamma^{\alpha}_{D} \cdot$ and the congruence $\Gamma^{\alpha}_{D\sqcup D^{-1}} \cdot$ for $\alpha \in \{-1/2,\, 1/2\}$ since the matrices D, D^{-1} and $D \sqcup D^{-1}$ are diagonal.

Remark 4.2. More generally, the invariant join is differentiable at the matrix $X \oplus Y$ if the matrix $X - Y$ is invertible.

4.2.4 Local and global Lipschitz constants

We now show that the global Lipschitz constant of the invariant join, with respect to the Thompson or Riemann metric, is equal to the supremum over local Lipschitz constants at its differentiability points, with respect to the same metric. In this section, we restrict the analysis to the case $m \in \{T, R\}$.

First, note that the inequality $\mathrm{Lip}^X_m \sqcup \leqslant \mathrm{Lip}_m \sqcup$ holds for all matrix X, since the quantity on the left-hand side only measures the rate of change of the map $\sqcup$ locally around the point X.

Then, we show the reverse inequality. Given positive definite matrices X_1, X_2, Y_1, Y_2, let γ denote the geodesic joining $X \coloneqq X_1 \oplus X_2$ to $Y \coloneqq Y_1 \oplus Y_2$ given by Equation (4.5). One can show that any value taken by γ is also a block diagonal matrix:

$$\gamma(s) = \gamma_1(s) \oplus \gamma_2(s) \,,$$

where γ_i is the geodesic joining X_i to Y_i, with $i \in \{1, 2\}$. We introduce the curve η joining $X_1 \sqcup X_2$ to $Y_1 \sqcup Y_2$ defined by

$$\eta(s) \coloneqq \gamma_1(s) \sqcup \gamma_2(s) \,.$$

Note in particular that for all $0 \leqslant s \leqslant 1$, we have $\gamma_1(s) \preccurlyeq \eta(s)$ and $\gamma_2(s) \preccurlyeq \eta(s)$.

We shall distinguish two cases in our analysis.

4.2.4.a Case 1: the map η has finitely many non-differentiability points

Let also $(x_k)_{0\leqslant k\leqslant p}$ denote the non-differentiability points of η on $[0, 1]$. On each open interval $]x_k, x_{k+1}[$, we have

$$\dot\eta(s) = d \sqcup_{\gamma(s)} \cdot \big(\dot\gamma(s)\big) \,,$$

so that, using Equation (4.1) and the definition of the local Lipschitz constant, we have

$$\begin{aligned} d_m(X_1 \sqcup X_2, Y_1 \sqcup Y_2) &\leqslant \sum_{k=0}^{p-1} \int_{x_k}^{x_{k+1}} \nu_m\big(\Gamma^{-1/2}_{\eta(s)}\, \dot\eta(s)\big)\, ds \\ &= \sum_{k=0}^{p-1} \int_{x_k}^{x_{k+1}} \nu_m\Big[\Gamma^{-1/2}_{\eta(s)}\, \big[d \sqcup_{\gamma(s)} \cdot \dot\gamma(s)\big]\Big]\, ds \,. \end{aligned}$$

By definition of the local Lipschitz constant $\mathrm{Lip}_m^{\gamma(s)}$, we have:

$$\nu_m\Big[\Gamma_{\eta(s)}^{-1/2}\left[d \sqcup_{\gamma(s)} \dot{\gamma}(s)\right]\Big] \leqslant \mathrm{Lip}_m^{\gamma(s)}\, \nu_m\Big[\Gamma_{\gamma(s)}^{-1/2}\left[\dot{\gamma}(s)\right]\Big].$$

Hence,

$$\begin{aligned} d_m(X_1 \sqcup X_2, Y_1 \sqcup Y_2) &\leqslant \sum_{k=0}^{p-1} \int_{x_k}^{x_{k+1}} \mathrm{Lip}_m^{\gamma(s)}\, \nu_m\Big[\Gamma_{\gamma(s)}^{-1/2}\left[\dot{\gamma}(s)\right]\Big]\, ds \\ &\leqslant \Big[\sup_{s\in[0,1]} \mathrm{Lip}_m^{\gamma(s)} \sqcup\Big] d_m(X_1 \oplus X_2, Y_1 \oplus Y_2)\,. \end{aligned}$$

Thus, we obtain

$$\frac{d_m(X_1 \sqcup X_2, Y_1 \sqcup Y_2)}{d_m(X_1 \oplus X_2, Y_1 \oplus Y_2)} \leqslant \sup_{s\in[0,1]} \mathrm{Lip}_m^{\gamma(s)} \sqcup \leqslant \sup_Z \mathrm{Lip}_m^Z \sqcup\,,$$

where the latter supremum is taken over all matrices Z at which the invariant join is differentiable.

4.2.4.b Case 2: the map η has infinitely many non-differentiability points First, we give a technical lemma which will greatly simplify the proof of the second case. We prove this lemma in Section 4.2.4.c.

Lemma 4.6. *Given $\varepsilon > 0$ and two positive definite matrices X, Y, there is a matrix Y^ε such that the invariant join only has finitely many non-differentiability points on the geodesic γ_X joining X to Y^ε and on the geodesic γ_Y joining Y^ε to Y. Moreover, the matrix Y^ε satisfies $d_m(Y, Y^\varepsilon) \leqslant \varepsilon d_m(X, Y)$.*

We apply this lemma to the two geodesics γ_1 and γ_2, which yields the matrix $Y^\varepsilon \coloneqq Y_1^\varepsilon \oplus Y_2^\varepsilon$. Next, we apply the results of the previous case twice, which yields the following sequence of inequalities:

$$\begin{aligned} d_m(X_1 \sqcup X_2, Y_1 \sqcup Y_2) &\leqslant d_m(X_1 \sqcup X_2, Y_1^\varepsilon \sqcup Y_2^\varepsilon) + d_m(Y_1^\varepsilon \sqcup Y_2^\varepsilon, Y_1 \sqcup Y_2) \\ &\leqslant \Big[\sup_Z \mathrm{Lip}_m^Z \sqcup\Big]\Big(d_m(X, Y^\varepsilon) + d_m(Y^\varepsilon, Y)\Big) \\ &\leqslant (1+2\varepsilon)\Big[\sup_Z \mathrm{Lip}_m^Z \sqcup\Big] d_m(X, Y)\,. \end{aligned}$$

Since the value of ε can be taken arbitrarily small, we get

$$\frac{d_m(X_1 \sqcup X_2, Y_1 \sqcup Y_2)}{d_m(X_1 \oplus X_2, Y_1 \oplus Y_2)} \leqslant \sup_Z \mathrm{Lip}_m^Z \sqcup\,,$$

where the latter supremum is taken over all matrices Z at which the invariant join is differentiable. This concludes the proof. □

4.2.4.c Proof of Lemma 4.6 Using the notation introduced earlier and Lemma 4.5, the map η has infinitely many non-differentiability points if the equation

$$\det\Big(\gamma_1(s) - \gamma_2(s)\Big) = 0 \quad \text{infinitely often for } s \in [0,1]\,. \tag{4.12}$$

By Equation (4.5), there are invertible matrices A, B and diagonal matrices D, Δ such that $\gamma_1(s) = A\exp(sD)A^T$ and $\gamma_2(s) = B\exp(s\Delta)B^T$. Thus, Equation (4.12) is equivalent to

$$\det\Big(\exp(sD) - M\exp(s\Delta)M^T\Big) = 0 \quad \text{infinitely often for } s \in [0,1]\,, \tag{4.13}$$

with $M = A^{-1}B$. The map $F \coloneqq s \mapsto \det\Big(\exp(sD) - M\exp(s\Delta)M^T\Big)$ is analytic, thus if it has infinitely many zeros on the compact set $[0,1]$, it must be identically zero. By the Leibniz formula, we have

$$F(s) = \sum_{I\subseteq[1,n]} (-1)^{|I|}\det\big(M\exp(s\Delta)M^T\big)_{I^c}\exp\big(s\sum_{i\in I} D_i\big)\,.$$

In the latter equation, we denote by $\det(X)_I$ the determinant of the square sub-matrix of X whose rows and columns indices are in the set I. By definition, we have $\det(X)_\emptyset = 1$.

Let us consider the term corresponding to $I = \{1, \cdots, n\}$, which is equal to

$$(-1)^n \exp(s\,\mathrm{trace}\,D) \neq 0\,.$$

Thus, in order for the map F to be zero, there must be at least another term in the sum which contains a term proportional to $\exp(s\,\mathrm{trace}\,D)$. The value of $\det\big(M\exp(s\Delta)M^T\big)_{I^c}$ is a multinomial in $\exp(s\Delta_i)$, thus the existence of another term proportional to $\exp(s\,\mathrm{trace}\,D)$ implies that some linear combination

$$\sum_{i\in I}\alpha_i\Delta_i + \sum_{j\in J} D_j = \sum_{1\leqslant i\leqslant n} D_i$$

is satisfied for some sets I, J and $\alpha_k \in \{1, \ldots, n\}$. There is only a finite number of such relations, and they are generically not verified. Thus, the map η generically has only a finite number number of non-differentiability points. Hence, there must an open neighborhood of the matrix Y such that, for any matrix Y^ε in this neighborhood, the geodesics joining X to Y^ε and Y^ε to Y contain a finite number of non-differentiability points of the map η. □

4.3 Nonexpansivity in the Riemann metric

We combine Equations (4.4), (4.6) and (4.10) with Remark 4.1 to obtain the expression of the local Lipschitz constant of the invariant join with respect to the Riemann metric at the matrix $D \oplus D^{-1}$:

$$\mathrm{Lip}_R^{D\oplus D^{-1}} \sqcup = \sup_{H,K\in\mathcal{S}_n} \frac{\|L^{\sqcup}_{D^2-I}\circ(\Gamma^{-1/2}_{D\sqcup D^{-1}}H) + L^{\sqcup}_{D^{-2}-I}\circ(\Gamma^{-1/2}_{D\sqcup D^{-1}}K)\|_F}{\|(\Gamma_D^{-1/2}H)\oplus(\Gamma_D^{1/2}K)\|_F}\,. \tag{4.14}$$

Without loss of generality, we may assume that the diagonal entries of the matrix D are ordered in decreasing order. Let $(\lambda_i)_{1\leqslant i\leqslant p}$ denote the entries of D that are larger than 1 and $(\mu_i)_{1\leqslant i\leqslant q}$ the entries that are smaller than 1. By Equation (4.11), we get the expressions:

$$L^{\sqcup}_{D^2-1} = \begin{pmatrix} J_p & Z \\ Z^T & 0_q \end{pmatrix} \quad \text{with} \quad Z_{ij} = \frac{\lambda_i^2 - 1}{\lambda_i^2 - \mu_j^2}\,, \tag{4.15}$$

$$L^{\sqcup}_{D^{-2}-I} = \begin{pmatrix} 0_p & W \\ W^T & J_q \end{pmatrix} \quad \text{with} \quad W_{ij} = \frac{\lambda_i^{-2} - 1}{\lambda_i^{-2} - \mu_j^{-2}}\,. \tag{4.16}$$

Note in particular that $0 \leqslant Z_{ij}, W_{ij} \leqslant 1$.

For brevity, we denote the numerator in Equation (4.14) as $N(H,K)$ and to the denominator as $D(H,K)$. Our goal is to show that $N(H,K)^2 \leqslant D(H,K)^2$. Recall that the Frobenius norm of the matrix X is given by

$$\|X\|_F^2 = \sum_{i,j} X_{ij}^2\,.$$

We prove that for each (i,j), the squared term appearing in the numerator with the variables H_{ij} and K_{ij} (also called ij-term) is no larger than the squared term appearing in the denominator with the same variables. To do this, we split the set of indexes of the matrix into 4 sets, each corresponding to a block in the matrices written in Equations (4.15) and (4.16).

4.3.1 The case of diagonal blocks

First, assume that $1 \leqslant i,j \leqslant p$. Due to the J and zero diagonal blocks present in the matrices in Equations (4.15) and (4.16), the ij-term in the numerator is simply written $(\lambda_i\lambda_j)^{-1}H_{ij}^2$ and the denominator is written $(\lambda_i\lambda_j)^{-1}H_{ij}^2 + \lambda_i\lambda_j K_{ij}^2$. Since all the data is non-negative, we obtain the desired inequality. The case $i,j \geqslant p+1$ is obtained by symmetry.

4.3.2 The case of off-diagonal blocks

By symmetry, we assume that $i > j$. We need to show that the following inequality holds:

$$\Big[\frac{Z_{ij}}{\sqrt{\lambda_i\mu_j^{-1}}}H_{ij} + \frac{W_{ij}}{\sqrt{\lambda_i\mu_j^{-1}}}K_{ij}\Big]^2 \leqslant \Big[\frac{1}{\sqrt{\lambda_i\mu_j}}H_{ij}\Big]^2 + \Big[\frac{1}{\sqrt{\lambda_i^{-1}\mu_j^{-1}}}K_{ij}\Big]^2. \tag{4.17}$$

We write $\lambda_i = \sqrt{1+a}$ and $\mu_j = \sqrt{1-b}$ with $0 < a$ and $0 < b < 1$, so after simplification, Equation (4.17) is equivalent to

$$\Big[\frac{a}{a+b}H_{ij} + \frac{b}{a+b}(1+a)K_{ij}\Big]^2 \leqslant H_{ij}^2 + \frac{1+a}{1-b}K_{ij}^2\,,$$

By the Jensen inequality, we already have

$$\Big[\frac{a}{a+b}H_{ij} + \frac{b}{a+b}(1+a)K_{ij}\Big]^2 \leqslant \frac{a}{a+b}H_{ij}^2 + \frac{b}{a+b}(1+a)^2K_{ij}^2\,,$$

thus it remains to be proven that $\frac{b}{a+b}(1+a) \leqslant \frac{1}{1-b}$. In other words, one must show that $a(1+b^2-b)+b^2 \geqslant 0$. This inequality is satisfied, since $a(1+b^2-b)+b^2 \geqslant a(1+b^2-2b)+b^2 = a(1-b)^2+b^2 \geqslant 0$. This concludes the proof. □

4.4 Lipschitz constant bounds in the Thompson metric

4.4.1 Upper bound in the Thompson metric

First, we prove the upper bound on the Lipschitz constant in Theorem 4.2. We combine Equation (4.6) with Remark 4.1 to obtain

$$\mathrm{Lip}_T^{D\oplus D^{-1}} \sqcup = \frac{1}{2} \sup_{H',K'\in\mathcal{S}_n} \frac{\left\|\Gamma_A\left[L_1\circ H'\right]+\Gamma_B\left[L_2\circ K'\right]\right\|}{\left\|H'\oplus K'\right\|}$$

with

$$\begin{aligned} A &\coloneqq \left(D\sqcup D^{-1}\right)^{-1/2} D^{1/2}\,, & B &\coloneqq \left(D\sqcup D^{-1}\right)^{-1/2} D^{-1/2}\,,\\ L_1 &\coloneqq L^{\sqcup}_{D^2-I}\,, & L_2 &\coloneqq L^{\sqcup}_{D^{-2}-I}\,,\\ H' &\coloneqq \Gamma_D^{-1/2} H\,, & K' &\coloneqq \Gamma_D^{1/2} K\,. \end{aligned}$$

By definition of the invariant join, we have $D\sqcup D^{-1} \succcurlyeq D, D^{-1}$. Thus the matrices A, B satisfy $AA^T \leqslant I$ and $BB^T \preccurlyeq I$, and the spectral norms of A, B are no larger than 1: $\|A\|, \|B\| \leqslant 1$.

By the triangle inequality and the fact that the spectral norm $\|\cdot\|$ is sub-multiplicative, we deduce that

$$\left\|\Gamma_A\left[L_1\circ H'\right]+\Gamma_B\left[L_2\circ K'\right]\right\| \leqslant \|A\|^2\left\|L_1\circ H'\right\| + \|B\|^2\left\|L_2\circ K'\right\|.$$

Moreover, we have $L^{\sqcup}_X = \frac{1}{2}J + \frac{1}{2}L^{|\cdot|}_X$ where J denotes the matrix with a 1 in each entry. We also have $\left\|H'\oplus K'\right\| = \max\left[\left\|H'\right\|, \left\|K'\right\|\right]$. Thus, we have

$$\mathrm{Lip}_T^{D\oplus D^{-1}} \sqcup \leqslant 1 + \frac{1}{2}\|L^{|\cdot|}_{D^2-I}\|^\circ + \frac{1}{2}\|L^{|\cdot|}_{D^{-2}-I}\|^\circ\,,$$

where we have introduced the *Schur multiplier norm* $\|\cdot\|^\circ$ of the matrices $L^{|\cdot|}_{D^2-I}$ and $L^{|\cdot|}_{D^{-2}-I}$ defined by

$$\|L\|^\circ \coloneqq \sup_{H\in\mathcal{S}_n} \frac{\|L\circ H\|}{\|H\|}\,.$$

Mathias has shown in [Mat93] that the following inequality holds for all diagonal matrix X:

$$\sup_{H\in\mathcal{M}_n} \frac{\|L^{|\cdot|}_X\circ H\|}{\|H\|} \leqslant 2\gamma_n + 1 \quad \text{with} \quad \gamma_n = \frac{1}{n}\sum_{j=1}^{n} |\cot\frac{(2j-1)\pi}{2n}|\,.$$

Moreover, we have $\gamma_n = \frac{2}{\pi}\log n + o(\log n)$. Since $\mathcal{S}_n \subset \mathcal{M}_n$, we immediately obtain

$$\mathrm{Lip}_T \sqcup = \sup_D \mathrm{Lip}_T^{D\oplus D^{-1}} \sqcup \leqslant 2 + \frac{4}{\pi}\log n + o(\log n)\,. \qquad \square$$

4.4.2 Lower bounds in the Thompson metric

We introduce the parametric collection of diagonal $(2n) \times (2n)$ matrices $\{D_t\}_{t>0}$ defined by

$$D_t = \operatorname{diag}\big(\sqrt{1+t^{2i+1}}\big)_{1\leqslant i\leqslant n} \oplus \operatorname{diag}\big(\sqrt{1-t^{2j}}\big)_{1\leqslant j\leqslant n},$$

as well as the $(2n) \times (2n)$ matrix H given by

$$H = \begin{pmatrix} 0 & W \\ W^T & 0 \end{pmatrix} \quad \text{with} \quad W_{ij} = \begin{cases} 0 & \text{if } i = j \\ (i-j)^{-1} & \text{otherwise} \end{cases}.$$

The Löwner matrix L_t of the invariant join at the matrix $D_t^2 - I$ is

$$L_t = \begin{pmatrix} J & M_t \\ M_t^T & 0 \end{pmatrix} \quad \text{with} \quad \big(M_t\big)_{ij} = \frac{1}{2}\frac{t^{2i+1}-t^{2j}}{t^{2i+1}+t^{2j}}.$$

Finally, let L_0 denote the limit of the Löwner matrices L_t which is given by

$$L_0 \coloneqq \begin{pmatrix} J & M_0 \\ M_0^T & 0 \end{pmatrix}.$$

We give a lower bound on local Lipschitz constant of the invariant join at the matrix D_t by bounding the value of the Schur multiplier norm of L_t with respect to the local (spectral) norm induced by $D_t \sqcup D_t^{-1}$. To this end, by means of the matrix H, we show that the value $\|(D_t \sqcup D_t^{-1})^{-1}(L_t \circ H)\|$ is larger than $(\log(n)/\pi - 0.92 + O(t))\|H\|$.

As the parameter t tends to 0, the matrix M_t tends to the matrix M_0 given by $M_0 \coloneqq T - J/2$, where T is the *triangular truncation operator* defined by $T_{ij} = 1$ if $i > j$ and 0 otherwise. It has already been shown in [ACN92] that $\|T \circ W\| \geqslant \frac{(\log n - 1)}{\pi}\|W\|$.

Moreover, there are positive constants α_n, β_n depending only on the dimension n such that

$$|L_0 - L_t|_{ij} \leqslant \alpha_n t, \qquad |(D_t \sqcup D_t^{-1})^{-1/2} - I|_{ij} \leqslant \beta_n t,$$

for all indexes i, j. Hence, since all norms on $\mathbb{R}^n$ are equivalent, there is a constant γ_n, depending only on n such that $\|(D_t \sqcup D_t^{-1})^{-1}\big[L_t \circ H\big] - L_0 \circ H\| \leqslant \gamma_n t\|H\|$. Moreover, we have $\|H\| = \|W\|$ and $\|L_0 \circ H\| = \|T \circ W - W/2\|$. We deduce that

$$\begin{aligned} \|(D_t \sqcup D_t^{-1})^{-1}\big[L_t \circ H\big]\| &\geqslant \|T \circ W\| - \frac{1}{2}\|W\| - \gamma_n t \\ &\geqslant \Big[\frac{\log n}{\pi} - \frac{1}{\pi} - \frac{1}{2} + O(t)\Big]\|H\| \\ &= \Big[\frac{\log(2n)}{\pi} - \frac{1+\log 2}{\pi} - \frac{1}{2} + O(t)\Big]\|H\|. \end{aligned}$$

Taking the limit when $t \to 0$, we obtain $\operatorname{Lip}_T \sqcup \geqslant \log(n)/\pi - 0.92$. □

Remark 4.3. It is worth noting that the growth in $\log n$ is not surprising, since the example taken here is reminiscent of the study of *matrix pinching*, see [Bha00].

4.5 Lipschitz constant in the Hilbert metric

Before proving Theorem 4.3, we recall a link between the Thompson and Hilbert metrics: for all positive definite matrices X, Y, we have

$$d_H(X,Y) = 2 \inf_{\lambda>0} d_T(\lambda X, Y)\,. \tag{4.18}$$

Moreover, this infimum is always reached for some positive value of λ.

We can now prove the equality of Lipschitz constants. Given positive definite matrices X_1, X_2, Y_1, Y_2, there is a positive λ such that

$$d_H(X_1 \oplus X_2, Y_1 \oplus Y_2) = 2d_T\Big((\lambda X_1) \oplus (\lambda X_2), Y_1 \oplus Y_2\Big)\,.$$

The invariant join is positively homogeneous, thus $(\lambda X_1) \sqcup (\lambda X_2) = \lambda(X_1 \sqcup X_2)$ and

$$d_H(X_1 \sqcup X_2, Y_1 \sqcup Y_2) \leqslant 2d_T\Big((\lambda X_1) \sqcup (\lambda X_2), Y_1 \sqcup Y_2\Big)\,.$$

Hence, the following inequality holds

$$\frac{d_H(X_1 \sqcup X_2, Y_1 \sqcup Y_2)}{d_H(X_1 \oplus X_2, Y_1 \oplus Y_2)} \leqslant \frac{d_T\big(\lambda(X_1 \sqcup X_2), Y_1 \sqcup Y_2\big)}{d_T\big(\lambda(X_1 \oplus X_2), Y_1 \oplus Y_2\big)}\,,$$

from which we deduce $\mathrm{Lip}_H \sqcup \leqslant \mathrm{Lip}_T \sqcup$.

Similarly, there is a positive μ such that

$$d_H(X_1 \sqcup X_2, Y_1 \sqcup Y_2) = 2d_T\Big((\mu X_1) \sqcup (\mu X_2), Y_1 \sqcup Y_2\Big)\,.$$

We also have

$$d_H(X_1 \oplus X_2, Y_1 \oplus Y_2) \leqslant 2d_T\Big((\mu X_1) \oplus (\mu X_2), Y_1 \oplus Y_2\Big)\,,$$

from which we deduce the opposite inequality

$$\frac{d_H\big(X_1 \sqcup X_2, Y_1 \sqcup Y_2\big)}{d_H\big(X_1 \oplus X_2, Y_1 \oplus Y_2\big)} \geqslant \frac{d_T\big(\mu(X_1 \sqcup X_2), Y_1 \sqcup Y_2\big)}{d_T\big(\mu(X_1 \oplus X_2), Y_1 \oplus Y_2\big)}\,,$$

and $\mathrm{Lip}_H \sqcup \geqslant \mathrm{Lip}_T \sqcup$. □

Part II

Ellipsoidal invariants for switched systems

CHAPTER 5

Introduction to switched systems, ellipsoids and abstract interpretation

We recall in this chapter several notions of interest in the sequel: switched systems, operations on ellipsoids and the abstract interpretation framework on which we rely.

5.1 Classes of switched systems

5.1.1 Affine switched systems

The following notation is common to all types of switched systems that we consider. We denote by x (resp. y) the vector containing the *state variables* $x_1, \ldots, x_n$ (resp. $y_1, \ldots, y_n$). We assume that we are given a bounded set of initial states $\mathcal{I}$. We also denote by u the vector of *input variables* $u_1, \ldots, u_m$, that may represent values measured from a sensor. These values are measured from a bounded set $\mathcal{U}$. The operations $+$ and $\times$ are vector or matrix operations. A vector assignment is denoted $(x_1, \ldots, x_n) := (y_1, \ldots, y_n)$, or $x := y$ for short. The non-deterministic choice of a value for the vector x inside a set $\mathcal{X}$ is denoted $x \leftarrow \mathcal{X}$. Each occurrence of the symbol $\diamond$ can be replaced by either $<$ or $\leqslant$, so that an expression of the form $\langle f, x\rangle \diamond g$ defines a affine guard condition, separating the state space into two half-spaces when the vector f is nonzero. The system switches between p modes that are labeled by integers in $\Sigma := \{1, \ldots, p\}$. In the following, for each $i \in \Sigma$, A_i denotes a $n \times n$ matrix, B_i denotes a $n \times m$ matrix and c_i denotes a vector of dimension n. Finally, *rand_bool* refers a non-deterministic choice in $\{true, false\}$.

In its most general form, a switched affine system takes the form of Program 3, involving several affine switching conditions within a loop. For simplicity, we assume that all variables have global scope.

Program 3: Switched affine program with guards

$y \leftarrow \mathcal{I}$;
while *rand_bool* **do**
  $x \coloneqq y$;
  $u \leftarrow \mathcal{U}$;
  if $\big(\langle f_{1,1}, x\rangle \diamond g_{1,1}\big) \wedge \cdots \wedge \big(\langle f_{1,P}, x\rangle \diamond g_{1,P}\big)$ **then**
    $y \coloneqq A_1 \times x + B_1 \times u + c_1$;
  end
  $\vdots$
  if $\big(\langle f_{p,1}, x\rangle \diamond g_{p,1}\big) \wedge \cdots \wedge \big(\langle f_{p,P}, x\rangle \diamond g_{p,P}\big)$ **then**
    $y \coloneqq A_p \times x + B_p \times u + c_p$;
  end
end

In most applications we have in mind, the switching conditions are mutually exclusive, meaning exactly one switching condition is valid for any value of the variable vector x. We point out that this assumption can be made without loss of generality, up to adding more conditional statements within the loop and adjusting the assignments accordingly.

We shall also consider the somehow simpler variant of this program in which every guard condition $\langle f_{i,j}, x\rangle \diamond g_{i,j}$ is replaced by the test of a random boolean, and refer to it as a *non-deterministic switched system*. Note that in this case, the switching conditions are mutually exclusive. An example is shown in Program 4.

Program 4: Non-deterministic switched affine program

$x \leftarrow \mathcal{I}$;
while *rand_bool* **do**
  $u \leftarrow \mathcal{U}$;
  $b \leftarrow \Sigma$;
  if $b = 1$ **then**
    $x \coloneqq A_1 \times x + B_1 \times u + c_1$;
  end
  $\vdots$
  if $b = p$ **then**
    $x \coloneqq A_p \times x + B_p \times u + c_p$;
  end
end

An invariant for Program 3 is defined as a set $\mathcal{X}$ that satisfies $\mathcal{I} \subseteq \mathcal{X}$, and, for all i, j, $x \in \mathcal{X}$ and $u \in \mathcal{U}$,

$$A_i x + B_i u + c_i \in \mathcal{X} \quad \text{whenever } \langle f_{i,j}, x\rangle \diamond g_{i,j}\,.$$

In the non-deterministic case, the condition $\langle f_{i,j}, x\rangle \diamond g_{i,j}$ is dropped.

5.1.2 Linear switched systems

Joint spectral radius We specialize Program 4 by dropping the affine part and influence of an external control, by setting $B_i = 0$ and $c_i = 0$ for all i. This yields an instance of a linear switched system, depicted in Program 5. We do not consider linear systems with state-dependent switching laws, thus we drop the term "non-deterministic" in this setting.

Program 5: Switched linear program

```
x ← I;
while rand_bool do
    b ← Σ;
    if b = 1 then
        x := A_1 × x;
    end
    ⋮
    if b = p then
        x := A_p × x;
    end
end
```

Alternatively, a discrete-time switched linear system is described by:

$$x(k+1) = A_{\sigma(k)}x(k),\ \sigma(k) \in \Sigma$$

where $x(k) \in \mathbb{R}^n$ denotes the trajectory of the system, and σ is the switching mechanism, which selects one of the matrices in $\mathcal{A} = \{A_1, \ldots, A_p\}$ at each instant. It is known that the system described in Program 5 is stable if and only if the *joint spectral radius* of the set of matrices $\mathcal{A}$ is no more than 1, see [Jun09]. The latter is defined by

$$\rho(\mathcal{A}) \coloneqq \lim_{k\to+\infty} \max_{1\leqslant i_1,\ldots,i_k\leqslant p} \|A_{i_1} \ldots A_{i_k}\|^{1/k} \ .$$

Note is particular that since the spectral norm $\|\cdot\|$ of a matrix and of its transpose are the same, the joint spectral radius of $\mathcal{A}$ and $\mathcal{A}^T \coloneqq \{A_1^T, \ldots, A_p^T\}$ coincide. Moreover, the definition of the join spectral radius is independent of the norm, since all norms on $\mathbb{R}^n$ are equivalent and $\lim_{k\to+\infty} \alpha^{1/k} = 1$ when $\alpha > 0$.

A fundamental result of Barabanov [Bar88] shows that if $\mathcal{A}$ is irreducible, meaning that there is no nontrivial subspace of $\mathbb{R}^n$ that is left invariant by every matrix in $\mathcal{A}$, then there is a norm v on $\mathbb{R}^n$ such that

$$\lambda v(x) = \max_{1\leqslant i\leqslant p} v(A_i x), \qquad \forall x \in \mathbb{R}^n \ ,$$

for some $\lambda > 0$. The scalar λ is unique and it coincides with the joint spectral radius $\rho(\mathcal{A})$. This shows that, when $\mathcal{A}$ is irreducible, all the trajectories of the switched linear system converge to zero if and only if $\rho < 1$. The norm v is known as a *Barabanov norm*. A norm which satisfies the inequality $\max_i v(A_i x) \leqslant \rho(\mathcal{A})v(x)$ for all $x \in \mathbb{R}^n$ is called an *extremal norm*.

A closely related result by Dranishnikov, Konyagin and Protasov [Pro96] shows that when the set of matrix $\mathcal{A}$ is irreducible, there is a symmetric convex set with non-empty interior (i.e. a symmetric convex body) M and a positive λ such that

$$\operatorname{conv} \cup_i \{A_i \cdot M\} = \lambda M$$

The scalar λ is unique a coincides again with the joint spectral radius of $\mathcal{A}$. Any such body is called an invariant body and is obtained as the polar set of ball B of a Barabanov norm of the set of adjoint matrices $\mathcal{A}^T$:

$$x \in M \iff \forall y \in B,\, |\langle x, y\rangle| \leqslant 1\,.$$

When $\rho(\mathcal{A}) \leqslant 1$, the sub-level set $\{x \in \mathbb{R}^n \colon v(x) \leqslant \alpha\}$ for $\alpha > 0$ is mapped into itself by each dynamic $x \mapsto A_i x$, so that choosing $\alpha \coloneqq \sup_{y \in \mathcal{I}} v(y)$ yields an invariant for the switched linear system in Program 5. The same property holds for the Protasov ball M. We point out that the invariant that is obtained in the linear case in convex. This does not hold for affine systems.

Path-complete graph Lyapunov functions In [AJPR14], Ahmadi and al. developed a method to compute an overapproximation of the joint spectral radius of a finite set of matrices, to which we shall compare our method.

Given a set of states $\mathcal{W}$ and an alphabet Σ, an *edge* of a *labeled graph* is a triple $(i, \sigma, j) \in \mathcal{W} \times \Sigma \times \mathcal{W}$. The set of edges is denoted E. Such a graph is called *path-complete* if for every state i and letter σ, there is some state j such that (i, σ, j) is an edge.

Let $\mathcal{A} = \{A_\sigma\}_{\sigma \in \Sigma}$ denote a finite set of $n \times n$ matrices and ρ a non-negative real number. In [AJPR14], the authors examine graphs, denoted $\mathcal{G}(X, \rho)$, whose states are positive definite matrices $\{X_i\}_i$ and whose edges are determined by

$$(i, \sigma, j) \in E \iff A_\sigma^T X_i A_\sigma \preccurlyeq \rho^2 X_j\,.$$

The main theorem in [AJPR14] shows that the construction of a path-complete graph $\mathcal{G}(X, \rho)$ gives an upper bound of the joint spectral radius:

Theorem 5.1 (Theorem 2.4 [AJPR14])**.** *If the graph $\mathcal{G}(X, \rho)$ is path-complete for some set of positive definite matrices $\{X_i\}_i$, then $\rho(\mathcal{A}) \leqslant \rho$. Moreover, the map $V \colon z \mapsto \max_i z^T X_i z$ is a Lyapunov-type function: it satisfies $V(A_\sigma x) \leqslant \rho^2 V(x)$ for all $\sigma \in \Sigma$ and $x \in \mathbb{R}^n$.*

Moreover, for every $\varepsilon > 0$, there is an automaton $(\Sigma, \mathcal{W}, \tau)$ such that an LMI $(\mathcal{P}_\rho)$ built with this automaton has a solution with $\rho \leqslant \rho(\mathcal{A}) + \varepsilon$.

In practice, for a fixed value of ρ and a given path-complete graph $\mathcal{G}$, checking the existence of a path-complete graph $\mathcal{G}(X, \rho)$ whose edges coincide with $\mathcal{G}$ amounts to checking the feasibility of the LMI $(\mathcal{P}_\rho)$:

$$\begin{aligned} \rho^2 X_j &\succcurlyeq A_\sigma^T X_i A_\sigma\,, \quad \forall (i, \sigma, j) \colon \tau(i, \sigma) = j\,, \\ X_i &\succ 0\,, \quad \forall i\,. \end{aligned} \tag{$\mathcal{P}_\rho$}$$

A bisection scheme is then implemented to refine ρ. For brevity, we shall refer to this method as the LMI method.

A class of graphs which provides good theoretical and experimental approximations is the class of *De Bruijn* graphs. The set of states of the De Bruijn graph of order d is the set Σ^d of words built on Σ which have length d. There is an edge (i, σ, j) between states i and j if and only if $i = \sigma_1 \dots \sigma_d$ and $j = \sigma_2 \dots \sigma_d \sigma$. This graph, denoted by D_d, is path-complete by construction.

5.1.3 Optimal switching problem

We also consider the following problem of optimal switching between linear quadratic models, studied by McEneaney [McE07], namely approximating the value function V of an optimal control problem having both a control u taking values in $\mathbb{R}^m$ and a discrete control (switches between different modes) μ taking values in $\Sigma \coloneqq \{1, \dots, p\}$:

$$V(x) = \sup_{u \in \mathcal{U}} \sup_{\mu \in \mathcal{D}} \sup_{t>0} \int_0^t \frac{1}{2} \xi(s)^T D^{\mu(s)} \xi(s) - \frac{\gamma^2}{2} |u(s)|^2 \, ds \,.$$

Here, $\mathcal{D}$ denotes the set of measurable functions from $[0, +\infty)$ to Σ (i.e. switching functions), $\mathcal{U} \coloneqq L^2([0, +\infty), \mathbb{R}^m)$ is the space of $\mathbb{R}^m$-valued control functions, and the state ξ is subject to

$$\dot{\xi}(s) = A^\sigma \xi(s) + B^\sigma u(s) \,, \quad \xi(0) = x \,,$$

where $\sigma = \mu(s)$ denotes the mode that is selected at time s.

In contrast with the programs considered earlier, where the switching occurred as a consequence of the state crossing some hyperplane or a random boolean changing its value, the discrete switching process μ is in this case a parameter on which the user can act.

It is known [McE07] that, under some assumptions on the parameters, the value function V takes finite values and is the unique viscosity solution of the stationary Hamilton-Jacobi-Bellman PDE:

$$H(x, \nabla V) = 0 \,, \quad x \in \mathbb{R}^n \ .$$

The Hamiltonian $H(x, p)$ in the latter equation is the point-wise maximum of simpler Hamiltonians $H^\sigma(x, p)$ given for $\sigma \in \Sigma$ by

$$H^\sigma(x, p) = (A^\sigma x)^T p + \frac{1}{2} x^T D^\sigma x + \frac{1}{2} p^T Q^\sigma p \,,$$

and $Q^\sigma = \gamma^{-2} B^\sigma (B^\sigma)^T$.

We associate with this problem the *Lax-Oleinik* semi-group $\{S_t\}_{t \geqslant 0}$ defined by

$$S_t[V^0](x) = \sup_{u \in \mathcal{U}} \sup_{\mu \in \mathcal{D}} \int_0^t \frac{1}{2} \xi(s)^T D^{\mu(s)} \xi(s) - \frac{\gamma^2}{2} |u(s)|^2 \, ds + V^0\big(\xi(t)\big) \,.$$

It is a semi-group in the sense that $S_t \circ S_s = S_{t+s}$ for all $s, t \geqslant 0$. Moreover, it is known [Mas87, AQV+98] that it is *max-plus linear*:

$$S_t[\max(f, g)] = \max\big(S_t[f], S_t[g]\big) \qquad S_t[f + \lambda] = S_t[f] + \lambda \,,$$

where max denotes the point-wise maximum and λ is a constant.

McEneaney showed in [McE07] that $V(x)$ coincides with $\lim_{t \to +\infty} S_t[V^0](x)$ and that the latter limit is uniform on compact sets if V^0 satisfies a quadratic growth condition (one requires that $\epsilon |x|^2 \leqslant V^0(x) \leqslant \lambda |x|^2$ for some positive constants ϵ, λ that are determined from the parameters).

We also associate with every value $\sigma \in \Sigma$ the semi-group $\{S_t^\sigma\}_{t \geqslant 0}$ corresponding to the unswitched control problem obtained by setting $\mu(s) \equiv \sigma$, i.e.,

$$S_t^\sigma[V^0](x) = \sup_{u \in \mathcal{U}} \int_0^t \frac{1}{2} \xi(s)^T D^\sigma \xi(s) - \frac{\gamma^2}{2} |u(s)|^2 \, ds + V^0\big(\xi(t)\big) \,. \tag{5.1}$$

Computing $S_t^\sigma[V^0]$ when $V^0(x) = x^T P_0 x$, reduces to solving the following indefinite Riccati differential equation,

$$\dot{P} = (A^\sigma)^T P + P A^\sigma + P Q^\sigma P + D^\sigma\,, \quad P(0) = P_0\,,$$

with $P(s) \in \mathcal{S}_n$. Indeed, we have $S_t^\sigma[V^0](x) = x^T P(t) x$. We denote by $\mathrm{ricc}_{t,\sigma}$ the flow of this equation, so that $\mathrm{ricc}_{t,\sigma}[P_0] \coloneqq P(t)$.

This modified control problem is written like a switched system in discrete time in Program 6.

Program 6: Switched control problem

$\mathtt{x} \leftarrow \mathcal{I}$;
while *rand_bool* **do**
  Choose $\sigma \in \Sigma$ and $\mathtt{u} \in L^2([0,t), \mathbb{R}^m)$ that maximize the value of problem (5.1);
  $\mathtt{x} \coloneqq \exp\left[tA^\sigma\right]\mathtt{x} + \int_0^t \exp\left[(t-s)A^\sigma\right] B^\sigma \mathtt{u}(s)\, ds$;
end

5.1.4 McEneaney's curse of dimensionality attenuation scheme

We assume that V^0 is a quadratic function $V^0(x) = x^T P^0 x$. The method of [McE07] that solves the linear quadratic optimal control problem described in Section 5.1.3 approximates the value function V by a finite supremum of quadratic forms

$$V \approx \sup_{\sigma_1,\ldots,\sigma_N \in \Sigma} S_t^{\sigma_1} \cdots S_t^{\sigma_N}[V^0]\,, \tag{5.2}$$

where t is a (small) time discretization step and N is a maximal number of switches. The latter supremum represent the value of a modified optimal control problem, in horizon tN, in which switches occur only at times multiple of t.

The key ingredient in this approach is the explicit computation of $S_t^\sigma[V^0]$ when when $V^0(x) = x^T P_0 x$, reduces to solving the following indefinite Riccati differential equation,

$$\dot{P} = (A^\sigma)^T P + P A^\sigma + P Q^\sigma P + D^\sigma\,, \quad P(0) = P_0\,,$$

with $P(s) \in \mathcal{S}_n$. Indeed, we have $S_t^\sigma[V^0](x) = x^T P(t) x$. We denote by $\mathrm{ricc}_{t,\sigma}$ the flow of this equation, so that $\mathrm{ricc}_{t,\sigma}[P_0] \coloneqq P(t)$.

We have $S_t^{\sigma_1} \cdots S_t^{\sigma_N}[V^0](x) = x^T Q x$, where $Q = \mathrm{ricc}_{t,\sigma_1} \circ \cdots \circ \mathrm{ricc}_{t,\sigma_N}(P_0)$, can be computed by integrating successive Riccati equations, which allows us to evaluate the expression in Equation (5.2).

The propagation of a quadratic form by the Lax-Oleinik semi-group has only a cubic cost in terms of the dimension n, contrary to classical grid-based methods whose cost is exponential in the dimension. In this sense, the curse of dimensionality has been reduced. However, the memory footprint of this method is exponential in the number of switches, since m^N quadratic forms are computed after N iterations. Several pruning schemes have been proposed in [GMQ11] to limit this growth. This is a costly operation, indeed, 99% of the computation time is spent solving LMIs inside the pruning procedure [GMQ11].

5.2 The space of ellipsoids

5.2.1 Uncentered ellipsoids

In this section, we introduce the domain of uncentered ellipsoids and several operations on ellipsoids that are needed in our analysis. Some definitions have already been given in Chapter 3; they are repeated here for the sake of completeness. For readability, we drop the term "uncentered" if it is clear from the context.

Let $\mathcal{B}_n = \{x \in \mathbb{R}^n \colon x^T x \leqslant 1\}$ denote the unit ball of $\mathbb{R}^n$. An *uncentered ellipsoid* is defined as the image of the unit ball $\mathcal{B}_n$ under an affine map $x \mapsto Lx + q$, where $L \in \mathcal{M}_n$ and $q \in \mathbb{R}^p$. When the matrix L is invertible, the matrix $Q = LL^T$ is positive definite and the ellipsoid $\mathcal{E}(Q, q)$ is given by

$$\mathcal{E}(Q, q) \coloneqq \{y \in \mathbb{R}^n \colon (y - q)^T Q^{-1} (y - q) \leqslant 1\}.$$

When the matrix L is not invertible, the matrix Q is only positive semi-definite, so we have $\mathcal{E}(Q, q) = \{y \in \mathbb{R}^p \colon (y - q)(y - q)^T \preccurlyeq Q\}$ as a consequence of Lemma 3.8. When the matrix Q is positive definite, the ellipsoid $\mathcal{E}(Q, q)$ is full-dimensional, whereas the ellipsoid $\mathcal{E}(Q, q)$ is "flat" when Q is only positive semi-definite. The volume of a full-dimensional ellipsoid is proportional to $\det L = (\det Q)^{1/2}$ (the proportionality constant only depends on the dimension).

We denote the set of uncentered ellipsoids in $\mathbb{R}^n$ by $\mathfrak{E}_n$. This set is equipped with the inclusion order. By extension of Theorem 2.1 is constitutes an anti-lattice.

5.2.2 The Löwner ellipsoid

Let us recall a famous result by Löwner in the setting of uncentered ellipsoids.

Theorem 5.2 (Löwner [Bus50, Bal97])**.** *Given a compact and full-dimensional set $\mathcal{X} \subset \mathbb{R}^n$, there is a unique ellipsoid $\mathcal{E}(Q, q)$ that contains $\mathcal{X}$ and that has minimum volume.*

This ellipsoid is called the *Löwner ellipsoid* of the set $\mathcal{X}$, and we denote it by $\mathrm{Löw}(\mathcal{X})$. The Löwner ellipsoid has a special property.

Proposition 5.3. *The Löwner ellipsoid commutes with invertible affine transformations: given a compact full-dimensional $\mathcal{X} \subset \mathbb{R}^n$ and an invertible affine map f, we have*

$$f\big(\mathrm{Löw}(\mathcal{X})\big) = \mathrm{Löw}\big(f(\mathcal{X})\big).$$

Proof. Let $\mathcal{E}$ denote an ellipsoid containing $\mathcal{X}$. We denote the invertible affine map f by $x \mapsto Ax + b$. The map f is monotone, so that $\mathcal{E} \supseteq f(\mathcal{X}) \iff f^{-1}(\mathcal{E}) \supseteq \mathcal{X}$. Moreover, the volume of the ellipsoid $\mathcal{E}$ is equal to the volume of the ellipsoid $f^{-1}(\mathcal{E})$ multiplied by $\det A$. Combined with the uniqueness of the Löwner ellipsoid, we deduce that $\mathcal{E} = \mathrm{Löw}\big(f(\mathcal{X})\big)$ if and only if $f^{-1}(\mathcal{E}) = \mathrm{Löw}(\mathcal{X})$. □

The affine automorphism group $\mathrm{Aut}\,\mathcal{X}$ of a set $\mathcal{X}$ is the set of affine transformations $T \colon x \mapsto Ax + c$ that leave $\mathcal{X}$ invariant, i.e. $T(\mathcal{X}) = \mathcal{X}$. The linear automorphism group is defined similarly for linear transformations. Danzer, Laugwitz and Lenz have shown in [DLL57] that the Löwner ellipsoid $\mathcal{E}$ of a set $\mathcal{X}$ retains the symmetry properties of the set $\mathcal{X}$.

Theorem 5.4. *The group of affine automorphisms of a full-dimensional set $\mathcal{X}$ is a subset of the group of automorphisms of* $\mathrm{Löw}\,\mathcal{X}$*:*

$$\mathrm{Aut}\,\mathcal{X} \subseteq \mathrm{Aut}\,\mathrm{Löw}(\mathcal{X})\,.$$

Proof. This is a direct consequence of Proposition 5.3. Given $f \in \mathrm{Aut}\,\mathcal{X}$, we have $f\big(\mathrm{Löw}\,\mathcal{X}\big) = \mathrm{Löw}\,\big(f(\mathcal{X})\big) = \mathrm{Löw}\,\mathcal{X}$, hence $f \in \mathrm{Aut}\,\mathrm{Löw}\,\mathcal{X}$. □

5.2.3 Operations on uncentered ellipsoids

We now describe operations on ellipsoids that are useful in the sequel.

5.2.3.a Testing the inclusion of ellipsoids We can check if an ellipsoid $\mathcal{E}(Q_1, q_1)$ is included in the ellipsoid $\mathcal{E}(Q_2, q_2)$. In the case where the ellipsoid $\mathcal{E}(Q_2, q_2)$ is full dimensional, this problem is equivalent to checking if an LMI has a solution: it contains the ellipsoid $\mathcal{E}(Q_1, q_1)$ if and only there is some real number λ such that When the matrix Q_2 is positive definite, checking the inclusion $\mathcal{E}(Q_1, q_1) \subseteq \mathcal{E}(Q_2, q_2)$ is equivalent to checking if the following LMI has a solution as proved in [BTN01]:

$$\mathcal{E}(Q_1, q_1) \subseteq \mathcal{E}(Q_2, q_2) \iff \exists \lambda \in \mathbb{R}\colon \begin{pmatrix} Q_2 & q_1 - q_2 & L_1 \\ (q_1 - q_2)^T & 1 - \lambda & 0_{1,n} \\ L_1^T & 0_{n,1} & \lambda I_n \end{pmatrix} \succcurlyeq 0\,, \tag{5.3}$$

where L_1 satisfies $Q_1 = L_1 L_1^T$. Checking the inclusion is thus equivalent to solving an LMI. In the special case where $q_1 = q_2$, this amounts to checking if $Q_1 \preccurlyeq Q_2$. We shall see later on that the case where the ellipsoid $\mathcal{E}(Q_2, q_2)$ is not full dimensional does not arise in our analysis. If $\mathcal{E}(Q_2, c_2)$ is not full-dimensional, it is first necessary to check that $\mathrm{ran}\,Q_1 \subseteq \mathrm{ran}\,Q_2$ and that $c_2 - c_1 \in \mathrm{ran}\,Q_2$. If this is true, the inclusion between the ellipsoids holds if and only if the inclusion of the image of those ellipsoids under the projection onto $\mathrm{ran}\,Q_2$ holds.

5.2.3.b Image under an affine map The image of an ellipsoid by an affine map is again an ellipsoid:

$$(x \mapsto Ax + c) \cdot \mathcal{E}(Q, q) = \mathcal{E}(AQA^T, Aq + c)\,.$$

Indeed, let L such that $Q = LL^T$. We know that $\mathcal{E}(Q, q)$ and $\mathcal{E}(AQA^T, Aq + c)$ respectively correspond to the image of the unit ball $\mathcal{B}_n$ under the affine maps $x \mapsto Lx + q$ and $x \mapsto A(Lx + q) + c$. The expected result follows straightforwardly.

5.2.3.c Union of ellipsoids We over-approximate the union of a finite number of ellipsoids by the Löwner ellipsoid $\mathrm{Löw}\,\big(\cup_k \mathcal{E}_k\big)$. The latter can be computed as $\mathcal{E}(Y^{-2}, Y^{-1}y)$, where (Y, y) is the optimal solution of the following semi-definite program [BTN01]:

$$\begin{aligned} \underset{Y,y}{\mathrm{argmin}} \quad & -\log\det Y \\ \text{subject to} \quad & \begin{pmatrix} I_n & (Yq_k - y) & YL_k \\ (Yq_k - y)^T & 1 - \lambda_k & 0_{1,n} \\ L_k^T Y & 0_{n,1} & \lambda_k I_n \end{pmatrix} \succcurlyeq 0\ ,\ \forall k \\ & Y \succcurlyeq 0 \end{aligned} \tag{5.4}$$

and $\mathcal{E}_k = \mathcal{E}(L_k L_k^T, q_k)$. This is only true when the convex hull of $\cup_k \mathcal{E}_k$ is full-dimensional (the opposite case will not arise in our analysis). For the sake of brevity, we denote this Löwner ellipsoid by

$$\sqcup_k \mathcal{E}_k = \mathcal{E}_1 \sqcup \cdots \sqcup \mathcal{E}_p \coloneqq \text{Löw}\left(\cup_k \mathcal{E}_k\right). \tag{5.5}$$

It will be convenient to write in infix form, $\mathcal{E}_1 \sqcup \cdots \sqcup \mathcal{E}_p$ instead of $\sqcup_k \mathcal{E}_k$, noting that this is an abuse of notation, since the operation $\sqcup$ is not associative. , i.e., $\mathcal{E}_1 \sqcup (\mathcal{E}_2 \sqcup \mathcal{E}_3) \neq (\mathcal{E}_1 \sqcup \mathcal{E}_2) \sqcup \mathcal{E}_3$.

This notation is coherent with the notation $\sqcup$ introduced in Chapter 3. Indeed, given ellipsoids ($\mathcal{E}(Q_i, q)$ that have the same center q, their Löwner ellipsoid also has the same center, since the Löwner ellipsoid commutes with the affine map $x \mapsto 2q - x$ and the ellipsoids $\mathcal{E}(Q_i, q)$ are left invariant by this map.

5.2.3.d Minkowski sum of ellipsoids Recall that the Minkowski sum of two sets X, Y is the set $X + Y \coloneqq \{x + y \colon x \in X, y \in Y\}$. and that the Minkowski sum of two ellipsoids may not be an ellipsoid. For instance, the Minkowski sum of the two flat ellipsoids $\mathcal{E}(\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right), 0)$ and $\mathcal{E}(\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right), 0)$ is the unit square $\{(x, y) \in \mathbb{R}^2 \colon \max(|x|, |y|) \leqslant 1\}$. Similarly, replacing the previous two ellipsoids by "nearly flat" yields a slightly rounded square. Like the case of the union of ellipsoids, we over-approximate the Minkowski sum of two ellipsoids $\mathcal{E}_0, \mathcal{E}_1$ by its Löwner ellipsoid, and denote it by

$$\mathcal{E}_0 \boxplus \mathcal{E}_1 \coloneqq \text{Löw}\left(\mathcal{E}_0 + \mathcal{E}_1\right). \tag{5.6}$$

It has been shown in [BTN01] that, given two full-dimensional ellipsoids $\mathcal{E}(Q_1, q_1)$ and $\mathcal{E}(Q_2, q_2)$, the Löwner ellipsoid of $\mathcal{E}(Q_1, q_1) + \mathcal{E}(Q_2, q_2)$ is then equal to $\mathcal{E}(Z^{-1}, q_1 + q_2)$, where Z is the solution of the semi-definite program

$$\begin{aligned} \underset{Z,\lambda}{\text{argmin}} \quad & -\log\det Z \\ \text{subject to} \quad & \begin{pmatrix} \lambda Q_1^{-1} & 0_n \\ 0_n & (1-\lambda) Q_2^{-1} \end{pmatrix} \succcurlyeq \begin{pmatrix} Z & Z \\ Z & Z \end{pmatrix} \\ & Z \succcurlyeq 0,\ 0 \leqslant \lambda \leqslant 1\,. \end{aligned} \tag{5.7}$$

5.2.3.e Intersection of an ellipsoids with a half-space We denote by $\mathcal{H}(f, g)$ the half-space $\{x \in \mathbb{R}^n \colon f^T x \leqslant g\}$, where $f \in \mathbb{R}^n$ is a non-zero vector and g is a real number. In general, the intersection of an ellipsoid $\mathcal{E}$ with a half-space $\mathcal{H}$ is not an ellipsoid. Given an ellipsoid $\mathcal{E}$, we over-approximate its intersection with the half-space $\mathcal{H}$ by its Löwner ellipsoid, and we denote it by

$$\mathcal{E} \sqcap \mathcal{H} \coloneqq \text{Löw}\left(\mathcal{E} \cap \mathcal{H}\right).$$

When $\mathcal{E}$ is full-dimensional (we shall see later that it is always the case in our computations), the set $\mathcal{E} \sqcap \mathcal{H}$ can be computed analytically following [BGT81]. We give the formula below for the sake of completeness. Given an ellipsoid $\mathcal{E}(Q, q)$ and a half-space $\mathcal{H}(f, g)$, let α denote the quantity $\alpha \coloneqq (f^T Q f)^{-1}(g - f^T q)$. If $\alpha \geqslant 1/n$, we have $\mathcal{E} \sqcap \mathcal{H} = \mathcal{E}$. If $\alpha < -1$, then $\mathcal{E} \cap \mathcal{H} = \emptyset$. If $\alpha = -1$, the ellipsoid $\mathcal{E} \sqcap \mathcal{H}$ is reduced to the point q. If $-1 \leqslant \alpha \leqslant 1/n$, we have $\mathcal{E} \sqcap \mathcal{H} = \mathcal{E}(Q^+, q^+)$, with $q^+ = q - (1+n)^{-1}(1 - n\alpha) Q f$ and

$$Q^+ = \frac{n^2(1-\alpha^2)}{n^2 - 1}\Big(Q - 2\frac{1 - n\alpha}{(1+n)(1-\alpha)}(Qf)(Qf)^T\Big).$$

The intersection with several half-spaces is handled in a sequential way, meaning that we evaluate Löw $\big(\mathcal{E} \cap (\mathcal{H} \cap \mathcal{H}')\big)$ as $(\mathcal{E} \sqcap \mathcal{H}) \sqcap \mathcal{H}'$. It will again be convenient to do an abuse of notation, denoting the latter operation by $\mathcal{E} \sqcap (\mathcal{H} \cap \mathcal{H}')$. Although this evaluation remains sound, it may yield a very coarse over-approximation, since the maps $\mathcal{E} \mapsto \mathcal{E} \sqcap \mathcal{H}$ and $\mathcal{E} \mapsto \mathcal{E} \sqcap \mathcal{H}'$ do not commute in general. When several half-spaces are involved, Unfortunately, there is no tractable way to get finding a better over-approximation is a difficult and intractable problem, see [BTN01, Section 3.7].

5.2.4 Centered ellipsoids: definitions and operations

The applications presented in Sections 5.1.2 and 5.1.3 have linear dynamics and the invariants considered are centrally symmetric around the origin (we refer to such sets as *symmetric sets*). We specialize the former framework to centered ellipsoids to benefit from this symmetry property. We say that an ellipsoid $\mathcal{E}(Q, q)$ is *centered* if $q = 0$. It is then denoted by $\mathcal{E}(Q)$ and

$$x \in \mathcal{E}(Q) \iff xx^T \preccurlyeq Q\,. \tag{5.8}$$

The Löwner ellipsoid is left invariant by the affine automorphism group of its enclosing set. The same property holds in particular for the linear automorphism group of if the initial set is symmetric, whose Löwner ellipsoid is centered. Hence all operations defined with the Löwner ellipsoid in the affine case yield centered ellipsoids when applied on centered ellipsoids.

5.2.4.a Testing the inclusion of ellipsoids The LMI defined in Equation (5.3) reduces in the case of centered ellipsoids $\mathcal{E}(Q_1), \mathcal{E}(Q_2)$ to checking if the positive semidefinite matrices Q_1, Q_2 are comparable in the Löwner order:

$$\mathcal{E}(Q_1) \subseteq \mathcal{E}(Q_2) \iff Q_1 \preccurlyeq Q_2\,.$$

5.2.4.b Image under an affine map The image of a centered ellipsoid by an affine map is again a centered ellipsoid:

$$A \cdot \mathcal{E}(Q) = \mathcal{E}(AQA^T)\,.$$

5.2.4.c Union of ellipsoids We over-approximate the union of a finite number of centered ellipsoids by the Löwner ellipsoid Löw $\big(\cup_k \mathcal{E}_k\big)$. The latter can be computed as $\sqcup_k \mathcal{E}_k \coloneqq \mathcal{E}(Z^{-1})$, where Z is the optimal solution of the following semi-definite program [BTN01]:

$$\begin{aligned} \underset{Z}{\text{argmin}} \quad & -\log\det Z \\ \text{subject to} \quad & Z \preccurlyeq Q_k^{-1} \\ & Z \succcurlyeq 0 \end{aligned} \tag{5.9}$$

when the ellipsoids $\mathcal{E}_k = \mathcal{E}(Q_k)$ are full-dimensional. Otherwise, we refer to the semi-infinite program defined in Section 3.3. When only two ellipsoids are involved, it can be computed using Equation (3.4).

5.2.4.d Minkowski sum of ellipsoids The Minkowski sum of two symmetric sets is also symmetric, hence

$$\mathcal{E}(Q_1) \boxplus \mathcal{E}(Q_2) \coloneqq \text{Löw}\left(\mathcal{E}_0 + \mathcal{E}_1\right) = \mathcal{E}(Z^{-1})\,,$$

where Z is the solution of the semi-definite program in Equation (5.7).

5.2.4.e Intersection of an ellipsoids with a half-space This operation does not arise in the centered case.

5.3 Abstract interpretation for switched affine programs

Abstract interpretation describes a framework in which two semantics describing of a same system or program co-exist. The first description is *concrete* and usually mimics very closely the behavior of the original system/program. In Section 5.3.1 we exhibit a collecting semantics that simply extends the definition of the program's semantics to sets of variables. It is often not computationally feasible to handle objects in this representation directly. This role is served by the second description, dubbed *abstract* semantics, which has a structure well suited for computational purposes. The key ingredient is a so-called "concretization operator" that maps abstract objects to concrete ones. Once several technical requirements are satisfied, the interplay of those semantics allows one to disregard the concrete objects and solely manipulate abstract elements. In particular, for our purposes, the search for an invariant set is "lifted" to the abstract domain.

We recall in this section the construction of the abstract domain of *lower sets of ellipsoids* in the case of switched affine systems.

5.3.1 A collecting semantics

We extend the definition of the three main operations that occur in our implementations of switched systems to act on bounded subsets of $\mathbb{R}^n$. These operations are the assignments of a vector variable, testing an `if` statement guarded by a conjunction of affine inequalities or a random integer value and a while loop with non-deterministic exit.

In the sequel, the set $\mathcal{X}$ belongs to the complete lattice $\wp^{\text{bounded}}(\mathbb{R}^n) \cup \{\infty\}$ of bounded subsets of $\mathbb{R}^n$ to which we have added an element identifying non-bounded sets.

Affine assignment The action of an affine assignment of variables $\mathtt{x} \coloneqq A_\sigma \mathtt{x} + c_\sigma$ on a concrete element $\mathcal{X}$, i.e. a (bounded) subset of $\mathbb{R}^n$, is denoted by $[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + c_\sigma]\!]$ and defined by

$$[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + c_\sigma]\!](\mathcal{X}) \coloneqq \{A_\sigma x + c_\sigma \colon x \in \mathcal{X}\}\,.$$

If this affine assignment is rather of the form $\mathtt{x} \coloneqq A_\sigma \mathtt{x} + B_\sigma \mathtt{u} + c_\sigma$ where $\mathtt{u}$ takes values in the set $\mathcal{U}$, the concrete operator is

$$[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + B_\sigma \mathtt{u} + c_\sigma]\!](\mathcal{X}) \coloneqq \left\{A_\sigma x + c_\sigma \colon x \in \mathcal{X}\right\} + \left\{B_\sigma u \colon u \in \mathcal{U}\right\}. \tag{5.10}$$

If-then-else and switch statements Let us first consider the case of Program 3. We have assumed that the conditions are mutually exclusive, hence only one assignment operation occurs on the variable y in each loop iteration. Assume that condition $\sigma \in \Sigma$ is satisfied and that some operator op_σ acts within this branch (in our case, op_σ is an affine assignment). Then the concrete operation associated with this branch is

$$[\![\texttt{if x} \in \bigcap_l \mathcal{H}_{\sigma,l} \texttt{ then } \mathrm{op}_\sigma(\texttt{x}); \texttt{ end}]\!](\mathcal{X}) = \left\{ \mathrm{op}_\sigma(x) \colon x \in \mathcal{X} \text{ and } x \in \bigcap_l \mathcal{H}_{\sigma,l} \right\},$$

It is then necessary to merge (or join) the branches, hence the concrete operator for the whole switch statement takes the form:

$$[\![\texttt{switch } \sigma \texttt{ ; case(x} \in \bigcap_l \mathcal{H}_{\sigma,l}) \colon \mathrm{op}_\sigma(\texttt{x});]\!](\mathcal{X}) = \bigcup_\sigma \left\{ \mathrm{op}_\sigma(x) \colon x \in \mathcal{X} \text{ and } x \in \bigcap_l \mathcal{H}_{\sigma,l} \right\}.$$

Loops The body of the loops in the programs of interest contain a sequence of mutually exclusive "if-then" statements, in which a single vector-assignment is performed. The concrete operator s corresponding to the body of the loop of Program 3 is given by the map

$$s \colon \mathcal{X} \mapsto \bigcup_{\sigma \in \Sigma} \left\{ A_\sigma x + c_\sigma \colon x \in \mathcal{X} \cap \bigcap_l \mathcal{H}_{\sigma,l} \right\} + \left\{ B_\sigma u \colon u \in \mathcal{U} \right\}. \tag{5.11}$$

We deduce the expression of the loop-body abstraction for Program 4

$$s \colon \mathcal{X} \mapsto \bigcup_{\sigma \in \Sigma} \left\{ A_\sigma x + c_\sigma \colon x \in \mathcal{X} \right\} + \left\{ B_\sigma u \colon u \in \mathcal{U} \right\}.$$

The least fixed point of the operator $[\![p]\!] \colon \mathcal{X} \mapsto \mathcal{I} \cup s(\mathcal{X})$ then provides a invariant for the switched affine program. The existence of this least fixed point is a consequence of the monotonicity of the concrete operators and the Knaster-Tarski theorem [Tar55], see [Rou13] for a complete proof. In fact, any post-fixed point of the operator $[\![p]\!]$, i.e. a set $\mathcal{X}$ such that $[\![p]\!](\mathcal{X}) \subseteq \mathcal{X}$ is an invariant.

5.3.2 The abstract domain of lower sets

5.3.2.a Lower sets: definitions and properties We begin by recalling the definition of lower sets.

Given a partially ordered set $(\mathcal{X}, \leqslant)$, a subset L is an *lower set* is $x \in L$ and $y \geqslant x$ implies $y \in L$. The set of lower sets in $\mathcal{X}$ is denoted[1] by $\mathrm{O}(\mathcal{X})$. The reunion and the intersection of two lower sets is again an lower set, and these operations endow the space $\big(\mathrm{O}(\mathcal{X}), \subseteq\big)$ with a complete lattice structure, with

$$\sup_i L_i = \cup_i L_i \quad \text{and} \quad \inf_i L_i = \cap_i L_i \,.$$

An element x is a *maximal element* of a lower set L if $y \in L$ and $x \leqslant y$ imply $x = y$. A lower set L is *principal* if it has a unique maximal element l, in which case

$$L = \downarrow \{l\} \coloneqq \{x \in \mathcal{X} \colon x \leqslant l\}.$$

[1] This is a standard notation, not to be confused with an orthogonal group $\mathcal{O}(n)$.

More generally, the set $\downarrow S$ refers to the lower set generated by S:

$$\downarrow S = \bigcup_{s\in S} \downarrow \{s\}\,.$$

The set of maximal elements of a lower set L is denoted by $\operatorname{Max} L$. The latter set constitutes an *antichain*, i.e. two distinct elements of $\operatorname{Max} L$ are never comparable. Moreover, these maximal elements fully characterize the lower set L:

$$L = \downarrow (\operatorname{Max} L)\,.$$

5.3.2.b Application in abstract interpretation Lower sets provide a good abstract domain in the setting of abstract interpretation, see [CC94]. Given a lattice $(\mathcal{D}, \preccurlyeq)$ and a monotone concretization map $\gamma\colon \mathrm{O}(\mathcal{X}) \mapsto \mathcal{D}$, the operator $\alpha\colon \mathcal{D} \mapsto \mathrm{O}(\mathcal{X})$ defined for $d \in \mathcal{D}$ by

$$\alpha(d) \coloneqq \sup\{L \in \mathrm{O}(\mathcal{X})\colon \gamma(L) \subseteq d\}$$

defines a monotone abstraction operator. Indeed, let $d, e \in \mathcal{D}$ such that $d \preccurlyeq e$. Then $\gamma\big[\alpha(d)\big] \preccurlyeq d \preccurlyeq e$, hence by definition of α we have $\alpha(d) \supseteq \alpha(e)$. As a consequence, the pair (α, γ) defines a *Galois connection* between $\mathcal{D}$ and $\mathrm{O}(\mathcal{X})$:

$$\alpha(d) \leqslant L \iff d \preccurlyeq \gamma(L)\,,$$

see [CC94] for more background.

In our work we choose $(\mathcal{X}, \leqslant) = \big(\mathfrak{E}_n, \subseteq\big)$ and $\mathcal{D}$ the lattice of bounded subsets of $\mathbb{R}^n$ in the inclusion order. The concretization γ maps an lower set of ellipsoids to their reunion in $\mathbb{R}^n$ and the abstraction α returns the upper set of ellipsoids that are subsets of a bounded set

$$\gamma(L) \coloneqq \cup\{\mathcal{E} \in L\} \quad \text{and} \quad \alpha(S) \coloneqq \{\mathcal{E} \in \mathfrak{E}_n\colon \mathcal{E} \subseteq S\}\,.$$

Now, we can define the counterparts of the operations introduced in Section 5.3.1 on the abstract domain $\mathrm{O}(\mathfrak{E}_n)$. Since the pair (α, γ) is a Galois connection, then Cousot and Cousot showed in [CC77a] that the abstract counterpart $[\![\cdot]\!]^\#$ of each operator $[\![\cdot]\!]$ is simply given by

$$[\![\cdot]\!]^\# \coloneqq \alpha \circ [\![\cdot]\!] \circ \gamma\,.$$

Moreover these operators are set-valued hence they are monotone in the inclusion order, hence the abstract operator corresponding to the body of the loop of Program 3 is monotone and given by

$$s^\#(L) \coloneqq \alpha\Big[\bigcup_{\sigma\in\Sigma} \big\{A_\sigma x + c_\sigma\colon x \in \gamma(L) \cap \bigcap_l \mathcal{H}_{\sigma,l}\big\} + \big\{B_\sigma u\colon u \in \mathcal{U}\big\}\Big]$$

An *abstract* invariant is a lower set L such that $s^\#(L) \subseteq L$ and $\mathcal{I} \subseteq \cup\{\mathcal{E} \in L\}$. Such an invariant can be computed by a Kleene iteration scheme. Let $L_\mathcal{I} \in \mathrm{O}(\mathfrak{E}_n)$ such that $\gamma(L_\mathcal{I}) \supseteq \mathcal{I}$. Then the iteration

$$L^0 = L_\mathcal{I} \tag{5.12}$$

$$L^{k+1} = L_\mathcal{I} \cup s^\#(L^k) \tag{5.13}$$

converges towards the least fixed point of the map $s^\#$ that contains $L_\mathcal{I}$. Indeed, the sequence L^k is increasing and bounded above by thus it converges and its limit L must satisfy $L_\mathcal{I} \subseteq L$ and $s^\#(L) \subseteq L$.

CHAPTER 6

Disjunctive ellipsoidal invariants for switched affine systems

This chapter is based on the articles "A Scalable Algebraic Method to Infer Quadratic Invariants of Switched Systems" [AGS+16] and "A fast method to compute disjunctive quadratic invariants of numerical programs" [AGG+17]

6.1 Introduction

6.1.1 Context

Although the abstraction by lower sets presented in Chapter 5 has the same expressivity as subsets of $\mathbb{R}^n$, it is not possible to implement. We tackle this issue by discretizing the space of lower sets $O(\mathfrak{E}_n)$ by considering only lower sets that arise as the reunion of N principal lower sets, for some integer N. In other words, our abstract elements can be identified vectors of ellipsoids and the concretization map becomes $\gamma\colon (\mathcal{E}_i)_{1\leqslant i\leqslant N} \mapsto \cup_i \mathcal{E}_i$, i.e., we abstract a subset of $\mathbb{R}^n$ by its cover in terms of N ellipsoids. This is no longer an exact abstraction but it is still defined for every bounded set. Moreover, although this approach has the key advantage of being implementable, it lacks several important ingredients. First, it no longer constitutes a lattice. Second, the abstraction operator is no longer defined. Hence the abstract operators acting on this space are no longer defined like in Section 5.3.2.

6.1.2 Contribution

We provide alternatives to each of these problems by overapproximating concrete elements by their Löwner ellipsoids. We use these primitives in two algorithms that compute invariants as unions of ellipsoids. The first algorithm deals with affine switched systems and consists in

a modified Kleene iteration: a small perturbation parameter is added to ensure convergence in finite time and robustness to numerical imprecisions. Second, a power-like algorithm is introduced to compute invariants of linear systems. We provide numerical benchmarks for each algorithm and compare the results with state of the art methods.

6.2 The domain of unions of ellipsoids

6.2.1 Definitions and notation

In the sequel, we consider a switched program like one presented in Programs 3 to 5.[1]. We label the branches of the loop of the program by integers from 1 to p, and denote by $\Sigma \coloneqq \{1, \ldots, p\}$. We can thus identify the set of finite traces of the program with the set Σ^* of finite words built on the alphabet Σ. Let $\mathcal{W}$ denote a subset of Σ^* with cardinal N and τ denote a map from $\mathcal{W} \times \Sigma$ to $\mathcal{W}$. The triple $(\Sigma, \mathcal{W}, \tau)$ defines a deterministic finite automaton, whose alphabet is Σ, whose states are elements of $\mathcal{W}$ and whose transition function is τ. Every state in this automaton is both an initial and final state. The function τ being totally defined over $\mathcal{W} \times \Sigma$, the automaton $(\Sigma, \mathcal{W}, \tau)$ accepts every word of Σ^*. We say that $(i, \sigma, j) \in \mathcal{W} \times \Sigma \times \mathcal{W}$ is an *admissible transition* when $\tau(i, \sigma) = j$.

We are now ready to present union of ellipsoids in the framework of abstract interpretation. Following the terminology of abstract interpretation, our concrete domain is defined as $\wp^{\text{bounded}}(\mathbb{R}^n)$, the lattice of bounded subsets of $\mathbb{R}^n$ equipped with the subset partial order $\subseteq$.

The abstract domain is the set of functions $\underline{\mathcal{E}}$ from $\mathcal{W}$ to the set of ellipsoids $\mathfrak{E}_n$. We denote by $\mathcal{E}_w$ the ellipsoid associated with $w \in \mathcal{W}$ by this function. It will be convenient to identify $\underline{\mathcal{E}}$ to the vector $(\mathcal{E}_w)_{w \in \mathcal{W}}$ in $\mathfrak{E}_n^{\mathcal{W}}$ indexed by elements of w. Operators defined on the abstract domain $\mathfrak{E}_n^{\mathcal{W}}$ are vector mappings. Hence the w-th coordinate of the abstract operator $[\![\cdot]\!]^\#$ is denoted by $[\![\cdot]\!]_w^\#$.

We equip the abstract domain $\mathfrak{E}_n^{\mathcal{W}}$ with the coordinate-wise ordering:

$$\underline{\mathcal{E}}^1 \sqsubseteq \underline{\mathcal{E}}^2 \iff \mathcal{E}_w^1 \subseteq \mathcal{E}_w^2 \ , \ \forall w \in \mathcal{W} \, .$$

This is, up to a permutation in the indexes w, the same order on the lower sets generated by the reunion of $\downarrow \{\mathcal{E}_w\}$ for $w \in \mathcal{W}$.

The concretization operator γ, which maps an abstract element to a concrete one, is defined as the function $\gamma : \mathfrak{E}_n^{\mathcal{W}} \to \wp^{\text{bounded}}(\mathbb{R}^n)$ which associates a vector of ellipsoids $\underline{\mathcal{E}}$ to the union of these ellipsoids in $\mathbb{R}^n$. The concretization operator γ is order-preserving:

Lemma 6.1. *The map $\gamma \colon \underline{\mathcal{E}} \mapsto \cup_w \mathcal{E}_w$ is order-preserving.*

Proof. Given $\underline{\mathcal{E}}^1 \sqsubseteq \underline{\mathcal{E}}^2$, let $x \in \gamma(\underline{\mathcal{E}}^1)$, i.e., $x \in \mathcal{E}_w^1$ for some $w \in \mathcal{W}$. Hence $x \in \mathcal{E}_w^2$ and $x \in \gamma(\underline{\mathcal{E}}^2)$. □

Like several other abstract domains (convex polyhedra [CH78], zonotopes [GGP09], etc.), the domain of unions of ellipsoids cannot be equipped with an abstraction operator α. Indeed, this operator would be supposed to map any bounded subset S to the "smallest" vector of ellipsoids $\underline{\mathcal{E}}$ such that $S \subseteq \cup_w \underline{\mathcal{E}}_w$.[2] Such an element does not exist in general, since it implies for instance that the boundary of the closure of S is the reunion of finitely many quadratic hypersurfaces.

[1] Program 6 is treated separately.

[2] In this case, (α, γ) forms a *Galois connection*.

6.2.2 Affine assignment

Given an abstract element $\underline{\mathcal{E}}$, the abstract operator corresponding to the affine assignment of variables $\mathtt{x} \coloneqq A_\sigma \mathtt{x} + c_\sigma$, denoted by $[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + c_\sigma]\!]^\#$, is the coordinate-wise affine transformation of ellipsoids:

$$[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + c_\sigma]\!]^\#_v\big(\underline{\mathcal{E}}\big) \coloneqq \big(x \mapsto A_\sigma x + c_\sigma\big) \cdot \mathcal{E}_v \,.$$

Indeed, affine transformations are order-preserving. Hence, given a lower set of the form $\downarrow \{\mathcal{E}_w\}_{w\in\mathcal{W}}$, its image by the abstract affine assignment operator is exactly the lower set $\downarrow \big\{(x \mapsto A_\sigma x + c_\sigma) \cdot \mathcal{E}_w\big\}_{w\in\mathcal{W}}$

Lemma 6.2. *The abstract assignment operator $[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + c_\sigma]\!]^\#$ is exact:*

$$[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + c_\sigma]\!]\big[\gamma(\underline{\mathcal{E}})\big] = \gamma\Big[[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + c_\sigma]\!]^\#(\underline{\mathcal{E}})\Big] \,.$$

Proof. The image of an ellipsoid by an affine map is again an ellipsoid, hence

$$[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + c_\sigma]\!]\big[\gamma(\underline{\mathcal{E}})\big] = (x \mapsto A_\sigma x + c_\sigma) \cdot \bigcup_v \mathcal{E}_v = \gamma\big[[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + c_\sigma]\!]^\#(\underline{\mathcal{E}})\big] \qquad \square$$

If this affine assignment is rather of the form $\mathtt{x} \coloneqq A_\sigma \mathtt{x} + B_\sigma \mathtt{u} + c_\sigma$ where $\mathtt{u}$ takes values in the set $\mathcal{U}$ (that is not reduced to a single point), the abstract operator defined in Equation (5.10) no longer preserves finitely generated lower sets, as Minkowski sums of ellipsoids are generically not finite reunion of ellipsoids, see [KV06]. Hence we choose a coordinate-wise over-approximation of the Minkowski sum, using the operator $\boxplus$ defined in Equation (5.6). This yields the the abstract operator

$$[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + B_\sigma \mathtt{u} + c_\sigma]\!]^\#_v\big(\underline{\mathcal{E}}\big) \coloneqq \big(x \mapsto A_\sigma x + c_\sigma\big) \cdot \mathcal{E}_v \;\; \boxplus \;\; \big(u \mapsto B_\sigma u\big) \cdot \mathcal{E}_{\mathcal{U}} \,,$$

where $\mathcal{E}_{\mathcal{U}}$ is the Löwner ellipsoid of the set $\mathcal{U}$.

This operator is not exact due to the approximation with $\boxplus$. However it provides a sound over-approximation of the concrete version.

Lemma 6.3. *The affine assignment operator $[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + B_\sigma \mathtt{u} + c_\sigma]\!]^\#$ is sound:*

$$[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + B_\sigma \mathtt{u} + c_\sigma]\!]\big[\gamma(\underline{\mathcal{E}})\big] \subseteq \gamma\Big[[\![\mathtt{x} \coloneqq A_\sigma \mathtt{x} + B_\sigma \mathtt{u} + c_\sigma]\!]^\#(\underline{\mathcal{E}})\Big] \,.$$

Proof. By definition of $\boxplus$, we have $\mathcal{E}_1 + \mathcal{E}_2 \subseteq \mathcal{E}_1 \boxplus \mathcal{E}_2$. The result follows in a straightforward way. $\square$

6.2.3 If-then-else and switch statements

We have already argued to the intersection of an ellipsoid with a half-space may not be an ellipsoid, hence it is also the case for union of ellipsoids intersected with multiple half-spaces. In other words, we cannot expect the abstract counterpart of the guard condition $[\![\mathtt{if}\ \mathtt{x} \in \bigcap_l \mathcal{H}_{\sigma,l}\ \mathtt{then}\ \mathrm{op}_\sigma(\mathtt{x});\ \mathtt{end}]\!]$ to preserve finitely generated lower sets of ellipsoids. We adopt a similar approach as for the Minkowski sum and approximate the intersection of an

ellipsoid with several half-spaces with its Löwner ellipsoid. Hence, the abstract operator corresponding to a single branch is

$$[\![\texttt{if x} \in \bigcap_l \mathcal{H}_{\sigma,l} \texttt{ then } \mathrm{op}_\sigma(\texttt{x});\ \texttt{end}]\!]_v^\#(\underline{\mathcal{E}}) \coloneqq \mathrm{op}_\sigma^\# \circ \mathrm{guard}_\sigma\left(\mathcal{E}_v\right)$$

where $\mathrm{op}_\sigma^\#$ is the abstract counterpart of op_σ (an affine assignment in our applications) and guard_σ is a shorthand for the ellipsoidal mapping $\mathcal{E} \mapsto \mathcal{E} \sqcap \left(\bigcap_l \mathcal{H}_{\sigma,l}\right)$. The latter map over-approximates the subset of $\mathcal{E}$ contained in all half-spaces $\mathcal{H}_{\sigma,l}$ by its Löwner ellipsoid.

Merging the different branches is done in the lower set framework by taking the reunion over all branches. Doing this operations in the present framework would multiply the size of the generated set by the number of branches, leading to an exponential increase along the computation. We control this growth by means of the automaton and over-approximate by a single ellipsoid $\mathcal{E}_w$ all ellipsoids $[\![\texttt{if} \ldots \texttt{then}\, \mathrm{op}_\sigma(\texttt{x})\,;\ \texttt{end}]\!]^\#(\mathcal{E}_v)$ such that $w = \tau(v,\sigma)$. We choose $\mathcal{E}_w$ to be the Löwner ellipsoid of the former matrices. A good choice of automaton will merge ellipsoids that are already "close" and introduce a little approximation gap. This yields an abstract operator of the form:

$$[\![\texttt{switch } \sigma\ ;\ \texttt{case}(\texttt{x} \in \bigcap_l \mathcal{H}_{\sigma,l})\colon\ \mathrm{op}_\sigma(\texttt{x});]\!]_w^\#(\underline{\mathcal{E}}) = \bigsqcup_{\substack{\sigma\in\Sigma, v\in\mathcal{W}\\ \tau(v,\sigma)=w}} \mathrm{op}_\sigma^\# \circ \mathrm{guard}_\sigma\left(\mathcal{E}_v\right).$$

When the condition is evaluated by a random integer value, the guard operator vanishes and the abstract operator is obtained as a special case:

$$[\![\texttt{switch } b\ ;\ \texttt{case}(b = \sigma)\ :\ \ \mathrm{op}_\sigma(\texttt{x});]\!]_w^\#(\underline{\mathcal{E}}) = \bigsqcup_{\substack{\sigma\in\Sigma, v\in\mathcal{W}\\ \tau(v,\sigma)=w}} \mathrm{op}_\sigma^\#\left(\mathcal{E}_v\right).$$

By definition of the $\sqcup$ operator on ellipsoids, the operators defined above are sound:

Lemma 6.4. *The abstract operator* $[\![\texttt{switch } \sigma\ ;\ \texttt{case}(\texttt{x} \in \bigcap_l \mathcal{H}_{\sigma,l})\colon\ \mathrm{op}_\sigma(\texttt{x});]\!]^\#$ *is sound:*

$$[\![\texttt{switch } \sigma\ ;\ \texttt{case}(\texttt{x} \in \bigcap_l \mathcal{H}_{\sigma,l})\colon\ \mathrm{op}_\sigma(\texttt{x});]\!]\left[\gamma(\underline{\mathcal{E}})\right] \subseteq \gamma\left[[\![\texttt{switch } \sigma\ ;\ \texttt{case}(\texttt{x} \in \bigcap_l \mathcal{H}_{\sigma,l})\colon\ \mathrm{op}_\sigma(\texttt{x});]\!]^\#(\underline{\mathcal{E}})\right].$$

6.2.4 Body of loops

We combine the abstract operators for affine assignments and affine guards to obtain the abstract counterpart of the map s in Equation (5.11) denoted by $s^\#$. It maps $\mathfrak{E}_n^{\mathcal{W}}$ into itself and is defined coordinate-wise by

$$s_w^\#\colon \underline{\mathcal{E}} \mapsto \bigsqcup_{\substack{\sigma\in\Sigma, v\in\mathcal{W}\\ \tau(v,\sigma)=w}} \left(x \mapsto A_\sigma x + c_\sigma\right) \circ \mathrm{guard}_\sigma\left(\mathcal{E}_v\right) \boxplus \left(u \mapsto B_\sigma u\right)\left(\mathcal{E}_{\mathcal{U}}\right). \tag{6.1}$$

It is a sound abstraction of the loop-body operator s.

Lemma 6.5. *The abstract operator* $s^\#$ *is sound: for all* $\underline{\mathcal{E}} \in \mathfrak{E}_n^{\mathcal{W}}$*, we have*

$$s\left[\gamma(\underline{\mathcal{E}})\right] \subseteq \gamma\left[s^\#(\underline{\mathcal{E}})\right].$$

Proof. The map s is set-valued, hence $s\big[\gamma(\underline{\mathcal{E}})\big] = \cup_{w\in\mathcal{W}} s(\mathcal{E}_w)$. Following Lemmas 6.3 and 6.4, the abstract operators used in $s^\#$ overapproximate their concrete counterpart, hence $s(\mathcal{E}_w) \subseteq s^\#_w(\underline{\mathcal{E}})$ for all $w \in \mathcal{W}$. We deduce that $s\big[\gamma(\underline{\mathcal{E}})\big] \subseteq \cup_{w\in\mathcal{W}} s^\#_w(\underline{\mathcal{E}}) = \gamma\big[s^\#(\underline{\mathcal{E}})\big]$. □

We deduce the expression of the loop-body abstraction for non-deterministic switched affine programs

$$s^\#_w \colon \underline{\mathcal{E}} \mapsto \bigsqcup_{\substack{\sigma\in\Sigma, v\in\mathcal{W}\\ \tau(v,\sigma)=w}} \big(x \mapsto A_\sigma x + c_\sigma\big)\big(\mathcal{E}_v\big) \boxplus \big(u \mapsto B_\sigma u\big)\big(\mathcal{E}_{\mathcal{U}}\big)\,.$$

For non-deterministic switched linear programs, the map $s^\#$ acts on *centered ellipsoids*. Writing $\mathcal{E}_w = \mathcal{E}(Q_w)$ and $\underline{Q} = (Q_w)_{w\in\mathcal{W}}$, this operator can be rewritten as a map $s^\#$ acting on $(\mathcal{S}^+_n)^{\mathcal{W}}$ defined by

$$s^\#_w \colon \underline{Q} \mapsto \bigsqcup_{\substack{\sigma\in\Sigma, v\in\mathcal{W}\\ \tau(v,\sigma)=w}} A_\sigma Q_v A_\sigma^T \tag{6.2}$$

6.2.5 Loop invariants

We recall that a (loop) invariant for Program 3 is a set $\mathcal{X}$ that contains the set of initial points $\mathcal{I}$ such that $A_\sigma x + B_\sigma u + c_\sigma \in \mathcal{X}$ for all $x \in \mathcal{X} \cap (\bigcap_l \mathcal{H}_{\sigma,l})$, $u \in \mathcal{U}$ and $\sigma \in \Sigma$. In other words, an invariant is a set $\mathcal{X}$ such that $\mathcal{I} \subseteq \mathcal{X}$ and $s(\mathcal{X}) \subseteq \mathcal{X}$. In the spirit of abstract interpretation, we solve the problem of computing an invariant in the abstract domain. First, we abstract the set of initial states $\mathcal{I}$ by a finitely generated lower set, which yields the vector of ellipsoid $\underline{\mathcal{E}}_{\mathcal{I}} \coloneqq (\underline{\mathcal{E}}^{\mathcal{I}}_w)_w$ whose concretization covers $\mathcal{I}$: $\mathcal{I} \subseteq \cup_w \mathcal{E}^{\mathcal{I}}_w$. Second, we define the map T by

$$T_w(\underline{\mathcal{E}}) \coloneqq \underline{\mathcal{E}}^{\mathcal{I}}_w \sqcup s^\#_w(\underline{\mathcal{E}})$$

and solve for $\underline{\mathcal{E}}$ the post-fixed point equation

$$T(\underline{\mathcal{E}}) \sqsubseteq \underline{\mathcal{E}}\,. \tag{6.3}$$

A loop invariant is then obtained as the concretization of any solution of Equation (6.3) in the abstract domain:

Theorem 6.6. *Let $\underline{\mathcal{E}}$ denote a post-fixed point of the map T. Then the set $\gamma(\underline{\mathcal{E}}) = \cup_{w\in\mathcal{W}} \mathcal{E}_w$ is an invariant for Program 3.*

Proof. If $\underline{\mathcal{E}} \in \mathfrak{E}^{\mathcal{W}}_n$ satisfies $T(\underline{\mathcal{E}}) \sqsubseteq \underline{\mathcal{E}}$, then, by monotonicity of the concretization, we obtain $\gamma\big[T(\underline{\mathcal{E}})\big] \subseteq \gamma(\underline{\mathcal{E}})$. By Lemma 6.5 we also have $s\big[\gamma(\underline{\mathcal{E}})\big] \subseteq \gamma\big[s^\#(\underline{\mathcal{E}})\big]$, hence $s\big[\gamma(\underline{\mathcal{E}})\big] \subseteq \gamma(\underline{\mathcal{E}})$. □

6.2.6 A robust analysis

We have defined our abstract operators by overapproximating in each instance the concrete element by its Löwner ellipsoid. The remarkable geometric properties of the Löwner ellipsoid are a major advantage when applied to program verification. Not only does the Löwner ellipsoid preserve any symmetries that are present, it also allows our analysis to be invariant under affine rewriting of the program.

Program 7: Program X	**Program 8:** Program Y

switch *rand_bool* **:**
  case *true* **:**
  $\mid$ $x \coloneqq Ax + a$
  end
  case $false$ **:**
  $\mid$ $x \coloneqq Bx + b$
  end
	switch *rand_bool* **:**
	  case *true* **:**
	  $\mid$ $y \coloneqq UAU^{-1}(y - v) + Ua + v$
	  end
	  case $false$ **:**
	  $\mid$ $y \coloneqq UBU^{-1}(y - v) + Ub + v$
	  end

Figure 6.1: Affine change of variables in a switch statement

Let us illustrate the latter property on the programs given in Figure 6.1. The right-hand side program is obtained from the left-hand side one by applying the affine change of variables $\mathtt{y} \coloneqq U\mathtt{x} + v$. We denote by $\mathcal{E}_X^0$ the initial abstract state in the analysis of the left-hand side program, i.e., before the execution of the switch statement. Accordingly, we assume that the abstract state of the right-hand side program is given by $\mathcal{E}_Y^0 = (x \mapsto Ux + v) \cdot \mathcal{E}_X^0$. Following the definition of the abstract primitives, the analysis of the two programs respectively provides the following final invariants:

$$\mathcal{E}_X^f = (x \mapsto Ax + a) \cdot \mathcal{E}_X^0 \sqcup (x \mapsto Bx + b) \cdot \mathcal{E}_X^0$$
$$\mathcal{E}_Y^f = (y \mapsto A'y + a') \cdot \mathcal{E}_Y^0 \sqcup (y \mapsto B'y + b') \cdot \mathcal{E}_Y^0$$

where $X' = UXU^{-1}$ and $x' = Ux + v - UXU^{-1}v$ for $(X, x) \in \{(A, a), (B, b)\}$. Then it can be verified using Proposition 3.9 that the final invariant of the second program corresponds to a rewriting of the invariant of the first program, i.e.,

$$\mathcal{E}_Y^f = (x \mapsto Ux + v) \cdot \mathcal{E}_X^f \,.$$

6.3 Non-monotone Kleene algorithm for switched affine systems

6.3.1 The non-monotone Kleene iteration

In this section, we present to compute an invariant for Program 3 as the union of finitely many ellipsoids. It is obtained as the fixed point of a non-monotone map that is based on an automaton, whose states represent finite execution traces (in the case of De Bruijn automata, see Section 6.3.5, these distinguish between different suffixes of traces of the same length). As a consequence, we expect that the more states this automaton has (i.e., the more execution traces are taken into account during the computation), the more accurate the invariant to be. Moreover, since an ellipsoid is associated with each state of the automaton, the number of disjunctions in the invariant is constant during the computation.

We introduce an action of Σ^* on the set of ellipsoids $\mathfrak{E}_n$, denoted by $\cdot$ and defined for $\sigma \in \Sigma$ and $\mathcal{E} \in \mathfrak{E}_n$ by:

$$\sigma \cdot \mathcal{E} \coloneqq \big(f_\sigma \circ \mathrm{guard}_\sigma(\mathcal{E})\big) \boxplus \big(B_\sigma \mathcal{E}_{\mathcal{U}}\big) \,,$$

where f_σ denotes the affine map $x \mapsto A_\sigma x + c_\sigma$. In this way, the map $\mathcal{E} \mapsto \sigma \cdot \mathcal{E}$ represents the abstract operator associated with the branch σ of the program.

The abstract operator for the whole loop is then written coordinate-wise in condensed form as

$$T_w(\underline{\mathcal{E}}) \coloneqq \mathcal{E}_w^{\mathcal{I}} \sqcup \bigsqcup_{\tau(v,\sigma)=w} \sigma \cdot \mathcal{E}_v \,. \tag{6.4}$$

The fact that semi-definite programs can only be solved up to a prescribed accuracy is a well known source of difficulties in numerical program verification. If the approximate invariant which is found is mapped to its interior, meaning that some strictly feasible solution is returned by the solver, then, an exact invariant can be obtained a posteriori by some rounding procedure, see the discussion in [RJGF12]. In order to make such methods applicable in the present setting, we introduce a small margin $\varepsilon > 0$ which will absorb numerical imprecisions as suggested by the authors in [RVS16]. Hence, we define the perturbed map T^ε from $\mathfrak{E}_n^{\mathcal{W}}$ to itself, whose w-th coordinate is obtained by adding a "padding" εI_n to each ellipsoid:

$$T_w^\varepsilon(\underline{\mathcal{E}}) \coloneqq \mathcal{E}\big(Q_w + \varepsilon I_n, q_w\big) \quad \text{where} \quad \mathcal{E}(Q_w, q_w) = T_w(\underline{\mathcal{E}}) \,.$$

Remark 6.1. The ellipsoids $\mathcal{E}(Q_w, q_w)$ and $\mathcal{E}(Q_w + \varepsilon I_n, q_w)$ have the same center and $Q_w \preccurlyeq Q_w + \varepsilon I_n$, hence by monotony of the translation $x \mapsto x - q_w$, we may assume[3] that they are centered and the inequality $Q_w \preccurlyeq Q_w + \varepsilon I_n$ implies that $T_w^\varepsilon(\underline{\mathcal{E}}) \supseteq T_w(\underline{\mathcal{E}})$.

This is a sound over-approximation of the abstract loop operator: $s_w^\#(\underline{\mathcal{E}}) \subseteq T_w^\varepsilon(\underline{\mathcal{E}})$. Introducing the parameter ε induces a trade-off between speed (ε large) and precision (ε small). The speed-up effect is shown in Equation (6.6), resulting from the complexity analysis the next section. The loss of precision is due to the fact that the map T^ε is a "deformation" of the true map T. In the experiments, we have chosen $0.01 \leqslant \varepsilon \leqslant 0.2$.

These operators enable the computation of invariants as unions of ellipsoids:

Theorem 6.7. *Let $\underline{\mathcal{E}} = (\mathcal{E}_w)_{w\in\mathcal{W}}$ denote a fixed point of the map T^ε. Then the set $\cup_{w\in\mathcal{W}}\mathcal{E}_w$ is an invariant for Program 3.*

Proof. Let $\underline{\mathcal{E}}$ denote such a fixed point, i.e., $T_w^\varepsilon(\underline{\mathcal{E}}) = \mathcal{E}_w$ for all $w \in \mathcal{W}$. By Remark 6.1, we have $T_w^\varepsilon(\underline{\mathcal{E}}) \supset T_w(\underline{\mathcal{E}})$ for all w, hence $T(\underline{\mathcal{E}}) \sqsubseteq \underline{\mathcal{E}}$, which show that $\gamma(\underline{\mathcal{E}})$ is an invariant by Theorem 6.6 □

The same is true if $\underline{\mathcal{E}}$ is only a post-fixed point of the map T^ε, i.e., if for all $w \in \mathcal{W}$, we have $T_w^\varepsilon(\underline{\mathcal{E}}) \subseteq \mathcal{E}_w$.

The map T^ε is the analogue of the fixed point functional in abstract interpretation. Classical abstract interpretation requires the fixed point functional to be a monotone map defined on a complete lattice [CC77b]. Then, a program invariant can be obtained as the least fixed point of this functional, which can be computed by a standard fixed point scheme, *Kleene iteration*. The present setting is more complex, for the space $\mathfrak{E}_n^{\mathcal{W}}$ is not a lattice, and the operators $\sqcup$, $\sqcap$ and $\boxplus$ are not monotone (this can be quickly verified, even for ellipsoids of dimension 2). This entails that the map T^ε is not monotone. However, we can still formulate an iteration scheme *a la* Kleene in the present setting, defining

$$\begin{aligned} \underline{\mathcal{E}}^0 &= (\mathcal{E}_1^{\mathcal{I}}, \ldots, \mathcal{E}_N^{\mathcal{I}}) \\ \underline{\mathcal{E}}^{k+1} &= T^\varepsilon(\underline{\mathcal{E}}^k) \,. \end{aligned} \tag{6.5}$$

[3] Alternatively, one can check that choosing $\lambda = 1$ validates the LMI in Equation (5.3).

6.3.2 On the convergence of the scheme

We assume in the sequel that the set of initial states $\mathcal{I}$ and the set of controls $\mathcal{U}$ are full-dimensional. If this is not the case, we may approximate these ellipsoids by unions of "nearly flat" ellipsoids. Then, it is easily shown by induction that all occurrences of the operators $\sqcup$, $\sqcap$ and $\boxplus$ in the Kleene iteration can be computed with the means presented in Section 5.2:

Lemma 6.8. *Assume that the ellipsoids $\mathcal{E}_I$ and $\mathcal{E}_\mathcal{U}$ are full-dimensional. Then all the ellipsoids $\mathcal{E}_w^k$ computed at each step of the Kleene iteration in Equation* (6.5) *are full-dimensional.*

The lack of monotonicity of the operator T^ε, and the fact that the space of ellipsoids does not constitute a lattice, make more difficult the analysis of the Kleene iteration scheme than in the classical case of abstract analysis. In particular, we have to replace some order theoretical arguments by metric fixed point properties. We establish the convergence of the Kleene iteration in the *linear case* — i.e., when the assignments are linear ($B_\sigma = 0$ and $c_\sigma = 0$), the switching process is non-deterministic and the ellipsoids are centered — if the "stability margin" is sufficient.

Theorem 6.9. *Assume that the assignments are linear ($B_\sigma = 0$ and $c_\sigma = 0$), the switching process is non-deterministic and that the ellipsoids $\mathcal{E}_w^\mathcal{I}$ are centered and full-dimensional. Then there is a positive constant $\mu_{n,I}$ depending on the dimension n and the initial states $\mathcal{E}_I$ such that if the spectral norms of the matrices $(A_\sigma)_{\sigma\in\Sigma}$ are smaller than $\mu_{n,I}$, then the Kleene iteration $\underline{\mathcal{E}}^{k+1} = T^\varepsilon(\underline{\mathcal{E}}^k)$ converges.*

Proof. Since all ellipsoids are centered, we use the notation from Chapter 3 and write $Q_1 \sqcup Q_2$ instead of $\mathcal{E}(Q_1, 0) \sqcup \mathcal{E}(Q_2, 0)$, and abuse the notation to write $T_w^\varepsilon(\underline{Q})$ instead of $T_w^\varepsilon(\underline{\mathcal{E}})$. We denote $\mathcal{E}_w^\mathcal{I} = \mathcal{E}(Q_w^\mathcal{I}, 0)$. The map T^ε is written

$$T_w^\varepsilon(\underline{Q}) = Q_w^\mathcal{I} \sqcup \bigsqcup_{\tau(v,\sigma)=w} (A_\sigma Q_v A_\sigma^T + \varepsilon I_n)\,.$$

First, we show that there are positive reals $\lambda_0 < \lambda_1$ such that the set of X such that for all $\underline{Q} \in (\mathcal{S}_n^{++})^\mathcal{W}$,

$$\Big(\forall w.\ \lambda_0 I_n \preccurlyeq Q_w \preccurlyeq \lambda_1 I_n\Big) \implies \Big(\forall w.\ \lambda_0 I_n \preccurlyeq T_w^\varepsilon(\underline{Q}) \preccurlyeq \lambda_1 I_n\Big)\,.$$

The map T_w^ε is bounded below by $Q_w^\mathcal{I}$, which is positive definite, so there is $\lambda_0 > 0$ such that $T_w^\varepsilon(\underline{Q}) \succcurlyeq \lambda_0 I_n$ for all X. We deduce from Theorem 3.17 that if the spectral norms of the matrices A_σ are strictly less than $N^{-1/2}$, then the desired property is satisfied for $\lambda_1 \coloneqq \varepsilon + \|Q_I\|_2(1 - \sum_\sigma \|A_\sigma\|)^{-1}$.

The set $\mathcal{K} \coloneqq \{X \in \mathcal{S}_n^{++} \colon \lambda_0 I_n \preccurlyeq X \preccurlyeq \lambda_1 I_n\}$ is bounded in Riemann's metric (its diameter is less than $n \log(\lambda_1 \lambda_0^{-1})$). Moreover, one can obtain the very coarse bound

$$(\lambda_1/\lambda_0^2)^{-1}\|X - Y\|_2 \leqslant d_R(X, Y) \leqslant (\lambda_1^2/\lambda_0)\|X - Y\|_2\,.$$

Hence the $\sqcup$ operator must be Lipschitz with respect to the euclidean norm on the set $\mathcal{K}$, with a Lipschitz constant no larger than $(\lambda_1/\lambda_0)^3$. Moreover, we recall that the Riemann is invariant by a congruence by an invertible matrix P: $d_R(PXP^T, PYP^T) = d_R(X, Y)$. Combining these results, we deduce, for $\underline{Q}, \underline{Q'} \in (\mathcal{S}_n^{++})^\mathcal{W}$ and $\alpha \coloneqq \max_\sigma \|A_\sigma\|_2^2$,

$$\|T_w^\varepsilon(\underline{Q}) - T_w^\varepsilon(\underline{Q'})\|_2 \leqslant \alpha\Big(\frac{\lambda_1}{\lambda_0}\Big)^{3|\mathcal{W}|}\sqrt{p} \max_{\tau(v,\sigma)=w} \|Q_v - Q'_v\|_2\,.$$

Hence, the Kleene iteration converges if the spectral norms of the matrices A_σ are small enough.

It is sufficient that $\|T^\varepsilon_w(\underline{Q}) - T^\varepsilon_w(\underline{Q'})\|_2 \leqslant \varepsilon$ to obtain an invariant, thus we deduce the number of iterations given in Equation (6.6). □

The bound $\mu_{n,I}$ that is given is very conservative. However, we shall see in Section 6.3.5 that our algorithm converges although the condition is not satisfied.

We also deduce that

$$O\Big(\frac{\log\varepsilon - \log\max_w \|Q^0_w - Q^\infty_w\|_2}{3|\mathcal{W}|\log(\lambda_1/\lambda_0) + \sqrt{p} + 2\log\max_\sigma \|A_\sigma\|_2}\Big) \tag{6.6}$$

iterations suffice to compute an invariant. As shown in Section 6.3.2, the values $\lambda_0 < \lambda_1$ depend only on the set of initial states and the matrices A_σ.

Establishing the convergence of the Kleene iteration in Equation 6.5 in the general case is difficult problem and remains open.

Open problem 6.10. *Does the iterative scheme in Equation 6.5 converge if the matrices A_σ are suitably small, when either affine assignments ($c_\sigma \neq 0$), guards (switching is state-dependent) or non-constant controls are present ($B_\sigma \neq 0$) ?*

6.3.3 Two implementations for affine and linear programs

The "small-LMI" approach When implementing the Kleene iteration scheme in Equation (6.5), a semi-definite program needs to be solved for each evaluation of the operator $\sqcup$ and $\boxplus$. The number of variables in each of these semi-definite programs only depends on the dimension n of the problem, not on the size $N = |\mathcal{W}|$ of the automaton, in contrast with alternative approaches detailed in Section 6.3.4, inspired by state of the art methods. For this reason, we call the method to compute invariants based on Theorem 6.7 and on Kleene iteration the "small-LMI" approach.

Note that Theorem 6.7 is also valid for the map T. As a consequence, we may be tempted to use this map rather than T^ε in the Kleene scheme. However, in practice, most operators are evaluated by solving semi-definite programs, which only return approximate optimal solutions. Introducing the parameter ε counters several hurdles that may be encountered and could endanger confidence in the final invariant. First, the parameter ε absorbs approximation errors that appear throughout the computation, and thus gives a margin of safety if computations are done using finite precision. Moreover, padding each inclusion constraint ensures that the set of feasible points has non-empty interior, so that the a posteriori numerical check presented in [RVS16] can be used. Finally, if the parameter ε was not present, a fixed point would only be reached *ultimately*, i.e., after an infinite number of iterations. Now, a post-fixed point can be reached in a finite number of iterations.

The "no-LMI" approach When the assignments in each branch are linear ($B_\sigma = 0$ and $c_\sigma = 0$) and when the switching condition is non-deterministic (meaning that the test is replaced by a random boolean), it is possible to get rid of LMIs altogether, still building on the same principles. Indeed, it was shown in [AGS+16] that the Löwner ellipsoid of the union of *two* centered ellipsoids can be computed from a Cholesky decomposition, avoiding the use of LMI, resulting in an important speed-up. We will exploit here the latter result, by relaxing

the computation of $\sqcup_k \mathcal{E}_k$ to a sequential computation $\big((\mathcal{E}_1 \sqcup \mathcal{E}_2) \cdots \sqcup \mathcal{E}_w\big)$. The variant of the map T^ε obtained in this way can now be computed without solving semi-definite programs. In the sequel, we will refer to this variant of the present "small-LMI" approach as the "no-LMI" approach. We explore the "no-LMI" in further detail in Section 6.4.

6.3.4 Alternative approaches: the "big-LMI" and "big-BMI" methods

For comparison, we next present two alternative approaches, derived from earlier works [AJPR14, RJGF12], leading to larger LMI or to non-convex programs.

The first approach has been studied in [AJPR14]. It is restricted to the special case of *linear assignments* under *non-deterministic switching*, where the initial state has been approximated by a *centered ellipsoid*. In other words, it requires that $B_1 = B_2 = 0$, $c_1 = c_2 = 0$, the guard condition $f^T x \leqslant g$ has been replaced by a non-deterministic switching mechanism and $\mathcal{E}_I = \mathcal{E}(Q_I, 0)$. Then, the post-fixed point problem can be rewritten as a single LMI involving the whole collection of design variables $(Q_w)_{w\in\mathcal{W}}$:

$$\begin{aligned} Q_I &\preccurlyeq Q_w, \ \forall w \in \mathcal{W}\,, \\ A_\sigma Q_v A_\sigma^T &\preccurlyeq Q_w, \ \forall w, v, \sigma \text{ such that } \tau(v,\sigma) = w\,. \end{aligned} \tag{6.7}$$

The variables in this LMI are highly coupled among themselves. Indeed, the variable Q_w appears N times on the right-hand-side of an inequality of the form above, but also N times on the left-hand-side. Thus, unless the transition map τ is constant ($\tau(v,\sigma) = v$ for all σ), it is not possible to solve Equation (6.7) for each word w separately. This case does not appear in practice: the transition map used in Section 6.3.5 induces a maximal coupling, where (almost) each word is related to p other words.

If $(Q_w)_{w\in\mathcal{W}}$ is a solution of this LMI, then the union of the ellipsoids $\cup_{w\in\mathcal{W}}\mathcal{E}(Q_w, 0)$ is an invariant for the associated program. It is in fact a variation on the method in [AJPR14], because the LMI $Q_w \succ 0$ has been replaced by $Q_w \succcurlyeq Q_I$. This only requires that the solution provided in [AJPR14] be scaled to contain the set of initial states.

The latter optimization problem has $|\mathcal{W}|n(n+1)/2$ independent variables. We say that this optimization problem is a "big LMI" because the number of variables depends on the size of the automaton. Finding an approximate solution via semi-definite programming thus has an arithmetic complexity of $\mathcal{O}(n^{6.5}|\mathcal{W}|^5)$, according to the formulas in Section 4.6.3 of [BTN01]. In comparison, the semi-definite programs used to compute the operators $\sqcup$ and $\boxplus$ both have an arithmetic complexity of $\mathcal{O}(n^{6.5})$ that does not depend on $|\mathcal{W}|$. Each iteration of the Kleene iteration only needs to compute $\mathcal{O}(|\mathcal{W}|)$ of these, for a total arithmetic complexity of $\mathcal{O}(n^{6.5}|\mathcal{W}|)$ per iteration. As a consequence, when the automaton used in the computation of the map T^ε has many states, the "big-LMI" approach becomes intractable, contrary to the "small-LMI" method.

The "big-LMI" method may be thought of as dual of a Lyapunov-type approach, also detailed in [AJPR14]: the latter is equivalent to representing the unit ball of the Barabanov norm by an intersection of ellipsoid. Instead, the "big-LMI" method computes an invariant set given by a union of ellipsoids. Both methods lead to semi-definite programs of a comparable nature and size.

In the case of a single centered ellipsoid, it is still possible to use LMIs if we assume that B_1 or B_2 is non-zero (but not both). We can use a method akin to the one used in [RJGF12], where bisection is used in order to successively compute a value for λ in Equation (5.7).

However, the more general case of affine assignments under non-deterministic switching cannot be dealt with LMIs, since the stability problem then involves *several bilinear inequalities* in terms of the design variables. This can dealt with by solving a *bilinear matrix inequality* (BMI for short), which has the form

$$A_0 + \sum_{i=1}^{d} x_i A_i + \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j A_{i,j} \succcurlyeq 0\,. \tag{6.8}$$

For instance, when $B_1 = B_2 = 0$ and the switching is non-deterministic, the post-fixed point problem can be rewritten as a BMI in the variables $(L_w)_{w\in\mathcal{W}}$, $(\lambda_w)_{w\in\mathcal{W}}$ and $(\mu_{v,\sigma})_{(v,\sigma)\in W\times\Sigma}$:

$$\begin{array}{l}
\forall w,\, \exists \lambda_w \in \mathbb{R}\colon \begin{pmatrix} L_w L_w^T & q_0 - q_w & L_0 \\ (q_0 - q_w)^T & 1-\lambda_w & 0_{1,n} \\ L_0^T & 0_{n,1} & \lambda_w I_n \end{pmatrix} \succcurlyeq 0\,, \\
\forall v,\sigma,\, \exists \mu_{v,\sigma} \in \mathbb{R}\colon \\
\begin{pmatrix} L_{\tau(v,\sigma)} L_{\tau(v,\sigma)}^T & A_\sigma q_v + c_\sigma - q_{\tau(v,\sigma)} & L_v \\ (A_\sigma q_v + c_\sigma - q_{\tau(v,\sigma)})^T & 1-\mu_{v,\sigma} & 0_{1,n} \\ L_v^T & 0_{n,1} & \mu_{v,\sigma} I_n \end{pmatrix} \succcurlyeq 0\,,
\end{array}$$

where $\mathcal{E}_I = \mathcal{E}(L_0 L_0^T, q_0)$. Any solution of this BMI yields as invariant the union of the ellipsoids $\mathcal{E}(L_w L_w^T, q_w)$. Unlike LMI, BMI have generally non-convex feasible sets, and therefore numerical solvers may return only locally optimal solutions. Despite the computational drawbacks of the "big-BMI" method, it is to our knowledge the only state of the art method that can deal with affine assignments with different equilibria.

6.3.5 Benchmarks

We present in this section numerical benchmarks of our method. The experiments are implemented in Matlab, running on one core of an 2.2GHz Intel Core i7 with 8Gb RAM. We use the SDPT-3 solver [TTT03], in conjunction with YALMIP [Löf04] to solve LMIs, and the PENLAB solver [JF13] to solve BMIs. In all subsequent pictures, the initial state is shown in magenta and the disjunctive invariant $\mathcal{I} := \cup_{w\in\mathcal{W}} \mathcal{E}_w$ is shown in red. We show in blue (resp. green) the image of the invariant $\mathcal{I}$ by the abstract operators of the branch 1 (resp. 2), i.e., $\cup_{w\in\mathcal{W}} 1\cdot\mathcal{E}_w$ (resp. $\cup_{w\in\mathcal{W}} 2\cdot\mathcal{E}_w$), which prove an over-approximation of the reachable set in branch 1 (resp. 2). In all examples, it was sufficient to compute 30 iterations to obtain a post-fixed point, and thus an invariant.

Switched linear system with guards We next show that automata that "keep in memory" the m last switches that happened produce better invariants than other types of automata. Moreover, we demonstrate that the invariants that are produced are more accurate the more switches are "remembered". We instantiate the elements from Program 9 as follows: $\mathcal{E}_I = \mathcal{E}(0.04 I_2, \left(\begin{smallmatrix}1\\0.5\end{smallmatrix}\right))$, $\mathcal{U} = \mathcal{E}(0.1 I_2, 0)$, $A_1 = 0.5\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$, $B_1 = I_2$, $c_1 = 0$, $A_2 = 0.5\left(\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right)$, $B_2 = I_2$, $c_2 = 0$, $f = \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$, $g = 1$. Note that it involves an affine guard, so the analysis of this example is out of reach of a "big-LMI", or "big-BMI",-type method.

We recall that the *De Bruijn automaton* on the set Σ^m is an automaton whose alphabet is Σ and whose states are precisely Σ^m. Its transition map τ deletes a word's first letter and appends the transition letter to its end. In other words, we have $\tau(v,\sigma) = w$ if and

Program 9: A simple switched affine program

```
x ← E_I;
while true do
    u ← E_U;
    if f^T x ⩽ g then
        x := A_1 x + B_1 u + c_1;
    else
        x := A_2 x + B_2 u + c_2;
    end
end
```

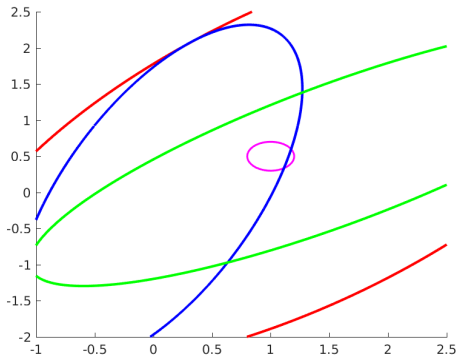
only if $v = \sigma_1\sigma_2 \dots \sigma_m$ and $w = \sigma_2 \dots \sigma_m\sigma$, with $\sigma_i \in \Sigma$. By construction, this automaton "remembers" the last m transitions. For this reason, we expect invariants computed using larger De Bruijn automata to be more precise. This has been verified experimentally and is shown in Figures 6.2a-6.2c. We have also experimented with non-De Bruijn automata, as shown in Figure 6.2d. Notice that as the more switches are "remembered", the more concise the invariant becomes throughout Figure 6.2. We also point out that the transition function for the automaton used in Figure 6.2d does not reflect a memory process. Although it performs slightly better that the De Bruijn automaton on Σ^0, using only one ellipsoid, the invariant computed with this automaton remains convex and thus less accurate than the previous ones.

Defocused switched affine systems We demonstrate again the fact that the "small-LMI" method provides better invariants the more states the underlying automaton has. We consider a discretized version of Example 6.3 in [NBSN13], to which we have added a guard condition. Using a discretization step $\delta t = 0.5$, we instantiate Program 9 with $\mathcal{E}_I = \mathcal{E}(0.04I_2, \left(\begin{smallmatrix}0.5\\0\end{smallmatrix}\right))$, $\mathcal{E}_\mathcal{U} = \mathcal{E}(0, 0)$, $A_1 = \left(\begin{smallmatrix}0.68 & -0.75\\0.19 & 0.68\end{smallmatrix}\right)$, $B_1 = 0$, $c_1 = \left(\begin{smallmatrix}0.5432\\-0.0724\end{smallmatrix}\right)$, $A_2 = \left(\begin{smallmatrix}0.72 & -0.39\\0.39 & 0.72\end{smallmatrix}\right)$, $B_2 = 0$, $c_2 = \left(\begin{smallmatrix}0\\0\end{smallmatrix}\right)$, $f = \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)$, $g = 0$. This system has two distinct fixed points. Since it involves an affine guard, the analysis of this example is out of reach of a "big-LMI", or "big-BMI",-type method. We show in Figure 6.3 two invariants computed by using the De Bruijn automata on Σ^2 (4 states) and Σ^4 (16 states). The invariant computed with the latter automaton is strictly better than the one computed with the former.

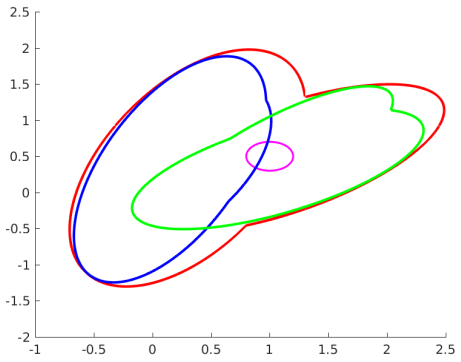
Observer based controller for a coupled mass system [RG13] We show that our method can also be used to analyze systems with saturations. The addition of saturation simulates sensors that measure a physical quantity precisely within some range, but cannot measure values outside this range. We study the stability of an affine dynamical system subject to saturation conditions on the first coordinate of the state vector:

$$x_1^{k+1/2} = \begin{cases} \beta & \text{if } f^T x^k > \beta \quad (6.9a)\\ -\beta & \text{if } f^T x^k < -\beta \quad (6.9b)\\ x_1^k & \text{otherwise} \quad (6.9c)\end{cases}$$

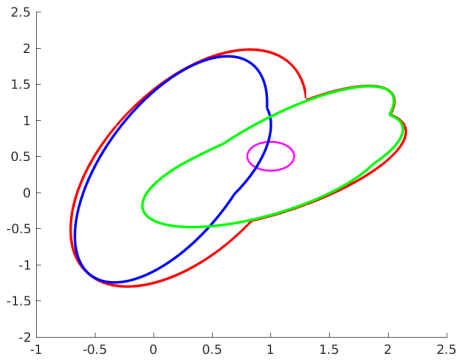
$$x^{k+1} = A_i x^{k+1/2} + B_i u^{k+1/2} + c_i \ , \ i \in \mathcal{I} \, ,$$



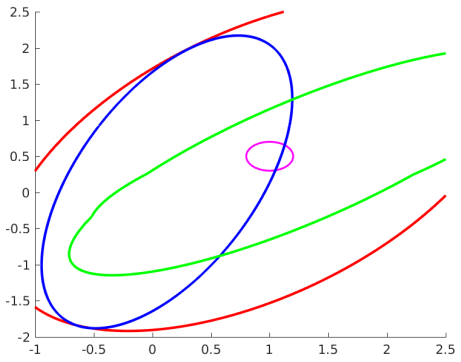
(a) De Bruijn automaton on Σ^0 (1 ellipsoid)



(b) De Bruijn automaton on Σ^1 (2 ellipsoids)



(c) De Bruijn automaton on Σ^2 (4 ellipsoids)



(d) Non-De Bruijn automaton with 3 states (3 ellipsoids)

Figure 6.2: Invariants (red) computed for a switched linear system w.r.t. the automaton, their image by the abstract operators (blue and green) and the initial state (magenta)

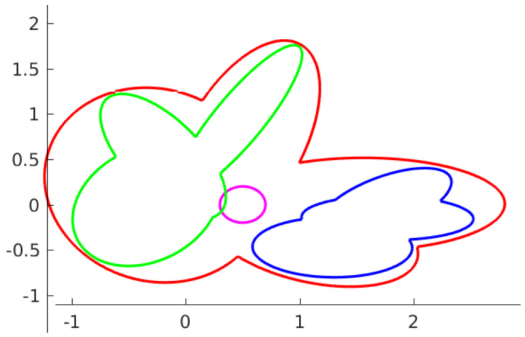
with a bounded control $u \in \mathcal{E}_{\mathcal{U}}$, first in a case without switching, and then in a case where switching occurs.

The semantics of a program implementing this system are the same of the program which resets the state vector to an initial value in the branches (6.9a) (also denoted by a) and (6.9b) (also denoted by b). In other words, we may choose $a \cdot \mathcal{E} := \mathcal{E}_I$ for all ellipsoid $\mathcal{E}$ (the same equation holds for b). Thus, the map T takes the usual form as in Equation (6.4) when the word w ends with c, and is written $T_w(\underline{\mathcal{E}}) = \mathcal{E}_I$ when w ends with 1 or 2.

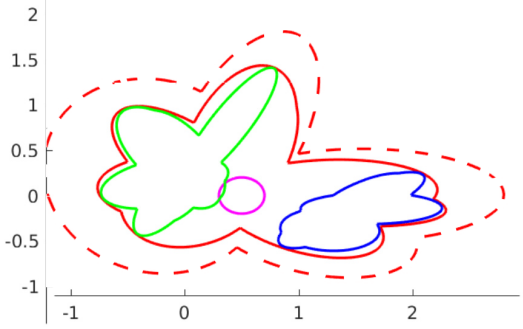
First, we demonstrate our method with $\mathcal{I} = \{1\}$, $c_1 = 0$, $f = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}^T$, $\beta = 0.5$,

$$A_1 = \begin{pmatrix} 0.6227 & 0.3871 & -0.113 & 0.0102 \\ -0.3407 & 0.9103 & -0.3388 & 0.0649 \\ 0.0918 & -0.0265 & -0.7319 & 0.2669 \\ 0.2643 & -0.1298 & -0.9903 & 0.3331 \end{pmatrix} \text{ and } B_1 = \begin{pmatrix} 0.3064 & 0.1826 \\ -0.0054 & 0.6731 \\ 0.0494 & 1.6138 \\ 0.0531 & 0.4012 \end{pmatrix}.$$

We use the De Bruijn graph on Σ^2 (4 ellipsoids). Our algorithm converges towards some collection of matrices $\underline{\mathcal{E}}$, and the resulting ellipsoids satisfy $\mathcal{E} \subset \mathcal{E}_{aa}$ for all $\mathcal{E} \in \underline{\mathcal{E}}$, meaning that the invariant is a single ellipsoid. Sections of this ellipsoid, as well as sections of the ellipsoid obtained in [RG13] are depicted in Figure 6.4.

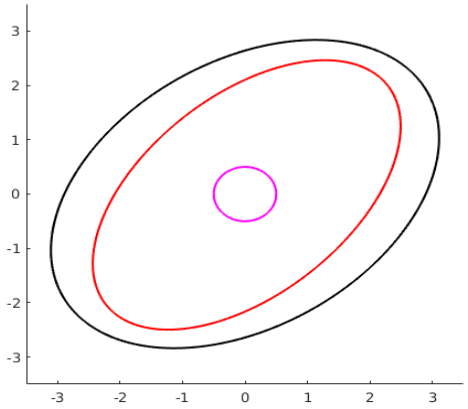


(a) De Bruijn automaton on Σ^2 (4 ellipsoids)

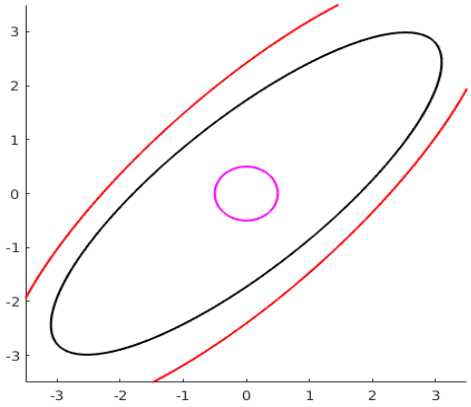


(b) De Bruijn automaton on Σ^4 (16 ellipsoids) – invariant with 4 ellipsoids in dashed

Figure 6.3: Invariants (red) computed for a defocused switched system w.r.t. the automaton, their image by the abstract operators (blue and green) and the initial state (magenta)



(a) Section in the (x_1, x_2) plane

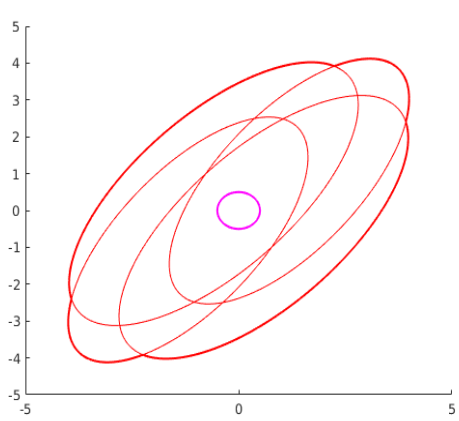


(b) Section in the (x_3, x_4) plane

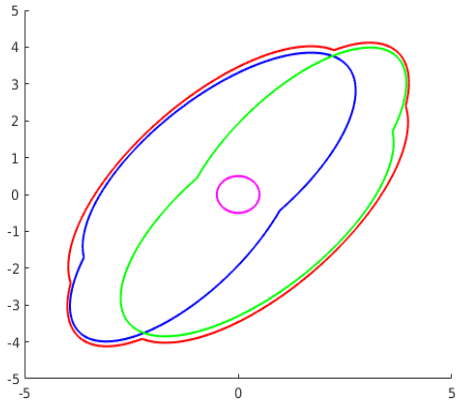
Figure 6.4: Sections of the invariant ellipsoid (red) for the coupled mass system and the reference in [RG13] (black)

We also demonstrate our algorithm on a variant of the system in [RG13], that combines a switching mechanism with a saturation constraint. More precisely, we shall use $\mathcal{I} = \{1, 2\}$, $A_1 = A_2$, $B_1 = B_2$, $c_1 = \begin{pmatrix} -1/2 & 0 & 0 & 0 \end{pmatrix}^T$ and $c_2 = \begin{pmatrix} 1/2 & 0 & 0 & 0 \end{pmatrix}^T$. The mode 1 is active if $x_1 > 0$ and the mode 2 is active otherwise. By design, this system has two fixed points (when $u = 0$). In this example, our method yields a non-convex invariant, whose section in the (x_1, x_2) plane is shown in Figure 6.5.

Comparison in the centered case: big-LMI versus no-LMI We compare the "big-LMI" and "no-LMI" methods numerically on systems switching between implementations of two damped harmonic oscillators $M_i\ddot{x} + C_i\dot{x} + K_i x = 0$, with a discretization time $\delta t = 0.1$. The matrices M_i, C_i, K_i are randomly generated positive definite matrices, for dimensions ranging from 2 to 30. In these examples, we have $B_1 = B_2 = 0$, $c_1 = c_2 = 0$ and the guard condition $f^T x \leqslant g$ has been replaced by a non-deterministic switching process. We use the De Bruijn



(a) The 4 ellipsoids representing the invariant



(b) Invariant (red) and its image by the abstract operators (green and blue)

Figure 6.5: Section in the (x_1, x_2) plane of the invariant for a switching variant of the coupled mass system (4 ellipsoids)

Dimension n	"big-LMI"	"no-LMI"	relative volume
5	0.4s	**0.2s**	1.24
10	0.8s	**0.3s**	1.12
15	5s	**0.4s**	1.15
20	22s	**0.7s**	-
25	2min	**1.0s**	-
30	6min	**1.4s**	-

Table 6.1: Invariant computation time of each method w.r.t. the dimension n of the matrices (8 ellipsoids)

automaton on Σ^3 in all computations, so the invariants that are computed are given as unions of 8 ellipsoids. The execution times for the "big-LMI" and "no-LMI" methods are shown in Table 6.1. We point out that the "no-LMI" method outperforms the "big-LMI" method by several orders of magnitude. The fact that the time-complexity relative to the dimension of the matrices is much smaller for the "no-LMI" method is also apparent.

Finally, we compare the relative accuracy of the "no-LMI" method with respect to the "big-LMI" approach by the *relative volume* of the computed invariants, defined by the n-th root of the ratio between the volume of the "no-LMI" invariant by the volume of the "big-LMI" invariant. We report a difference in relative volume no larger than 25% in Table 6.1, where the volumes have been estimated by a Monte-Carlo approximation, up to dimension 15 (no results for higher dimensions, due to lack of precision of the Monte-Carlo approach).

Comparison in the uncentered case: big-BMI versus small-LMI We compare the "big-BMI" and "small-LMI" methods numerically on systems switching between implementations of two damped harmonic oscillators with a non-deterministic control $M_i\ddot{x} + C_i\dot{x} + K_i x = u$, with a discretization time $\delta t = 0.1$ and the control u is bounded in a non-centered ellipsoid. The matrices M_i, C_i, K_i are randomly generated positive definite matrices, for dimensions ranging from 2 to 14. The switching process is non-deterministic. We have used the De Bruijn automaton on Σ^2 in all computations, so the invariants that are computed are given

Dimension n	"big-BMI"	"small-LMI"	relative volume
2	**3.5s**	24s	1.08
4	**9.2s**	26s	1.04
6	**34s**	36.3s	1.03
8	2.8min	**42.5s**	1.03
10	9.5min	**1.2min**	1.03
12	20.5min	**1.5min**	1.02
14	2.1h	**2.1min**	1.05
16	$>$ 3h	**3.8min**	1.04

Table 6.2: Invariant computation time of each method w.r.t. the dimension n of the matrices (4 ellipsoids)

as unions of 4 ellipsoids. The execution times for the "big-BMI" and "small-LMI" methods are shown in Table 6.2. Although the "big-BMI" is more efficient on lower dimensional examples, one can see that it is very time-costly for 14×14 matrices, as solving the BMI takes 2000 times longer than for 2×2 matrices. In contrast, the "small-LMI" method has a base time cost per iteration that only grows from 1s in dimension 2 to 4s in dimension 14. Finally, we compare the relative accuracy of the "small-LMI" method with respect to the "big-BMI" approach by the *relative volume* of the computed invariants. We report an difference in relative volume no larger than 8% in Table 6.2, where the volumes have again been estimated by a Monte-Carlo approximation.

6.4 A nonlinear power algorithm for linear systems

It is possible to split the computation of an invariant set into two parts in the linear case, i.e., when the operator $s^{\#}$ is written

$$s^{\#}_w : \underline{Q} \mapsto \bigsqcup_{\substack{\sigma \in \Sigma, v \in \mathcal{W} \\ \tau(v,\sigma)=w}} A_\sigma Q_v A_\sigma^T \tag{6.10}$$

We first look for a vector of matrices $\underline{Q}$ in the interior of $(\mathcal{S}_n^+)^{\mathcal{W}}$ (we say by extension that $\underline{Q}$ is positive definite) satisfying $s^{\#}(\underline{Q}) \sqsubseteq \underline{Q}$ (i.e., a post-fixed point of the abstract operator $s^{\#}$) using the power-like algorithm described in Section 6.4. The abstract element $\underline{Q}$ is then scaled in order to "contain" the abstract element $\underline{Q_{\mathcal{I}}}$ corresponding to the initial set $\overline{\mathcal{I}}$, i.e., we return the abstract element $\mu\underline{Q}$, where $\mu_T = \inf\{\mu \in \mathbb{R}_+ \mid \mu\underline{Q} \sqsupseteq \mathcal{Q}_{\mathcal{I}}\}$. This approach yields a sound invariant, because every abstract operator is positively homogeneous when dealing with the linear case.

As a consequence, the vector of matrices $\underline{Q}$ somehow serves as a template which is here computed in an automatic way. A similar scaling technique appeared in [Rou13], in which the template is computed using semidefinite programming.

6.4.1 Additive and multiplicative power iterations

In this section, we present two scalable algorithms which will allow us to find an ellipsoid invariant. We shall consider an auxiliary nonlinear spectral problem, which consists in finding

a positive definite $\underline{Q} \in (\mathcal{S}_n^{++})^{\mathcal{W}}$ and a scalar $\lambda > 0$ such that

$$s^{\#}(\underline{Q}) = \lambda \underline{Q} \ . \tag{6.11}$$

If we find an element $\underline{Q}$ for which $\lambda \leqslant 1$, then the original problem $s^{\#}(\underline{Q}) \sqsubseteq \underline{Q}$ is solved. An interest of introducing the extra degree of freedom λ is to allow for finite precision computations. If $s^{\#}(\underline{Q}) = \lambda \underline{Q}$ holds for $\lambda < 1$ and $\underline{Q}$ positive definite, then, the relation $s^{\#}(\underline{Q}) \sqsubseteq \underline{Q}$ remains valid under a small perturbation of $\underline{Q}$.

A simple idea to solve (6.11) is to choose an order preserving linear form $\psi : (\mathcal{S}_n^{+})^{\mathcal{W}} \to \mathbb{R}_+$, and to define the following fixed point scheme

$$\underline{Q}^{k+1} = \frac{1}{\psi \circ s^{\#}(\underline{Q}^k)} s^{\#}(\underline{Q}^k) \tag{6.12}$$

initialized with a positive definite $\underline{Q}^0$. A convenient choice of ψ is the trace functional: $\psi(\underline{Q}) = \sum_w \mathrm{trace}(\underline{Q}_w)$. The latter has the property that it does not vanish on $(\mathcal{S}_n^{+})^{\mathcal{W}}$ except at the zero vector. So, a division by zero will not occur in (6.12), unless $s(\underline{Q}^k)$ vanishes at some iteration, which will not be the case for the abstract operators considered here. By construction, $\psi(\underline{Q}^{k+1}) = 1$ holds for all k. If $\underline{Q}_k$ converges to a matrix $\underline{Q}$, we get $\underline{Q} = \frac{s^{\#}(\underline{Q})}{\psi(s^{\#}(\underline{Q}))}$ and so, $s^{\#}(\underline{Q}) = \lambda \underline{Q}$ with $\lambda = \psi(s^{\#}(\underline{Q}))$, which solves problem (6.11).

The algorithm (6.12) is a non-linear analogue of the power algorithm which is familiar in matrix theory [GVL13]. The latter allows one to compute an eigenvector associated to a dominant eigenvalue (eigenvalue of maximal modulus) of a real matrix M by computing the sequence

$$x_{k+1} = \frac{Mx_k}{\|Mx_k\|_2} \tag{6.13}$$

where x_0 is a non-zero vector. This is similar to (6.12), except that we replaced the Euclidean norm $\|\cdot\|_2$ by the linear functional ψ. The well known advantage of the power algorithm is its scalability. To implement it, the matrix M need not be explicitly stored, it suffices to have an oracle which takes x as input and return Mx, hence, it is adapted to instances of large dimension (e.g., the "pagerank" algorithm is a variant of the power iteration). The classical power iteration is known to converge for generic values of the initial vector x_0, provided that the matrix M has a unique eigenvalue of maximal modulus. This is the case in particular when the matrix M has positive entries. It is straightforward to find examples in which the power iteration (6.13) does not converge if the latter positivity condition is relaxed.

Therefore, in order to guarantee that the non-linear iteration (6.12) converges, we need to find an analogue of the classical positivity condition. Geometrically speaking, the latter means that the map $x \mapsto Mx$ sends the cone $\mathbb{R}_+^n$ to its interior. By analogy, it is natural to require that the abstract operator $s^{\#}$ sends the cone $(\mathcal{S}_n^{+})^{\mathcal{W}}$ to its interior, i.e., to require that $s^{\#}(\underline{Q})$ is positive definite as soon as $\underline{Q}$ is non-zero. We can always make sure that this assumption is satisfied by introducing a damping parameter $\varepsilon > 0$ and replacing the operator $s^{\#}$ coordinate-wise by

$$q \mapsto s_w^{\#}(q) + \varepsilon \psi\big[s^{\#}(q)\big]\, I \ .$$

This leads to the damped non-linear power iteration

$$\underline{Q}_w^{k+1} = \frac{s_w^{\#}(\underline{Q}^k) + \varepsilon \psi\big[s^{\#}(\underline{Q}^k)\big] I}{\psi\left[s_w^{\#}(\underline{Q}^k) + \varepsilon \psi\big[s^{\#}(\underline{Q}^k)\big] I\right]} \ . \tag{6.14}$$

We shall refer to (6.14) as the *non-linear additive power iteration* in the sequel, for the ε-perturbation acts in an additive way on s.

The choice of ε will be a trade off between making the perturbation small, which requires to choose a small ε, and ensuring a fast convergence, which is the case when ε is large. For the present experimental purposes, we will see that taking $\varepsilon \in [10^{-2}, 10^{-1}]$ leads to satisfactory results. The interest of the non-linear additive power iteration is its simplicity of implementation.

We also use a variant that uses a multiplicative perturbation instead of an additive one, which experimentally gives comparable results, with $0 < \varepsilon < 1$:

$$\underline{Q}_w^{k+1} = \frac{s_w^{\#}(\underline{Q}^k)^{1-\varepsilon}}{\psi\left[s_w^{\#}(\underline{Q}_k)^{1-\varepsilon}\right]} \ . \tag{6.15}$$

Recall that for all positive semidefinite matrices Y and for all $s > 0$, Y^s denotes the s-th power of Y defined to be the matrix $U^T D^s U$, where $Y = U^T D U$ and D^s is the diagonal matrix obtained by raising to the power s every diagonal entry of D. We say that A^s is the *s-th power* of A, as it coincides with the usual s-th power for integer values of s. For brevity, we write $s^{\#}(q)^{1-\varepsilon}$ for $\left(s^{\#}(q)\right)^{1-\varepsilon}$. We refer to (6.15) as the *non-linear multiplicative power iteration*.

We show in Section 7.3 that these iterations converge independently of their starting point provided ε is large enough.

6.4.2 Benchmarks

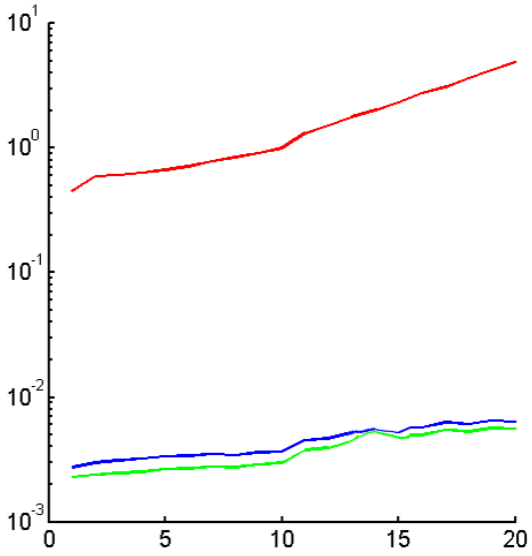
We now experiment the methods that we have introduced, and we compare them with alternative techniques based on LMI. The experiments are implemented in MATLAB, running on one core of an 2.2GHz Intel Core i7 with 8Gb RAM.

We show in Figure 6.6a the average time to find an invariant using LMIs (in red), the additive nonlinear power algorithm (in blue) and the multiplicative power algorithm (in green). These results were obtained on randomly generated programs of the form depicted in Figure 10, where A_1 and A_2 are invertible matrices. For the benchmarks, the power algorithms are always initialized at I_n and the LMI approach for finding an invariant is done by testing the feasibility of the following LMI:

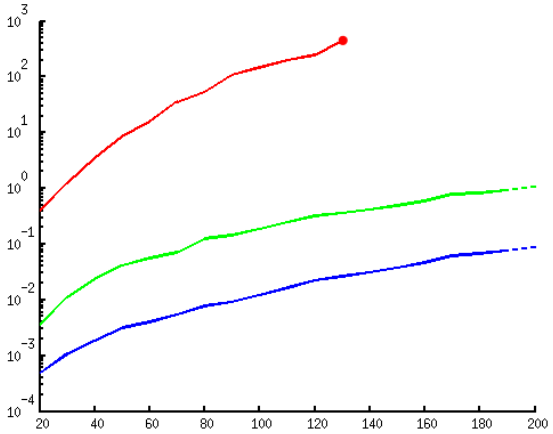
$$\begin{cases} X & \succcurlyeq A_1 X A_1^T \\ X & \succcurlyeq A_2 X A_2^T \\ X & \succ 0 \end{cases} . \tag{6.16}$$

Such a feasible element X is an invariant for the programs described above. We observe that the power type algorithms bring a significant speed-up over the LMI technique.

Furthermore, we compare in Figure 6.6b the execution time of the power algorithms with the resolution of an LMI on a set of high-dimensional linear systems without any switch. The linear systems correspond to parallel simulations of damped oscillators $\ddot{x}_i + c_i \dot{x}_i + k_i x_i = 0$, i.e., given by $A_0 = \begin{pmatrix} I_n & hI_n \\ -hK & I_n - hC \end{pmatrix}$, where $h = 0.05$, and $C, K \in \mathbb{R}^{n \times n}$ are diagonal matrices, respectively with positive diagonal elements c_i and k_i. Unlike Figure 6.6a, the additive power algorithm seems to be faster: here, there is no invariant join computation, hence the cost of the matrix power in the multiplicative algorithm becomes visible. This example highlights



(a) Program invariant, with LMI (red), additive power (blue) and multiplicative power (green)



(b) Program invariant in high dimensions, with an LMI (red), additive power (blue) and multiplicative power (green)

Figure 6.6: Computation times (in s) w.r.t. the dimension of the problem.

Program 10: Structure of programs used in the benchmarks

```
while rand_bool do
    if rand_bool then
        x := A_1 x
    else
        x := A_2 x
    end
end
```

another scalability aspect of the power algorithms: while the semidefinite program approach runs out of memory for systems of dimension 140 and beyond, the computation of an invariant through the power-methods is successful and still runs in less than 2s even when there are 200 variables. Note that when there is no switch, the present power algorithm essentially reduces to the classical power algorithm applied to the linear operator $X \mapsto AXA^T$.

In Table 6.3, we compare our method with an LMI-based approach on a specific set of instances. On top of providing the execution time of the analyses, we also provide the *relative stability margin* of the invariants that we obtain. Given an invariant X, the latter quantity is defined as $\lambda_{\min}\big(X - s(X)\big)/\lambda_{\max}(X)$, where $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ respectively denote the smallest and largest eigenvalues of the matrix M. This quantity is non-negative and well defined as an invariant X satisfies $X \succ 0$ and $X \succcurlyeq s(X)$. A large relative margin ensures that the invariant is stable with respect to rounding errors. Except in the last example, the invariants that we obtain using the two approaches are not comparable. However, we give an estimate of the precision of each invariant by using its largest eigenvalue once it has been rescaled to contain the identity matrix, or, in terms of ellipsoids, the unit ball: a size of 1

Example		Switched oscillator	Switched system		Symplectic operator
ε		0.05	0.1	0.8	0.1
Time (ms)	LMI	160	190	190	100
	add	**5**	**6**	**3**	**1**
	mult	80	15	14	3
Stability margin	LMI	4.10^{-3}	**0.36**	**0.36**	$\leqslant 5.10^{-3}$
	add	4.10^{-4}	0.07	**0.36**	$\leqslant 5.10^{-3}$
	mult	9.10^{-3}	0.02	**0.36**	$\leqslant 5.10^{-3}$
Invariant size	LMI	**1.52**	**1.56**	1.56	**1**
	add	1.91	23.37	2.19	1
	mult	2.48	9.78	**1.50**	1

Table 6.3: Benchmarks on specific examples

Program 11: Implementation of the switched oscillator using an explicit Euler integration scheme

$h = 0.01$, $\omega_0 = 1$, $\omega_1 = 0.8$ are constants
$x \leftarrow [-1, 1]$
$v \leftarrow [-1, 1]$
while *rand_bool* **do**
  if *rand_bool* **then**
    $\begin{pmatrix} x \\ v \end{pmatrix} \leftarrow \begin{pmatrix} x + hv \\ -(h\omega_0^2)x + (1 - h\omega_0)v \end{pmatrix}$
  else
    $\begin{pmatrix} x \\ v \end{pmatrix} \leftarrow \begin{pmatrix} x + hv \\ -(h\omega_1^2)x + (1 - h\omega_1)v \end{pmatrix}$
  end
end

means that the invariant is very close to the unit ball, while greater sizes mean that the ellipsoid spans far from the unit ball in some directions.

The switched oscillator refers to the example of Figure 11. We also consider another switched linear system, already studied in [SH10], characterized by the matrices

$$A_1 = \begin{pmatrix} -0.06515 & -0.4744 & 0.3041 \\ -0.4744 & 0.4872 & 0.3732 \\ 0.3041 & 0.3732 & -0.1271 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 0.04419 & 0.3155 & -0.04247 \\ 0.1451 & -0.04931 & -0.2805 \\ 0.2833 & -0.01418 & 0.1554 \end{pmatrix}.$$

This system allows us to show the importance of the parameter ε by its action on the final quality of the invariant. Indeed, if the power algorithms use $\varepsilon = 0.1$, then the quality of the invariants is quite bad relative to the one computed by the LMI. In contrast, if they use $\varepsilon = 0.8$, then, with even less computation time, the quality of the new invariants similar to the one computed by the LMI.

Finally, we apply the power algorithms to the simulation of the non-damped oscillator $\ddot{x} + c\dot{x} + x = 0$ with $c = 0$. In this case, the energy of the oscillator is preserved. However, the Euler scheme used in the example in Figure 11 is not energy-preserving and even diverges

when applied to this system. This is why we use a variant of a symplectic integration scheme $(x_{n+1}, v_{n+1})^T = S(x_n, v_n)^T$, where $A = \begin{pmatrix} 1-\tau^2/2 & \tau^3/4-\tau \\ \tau & 1-\tau^2/2 \end{pmatrix}$ and $\tau = 0.001$. This integration method preserves a quadratic energy function represented by a positive definite matrix Q, i.e., $(x, v)A^T Q A(x, v)^T = (x, v)Q(x, v)^T$. This means that there is no stability margin. In spite of that, all three methods return an invariant, scalar multiples of the same matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1-\tau^2/4 \end{pmatrix}$ which is very close to the identity matrix. It is remarkable that both power algorithms successfully compute that invariant, as other algorithms may not even find a bounded invariant [AGG12].

CHAPTER 7

Tropical Kraus maps for optimal control of switched systems

In quantum information theory [KBDW83], Kraus maps refer to trace-preserving completely positive linear maps on the cone $\mathcal{S}_n^+$ that take the form

$$X \mapsto \sum_i A_i X A_i^* \quad \text{with} \quad \sum_i A_i^* A_i = I_n \,.$$

Here positive semidefinite matrices of trace 1 are called density matrices and represent quantum probability measures. Kraus maps correspond to the transformations of these measures by quantum channels. A fixed point of a Kraus map (or, equivalently, an eigenvector associated with the eigenvalue 1) represents a quantum measure that is left invariant by the channel.

In program verification and control, a positive semidefinite matrix is used to represent a quadratic Lyapunov functions certifying the stability of a linear transformation if $A^T Q A \preccurlyeq Q$. Such matrices also give sufficient conditions for the stability of a switched linear system, acting as a common quadratic Lyapunov function by $A_i^T Q A_i \preccurlyeq Q$ for all mode i.

We introduce in the chapter the notion of "tropical Kraus maps" which is the analogue to classical Kraus in a control setting. They are multivalued transformations induced by a switched linear system on the space of positive semidefinite matrices.

We show that non-linear eigenvectors of these maps produce Lyapunov-type certificates and that their associated eigenvalue gives an upper bound on the joint spectral radius. We present a Krasnoselskii-Mann-type iteration to compute these eigenvectors and discuss the convergence of our method. Our method is related to the one in [AJPR14] except it completely avoids the recourse to semidefinite programming. This entails a large gain in scalability that is demonstrated in Section 7.4.

We also explore a variant of our approach that enables us to deal with the hybrid optimal control problem defined in Section 5.1.3.

This chapter is based on the articles "Tropical Kraus maps for optimal control of switched systems" [GS17] and "A Scalable Algebraic Method to Infer Quadratic Invariants of Switched Systems" [AGS+16].

7.1 Tropical Kraus maps

7.1.1 Notation and definitions

We define a *tropical Kraus map* as a *multivalued map* from $\mathcal{S}_n^+$ to $\wp(\mathcal{S}_n^+)$ by

$$X \mapsto \bigvee_i A_i X A_i^T \,. \tag{7.1}$$

It has a similar structure to the classical Kraus map K, except the sum has been replaced with the "supremum" operation $\bigvee$ (the set of minimal upper bounds). Moreover, it is convenient to omit the normalization requirement. This is why, in light of tropical algebra where the usual sum is replaced by the maximum, this map is christened a *tropical* Kraus map. As we work with real quadratic forms, instead of hermitian forms, the hermitian conjugate * is replaced by transposition ^{T}.

It will be useful to consider vector versions of tropical Kraus maps T on $(\mathcal{S}_n^+)^{\mathcal{W}}$, whose i-th coordinate for $i \in \mathcal{W}$ is the tropical Kraus map:

$$X \mapsto \bigvee_j A_{ij} X_j A_{ij}^T \,. \tag{7.2}$$

The latter map can be written as a standard tropical Kraus map by identifying the cone $(\mathcal{S}_n^+)^{\mathcal{W}}$ to the sub-cone of $n \times n$ block-diagonal matrices in $\mathcal{S}_{n|\mathcal{W}|}^+$ via the map:

$$(X_w)_{w\in\mathcal{W}} \mapsto \operatorname{diag}\left(X_w\right)_{w\in\mathcal{W}} .$$

So the coordinate i is obtained as the i-th diagonal block of the map

$$X \mapsto \bigvee_j B_{ij} X B_{ij}^T \quad \text{with} \quad \begin{cases} X & = \operatorname{diag}\left(X_w\right)_{w\in\mathcal{W}} \\ B_{ij} & = E_{ij} \otimes A_{ij} \end{cases} ,$$

where $\otimes$ is the Kronecker product. Indeed, the matrices B_{ij} preserve the block-diagonal structure of the matrices:

$$\begin{aligned} B_{ij} X B_{ij}^T &= (E_{ij} \otimes A_{ij})(E_{kk} \otimes \sum_k X_k)(E_{ij} \otimes A_{ij})^T \\ &= \sum_k (E_{ij} \otimes A_{ij})(E_{kk} \otimes X_k)(E_{ij} \otimes A_{ij}^T) \\ &= \sum_k (E_{ij} E_{kk} E_{ji}) \otimes (A_{ij} X_k A_{ij}^T) \\ &= E_{ii} \otimes (A_{ij} X_j A_{ij}^T) \,, \end{aligned}$$

because $E_{xy}E_{zw} = E_{xw}$ if $y = z$ and 0 otherwise.

Moreover, since the cone $\mathcal{S}_n^+$ is self-dual, the cone $(\mathcal{S}_n^+)^{\mathcal{W}}$ is also self-dual. Thus, by Theorem 1.15, a minimal upper bound Y_j of $\{A_{ij}X_jA_{ij}^T \colon 1 \leqslant i, j \leqslant N\}$ is selected by a positive definite matrices C_j for all j if and only if the matrix $Y = \operatorname{diag}(Y_1, \ldots, Y_N)$ is a minimal upper bound of $\{B_{ij}XB_{ij}^T \colon 1 \leqslant i, j \leqslant N\}$ selected by $C \coloneqq (C_1, \ldots, C_N)$. For this reason, we shall also refer to maps of the form in Equation (7.2) as tropical Kraus maps. An element X of the cone $(\mathcal{S}_n^+)^{\mathcal{W}}$ satisfies $X_i \succcurlyeq 0$ for all i, and we abbreviate this notation to simply write $X \succcurlyeq 0$.

Given a minimal upper bound selection, a tropical Kraus map specializes into a map from $(\mathcal{S}_n^+)^{\mathcal{W}}$ to $(\mathcal{S}_n^+)^{\mathcal{W}}$ by choosing for each X the minimal upper bound of $\{A_iXA_i^T \colon i \in \mathcal{W}\}$ selected by the selection process. We shall in particular deal with the invariant selection defined in Chapter 3, yielding the map from $\mathcal{S}_n^+$ to $\mathcal{S}_n^+$ by

$$X \mapsto \bigsqcup_i A_iXA_i^T \,.$$

7.1.2 Tropical Kraus map associated with a switched linear system

The latter map is identical to the coordinates of the operator $s^{\#}$ defined in Equation (6.10) to compute a disjunctive invariant of a linear switched system like Program 5 of the form $\cup_w \mathcal{E}_w$ where $\mathcal{E}_w$ are ellipsoids. In particular, the convex hull of the latter set $M \coloneqq \operatorname{conv}(\cup_w \mathcal{E}_w)$ is an invariant symmetric convex body, and by convexity we have

$$\operatorname{conv}\Big[\bigcup_\sigma A_\sigma \cdot M\Big] \subseteq \lambda M \,.$$

The analysis in Chapter 6 thus falls in the case of the computation of a Protasov ball. We recall that λ provides an upper bound on the join spectral radius $\rho(\mathcal{A})$ defined by

$$\rho(\mathcal{A}) \coloneqq \lim_{k \to +\infty} \max_{1 \leqslant \sigma_1, \ldots, \sigma_k \leqslant p} \|A_{\sigma_1} \ldots A_{\sigma_k}\|^{1/k} \;,$$

and that this definition is independent of the choice of the norm.

We develop in this chapter a method to compute *approximate Barabanov norms*, so we work with the collection of adjoint matrices $\mathcal{A}^T$. The transposition is not surprising: in addition to the remark made in Section 5.1.2, classical Kraus maps provide a forward propagation of density matrices, whereas we are interested in Lyapunov functions, whose propagation follows a backward scheme. I.e., the present tropical Kraus maps are analogues to the *adjoints* of classical Kraus maps.

We consider a switched linear program that switches between p modes, indexed by $\Sigma = \{1, \ldots, p\}$, whose σ-th mode is given by the matrix A_σ. We say that $(i, \sigma, j) \in \mathcal{W} \times \Sigma \times \mathcal{W}$ is an *admissible transition* when $\tau(i, \sigma) = j$. Our analysis deals with the tropical Kraus map associated with this switched linear program. Let $\mathcal{W}$ denote a subset of Σ^* and τ denote a map from $\mathcal{W} \times \Sigma$ to $\mathcal{W}$. The triple $(\Sigma, \mathcal{W}, \tau)$ defines a deterministic finite automaton, whose alphabet is Σ, whose states are elements of $\mathcal{W}$ and whose transition function is τ. The tropical Kraus map is defined from $(\mathcal{S}_n^+)^{\mathcal{W}}$ to $(\wp(\mathcal{S}_n^+))^{\mathcal{W}}$, and its j-th coordinate maps $X = (X_w)_{w \in \mathcal{W}} \in (\mathcal{S}_n^+)^{\mathcal{W}}$ to the subset of $\mathcal{S}_n^+$:

$$T_j^{\mathrm{lin}} \colon X \mapsto \bigvee \Big\{ A_\sigma^T X_i A_\sigma \colon (i, \sigma) \in \mathcal{W} \times \Sigma \,, i \cdot \sigma = j \Big\} \,.$$

We also consider a variant of the tropical Kraus map, adapted to the optimal control problem in Section 5.1.3 and Program 6, defining the map M_τ from $(\mathcal{S}_n^+)^{\mathcal{W}}$ to $(\wp(\mathcal{S}_n^+))^{\mathcal{W}}$ by:

$$(M^t)_j(X) \coloneqq \bigvee \left\{ \operatorname{ricc}_{t,\sigma} X_i \colon (i,\sigma) \in \mathcal{W} \times \Sigma \,, i \cdot \sigma = j \right\}.$$

We recall that $\operatorname{ricc}_{t,\sigma}$ denotes the Riccati flow in time t of the mode σ, which maps a positive semidefinite matrix P to the positive semidefinite matrix $Q = \operatorname{ricc}_{t,\sigma} P$ obtained as the solution of the optimal control problem

$$S_t^\sigma[V](x) = \sup_{u \in \mathcal{U}} \int_0^t \frac{1}{2}\xi(s)^T D^\sigma \xi(s) - \frac{\gamma^2}{2}|u(s)|^2 \, ds + V\big(\xi(t)\big) \,, \qquad \text{(5.1-recalled)}$$
$$V(x) = x^T P x \,, \qquad S_t^\sigma[V](x) = x^T Q x \,.$$

Remark 7.1. The operator $\operatorname{ricc}_{t,\sigma}$ maps the cone of positive semidefinite matrices into itself. This key property is not valid for the operator considered in Equation (6.1) for affine systems. This is why we do not extend the definition of tropical Kraus maps to affine systems but allow the former case.

7.1.3 Non linear eigenvalue and fixed point problems associated to Tropical Kraus Maps

The tropical Kraus map T^{lin} is positively homogeneous, meaning that $T^{\mathrm{lin}}(\alpha X) = \alpha T^{\mathrm{lin}}(X)$ for all $X \in \mathcal{S}_n^+$ and $\alpha \geqslant 0$. This suggests to consider a multivalued eigenproblem.

Definition 7.1. A *(non-linear) eigenvector* of T, associated to the *eigenvalue* λ is an element $X \in (\mathcal{S}_n^+)^{\mathcal{W}}$ such that $\lambda X_j \in T_j^{\mathrm{lin}}(X)$ holds for all $j \in \mathcal{W}$. We write $\lambda X \in T^{\mathrm{lin}}(X)$ for brevity.

This notation is licit since we can identify $T^{\mathrm{lin}}(X)$ which is an element of $(\wp(\mathcal{S}_n))^{\mathcal{W}}$ to an element of $\wp\big((\mathcal{S}_n)^{\mathcal{W}}\big)$.

The following result shows that a non-linear eigenvalue of the tropical Kraus map provides an upper bound for the joint spectral radius.

Theorem 7.1. *If the multivalued eigenvector problem $\lambda X \in T^{lin}(X)$ has a positive semidefinite solution such that the matrix $\sum_{w \in \mathcal{W}} X_w$ is positive definite, then, the map*

$$v(z) \coloneqq \sup_{w \in \mathcal{W}} (z^T X_w z)^{1/2}$$

is a norm, and $v(A_\sigma z) \leqslant \sqrt{\lambda} v(z)$ holds for all $z \in \mathbb{R}^n$ and $\sigma \in \Sigma$. In particular, the joint spectral radius of $\mathcal{A}$ does not exceed $\sqrt{\lambda}$.

Proof. Let (X, λ) denote such an eigenvector-eigenvalue pair and v the map defined above. The map v is clearly non-negative, positively homogeneous and satisfies the triangular inequality. We only show that it is positive definite: let $z \in \mathbb{R}^n$ such that $v(z) = 0$, i.e., $z^T X_w z = 0$ for all w. The matrices X_w are positive semidefinite, hence z belongs to the kernel of all the matrices X_w. In particular, z belongs to the kernel of their sum, which is positive definite, hence $z = 0$. Thus v defines a norm on $\mathbb{R}^n$.

The pair (X, λ) satisfies $A_\sigma^T X_i A_\sigma \preccurlyeq \lambda X_{\tau(i,\sigma)}$ for all $(i, \sigma) \in \mathcal{W} \times \Sigma$. Hence

$$v(A_\sigma z) = \sup_{w \in \mathcal{W}} \left[z^T A_\sigma^T X_w A_\sigma z\right]^{1/2} \leqslant \sup_{w \in \mathcal{W}} \lambda^{1/2} \left[z^T X_{\tau(w,\sigma)} z\right]^{1/2} \leqslant \sqrt{\lambda} v(z) \,. \qquad (7.3)$$

The norm v on $\mathbb{R}^n$ induces a subordinate norm $\|\cdot\|_v$ on $\mathcal{S}_n$ by

$$\|M\|_v \coloneqq \max_{z\neq 0} \frac{v(Mz)}{v(z)}\,.$$

We can bound the product $A_{\sigma_1}\dots A_{\sigma_k}$ in this norm by successive applications of Equation (7.3) as follows:

$$\|A_{\sigma_1}\dots A_{\sigma_k}\|_v \leqslant \sup_{z\neq 0} \frac{v(A_{\sigma_1}\dots A_{\sigma_k}z)}{v(z)} \leqslant \lambda^{k/2}\,.$$

Taking the k-root and the supremum over all products of length k as k tends to $+\infty$, we deduce that $\rho(\mathcal{A}) \leqslant \sqrt{\lambda}$. □

We have the an analogous result for the switched linear quadratic control problem.

Theorem 7.2. *If the multivalued fixed point problem $X \in M^t(X)$ has a solution, then the map $V \coloneqq z \mapsto \sup_{w\in\mathcal{W}} z^T X_w z$ determines a sub-invariant function of all the Lax-Oleinik semi-groups S_t^σ, meaning that:*

$$\max_{\sigma\in\Sigma} S_t^\sigma[V](z) \leqslant V(z) \text{ for all } z\,.$$

Proof. A solution X satisfies $\mathrm{ricc}_{t,\sigma} X_i \preccurlyeq X_{\tau(i,\sigma)}$ for all $(i,\sigma) \in \mathcal{W}\times\Sigma$. By max-plus linearity of the semi-group S_t, we have

$$\begin{aligned}
\max_{\sigma\in\Sigma} S_t^\sigma[V](z) &= \max_{\sigma\in\Sigma}\max_{i\in\mathcal{W}} S_t^\sigma[x\mapsto x^T X_i x](z)\\
&= \max_{\sigma\in\Sigma}\max_{i\in\mathcal{W}} z^T(\mathrm{ricc}_{t,\sigma} X_i)z\\
&\leqslant \max_{\sigma\in\Sigma}\max_{i\in\mathcal{W}} z^T X_{\tau(i,\sigma)}z\\
&\leqslant V(z)\,.
\end{aligned}$$

□

7.1.4 Non-linear eigenvectors of tropical Kraus maps and computation by a Krasnoselskii-Mann iteration

For a completely positive map, $X \mapsto \sum_i A_i X A_i^T$, the existence of a positive semidefinite eigenvector follows from the Perron-Frobenius theorem [LN12b]. Moreover, such an eigenvector is necessarily positive definite as soon as the map is *irreducible* in the Perron-Frobenius sense, meaning that the map does not leave invariant a non-trivial face of the closed cone $\mathcal{S}_n^+$. As shown in [Far96], the latter condition holds if and only if the set of matrices $\{A_i\}$ is *irreducible* in the algebraic sense, meaning that there is no non-trivial subspace invariant by each matrix in this set.

In order to show that tropical Kraus maps have eigenvectors, we specialize the multivalued map T defined in Section 7.1.1 by fixing a selection of minimal upper bound $\mho$. We obtain the map T defined on $(\mathcal{S}_n^+)^{\mathcal{W}}$ by

$$T_j(X) \coloneqq \mho\Big\{A_\sigma^T X_i A_\sigma : i\cdot\sigma = j\Big\}\,.$$

We will prove that the map T has a non-linear eigenvector if the selection $\mho$ is the invariant join $\sqcup$. We introduce the "non-commutative simplex"

$$\Delta_{\mathcal{W}} \coloneqq \{X \in (\mathcal{S}_n^+)^{\mathcal{W}} \colon \sum_w \langle I_n, X_w \rangle = 1\}$$

and the map $\widehat{T}$ sending $\Delta_{\mathcal{W}}$ to itself:

$$\widehat{T}(X) \coloneqq \frac{1}{2}\Big[\frac{1}{\sum_w \langle I_n, T_w(X)\rangle} T(X) + X\Big]\,. \tag{7.4}$$

Observe that, independently of the selection $\mho$, a fixed point $X \in \Delta_{\mathcal{W}}$ of the map $\widehat{T}$ yields an eigenvector for the map T associated with the eigenvalue $\sum_w \langle I_n, T_w(X)\rangle$. We can now state the theorem.

Theorem 7.3. *If the set of matrices $\{E_{ij} \otimes A_\sigma \colon \tau(i,\sigma) = j\}$ is irreducible, then the map $\widehat{T}$ has a positive definite fixed point.*

This is proved in Section 7.2. We point out that assuming the irreducibility of the set of matrices $\{E_{ij} \otimes A_\sigma \colon \tau(i,\sigma) = j\}$ is a stronger statement than assuming the irreducibility of $\mathcal{A}$.

We obtain as an immediate corollary:

Corollary 7.4. *If the set of matrices $\{E_{ij} \otimes A_\sigma \colon \tau(i,\sigma) = j\}$ is irreducible, then, the tropical Kraus map T has a positive definite eigenvector.*

In order to compute a fixed point of the map $\widehat{T}$, we compute successive iterates starting from a positive definite matrix $X^{(0)}$. We shall consider the following scheme

$$X^{k+1} = \frac{1}{2}\Big[\frac{1}{\sum_w \langle I_n, T_w(X^k)\rangle} T(X^k) + X^k\Big]\,.$$

This is a power-type iteration, involving a renormalization and a "damping term" (addition of $X^{(k)}$) to avoid oscillations. This is inspired by the classical Krasnoselskii-Mann iteration, which applies to non-expansive mappings T, and takes the form $X^{(k+1)} = (T(X^{(k)}) + X^{(k)})/2$, see [RZ00]. We discuss the convergence of this scheme in Section 7.2.5.

Remark 7.2. There is a multiplicative variant of the iteration, defined by

$$X^{k+1} = \Big[\frac{T(X^k)}{\sum_w \langle I_n, T_w(X^k)\rangle}\Big] \# X^k\,,$$

where $P \# Q \coloneqq P^{1/2}\big(P^{-1/2} Q P^{-1/2}\big)^{1/2} P^{1/2}$ denotes the *Riemannian barycenter* of the positive definite matrices P, Q, see [Bha07, Chapter 2] for more information. This multiplicative version does converge in the "commutative case", i.e., when $n = 1$. Indeed, the map $X \mapsto \sum_w \langle I_n, T_w(X)\rangle^{-1} T(X)$ is then nonexpansive in the Hilbert metric [LN12b], and then, the general result of [RZ00] can be applied. The additive version can also be shown to be converging when $n = 1$, by a reduction to the same result, but the proof is more involved.

We use a different iteration scheme to compute fixed points of the variant M^t defined in terms of Riccati flows. First, we specialize again the multivalued map M^t with a minimal upper bound selection $\mho$

$$X^{k+1} = (M^t)^{\mathrm{sel}}(X^k)\,. \tag{7.5}$$

Contrary to the map T, the operator M^t is not positively homogeneous, hence there is no need to renormalize as in Equation (7.4).

Although this iteration converge quite well for a variety of selection such as the invariant join or the minimum trace selection, see Section 7.4.3, proving the convergence remains an open problem. It is known that the (indefinite) Riccati flow is a contraction in the Thompson metric, with contraction rate $\alpha > 0$ determined by the parameters of the flow [GQ14b, Corollary 4.7], but this constant α depends on a compact invariant interval of the form $\{X \in \mathcal{S}_n^+ \colon \lambda_1 I_n \preccurlyeq X \preccurlyeq \lambda_2 I_n\}$.

Unfortunately, the selections mentioned earlier do not preserve this interval, i.e., it is possible to find two positive semidefinite matrices whose eigenvalues are smaller than λ_2, but the spectrum of their associated minimal upper bound crosses this threshold. It is not hopeless, since there are selections that preserve this interval:

Given a collection of matrices $\{Q_i\}_{1 \leqslant i \leqslant p} \subseteq \mathcal{S}_n^+$, let $\lambda \coloneqq \max_i \lambda_{\max}(Q_i)$ denote the largest eigenvalue of among the matrices Q_i. A selection that preserves such intervals is then given by the unique optimal solution of the following semidefinite program, for a positive definite matrix C:

$$\begin{aligned} \text{minimize} \quad & \langle C, X \rangle \\ & Q_i \preccurlyeq X\ ,\ \forall i \\ & X \preccurlyeq \lambda I_n \end{aligned}$$

If such a selection process is shown to be Lipschitz in the Thompson metric with constant L_n, we can combine these results to show that the iteration is guaranteed to converge when $\exp(\alpha t) > L_n$. This remains an open problem:

Open problem 7.5. *Is there a minimal upper bound selection in $\mathcal{S}_n^{++} \times \mathcal{S}_n^{++}$ that preserves intervals of the form $\{X \in \mathcal{S}_n^+ \colon \lambda_1 I_n \preccurlyeq X \preccurlyeq \lambda_2 I_n\}$ and has a finite Lipschitz constant in Thompson's metric ?*

7.1.5 "Relaxation" of the graph-Lyapunov-function approach

The iterative algorithm presented in Section 7.1.4 is a relaxation of a variant of the LMI $(\mathcal{P}_\rho)$, that has been considered by Ahmadi et al. that is recalled here:

$$\begin{aligned} \rho^2 X_j \succcurlyeq A_\sigma^T X_i A_\sigma\ , \quad & \forall (i, \sigma, j) \colon \tau(i, \sigma) = j\ , \\ X_i \succ 0\ , \quad & \forall i\,. \end{aligned} \tag{$\mathcal{P}_\rho$}$$

Indeed, if we pad the right-hand side of each inequality and minimize the trace of the matrices X_i satisfying these new inequalities, we obtain the semidefinite program

$$\begin{aligned} \text{minimize} \quad & \sum_i \operatorname{trace} X_i \\ & \rho^2 X_j \succcurlyeq A_\sigma^T X_i A_\sigma + \varepsilon I_n\ , \qquad \forall (i, \sigma, j) \colon \tau(i, \sigma) = j\ , \\ & X_i \succ 0\ , \qquad \forall i\,. \end{aligned} \tag{7.6}$$

Then optimal solutions of Equation (7.6) are nonlinear eigenvectors of a perturbed tropical Kraus map:

Proposition 7.6. *Let $\rho > 0$ such that Equation $(\mathcal{P}_\rho)$ has a strict solution and $\varepsilon > 0$. Problem* (7.6) *admits an optimal solution. Any optimal solution X satisfies*

$$\rho^2 X_j \in \bigvee_{\tau(i,\sigma)=j} A_\sigma^T X_i A_\sigma + \varepsilon I_n \,.$$

Proof. Introducing dual variables $\Lambda_{i,\sigma} \succcurlyeq 0$ and $\mu_i \succcurlyeq 0$, the Lagrangian of this LMI is written

$$L(X, \Lambda, \mu) = \sum_i \operatorname{trace} X_i - \sum_{i,\sigma} \langle \Lambda_{i,\sigma}, \rho^2 X_{\tau(i,\sigma)} - A_\sigma^T X_i A_\sigma - \varepsilon I \rangle - \langle \mu_i, X_i \rangle \,.$$

The dual problem is then written

$$\begin{aligned} \text{maximize} \quad & \varepsilon \sum_{i,\sigma} \operatorname{trace} \Lambda_{i,\sigma} \\ & \Lambda_{i,\sigma} \succcurlyeq 0 \\ & \mu_i \succcurlyeq 0 \\ I_n + \sum_\sigma A_\sigma \Lambda_{j,\sigma} A_\sigma^T &= \mu_j + \sum_{\tau(i,\sigma)=j} \Lambda_{i,\sigma} \,, \forall j \end{aligned}$$

The dual problem is strictly feasible, since for $\eta > 0$ small enough, choosing $\Lambda_{i,\sigma} = \eta I_n$ yields $\mu_j = I_n - O(\eta) \succ 0$. Hence strong duality holds and there is a primal-dual optimal solution to the pair of primal-dual problems, see [Bar]. Given such an optimal solution (X, Λ, μ), complementary slackness implies that $\langle \mu_j, X_j \rangle = 0$ holds for all i . However, we must have $\rho^2 X_i \succcurlyeq \varepsilon I_n \succ 0$. Hence $\mu_j = 0$ and $\sum_{\tau(i,\sigma)=j} \Lambda_{i,\sigma} \succcurlyeq I_n \succ 0$, i.e. $\sum_{\tau(i,\sigma)=j} \operatorname{ran} \Lambda_{i,\sigma} = \mathbb{R}^n$. By Theorem 1.15 and complementary slackness, it implies that $\rho^2 X_j$ is a minimal upper bound of $\{A_\sigma^T X_i A_\sigma + \varepsilon I_n \colon \tau(i,\sigma) = j\}$. □

7.2 Existence of nonlinear eigenvectors

We denote in this section by T the tropical Kraus map from $\mathcal{S}_n^+$ to $\mathcal{S}_n^+$ specialized with the invariant join defined by

$$T \colon X \mapsto \bigsqcup_{1 \leqslant i \leqslant N} A_i X A_i^T \,.$$

7.2.1 Inequalities between classical and tropical Kraus maps

The key ingredient for showing the existence of a nonlinear eigenvector is a double inequality which generalizes the inequalities $\frac{1}{N}\sum_k x_k \leqslant \max_k x_k \leqslant \sum_k x_k$ for non-negative reals $(x_k)_{1 \leqslant k \leqslant N}$ to the non-commutative case when the "maximum" is the invariant join.

Proposition 7.7. *The inequalities*

$$\frac{1}{N} K(X) \preccurlyeq T(X) \preccurlyeq K(X)$$

holds for all $X \in \mathcal{S}_n^+$.

Proof. The lower bound is a consequence of Remark 3.10. The upper bound follows from Theorem 3.17. □

This result shows that the distance between the values of the classical and tropical Kraus maps remains bounded in the Hilbert metric. It also proves that $T(X)$ and $K(X)$ belong to the same face of the cone $\mathcal{S}_n^+$ for all $X \in \mathcal{S}_n^+$.

A similar result holds for iterated versions of the classical and tropical Kraus maps:

Corollary 7.8. *The inequalities*

$$\frac{1}{N^q} K^q(X) \preccurlyeq T^q(X) \preccurlyeq K^q(X)$$

holds for all $X \in \mathcal{S}_n^+$.

Proof. We write $f \preccurlyeq g$ to mean $f(X) \preccurlyeq g(X)$ for all $X \in \mathcal{S}_n^+$. We prove this result by induction on q. The case $q = 1$ is true by Proposition 7.7. Assume that $N^{-1}K^q \preccurlyeq T^q \preccurlyeq K^q$. By Proposition 7.7, we have $T(T^q) \preccurlyeq K(T^q)$. Moreover, the map K is monotone, so $K(T^q) \preccurlyeq K(K^q)$, which shows that $T^{q+1} \preccurlyeq K^{q+1}$. Similarly, by Proposition 7.7, we have $T(T^q) \succcurlyeq N^{-1}K(T^q)$ and the monotony of the map K implies $K(T^q) \succcurlyeq N^{-q}T^{q+1}$, so that $T^{q+1} \succcurlyeq N^{-q-1}K^{q+1}$. □

It also follows from Proposition 7.7 and Corollary 7.8 that the map T (resp. T^q) sends $\mathcal{S}_n^+ \setminus \{0\}$ to its interior if and only if K (resp. K^q) does. When $q = 1$, we say that the maps T and K are strictly positive. Gaubert and Qu have shown [GQ14a] that checking the strict positivity for K is NP-hard, thus it is also the case for T. We say that T is primitive if there is some $q \geqslant 1$ such that T^q is strictly positive. Corollary 7.8 implies that T is primitive if and only if K is primitive. Finally, we extend the definition of irreducibility to the tropical Kraus map T, i.e. T is irreducible if there is no non-trivial face of the cone $\mathcal{S}_n^+$ that is left invariant by the map T. Again, Proposition 7.7 implies that T is irreducible if and only if K is irreducible.

7.2.2 Existence of non-linear eigenvectors

We now state a theorem regarding the existence of nonlinear eigenvectors under varying assumptions which are analogues of results guaranteeing the existence of eigenvectors of classical Kraus maps, respectively when the tropical Kraus map is strictly positive, primitive and irreducible.

Theorem 7.9. *There is a positive definite* Q *and* $\lambda > 0$ *such that* $T(Q) = \lambda Q$ *if the tropical Kraus map* T *is irreducible. In particular, this is the case when* T *is primitive or strictly positive.*

We obtain Theorem 7.3 as a corollary of Theorem 7.9.

7.2.3 Proof of Theorem 7.9

We first prove the theorem when T is strictly positive. We then prove the case when T is primitive. Finally, we prove the most general case when T is irreducible.

Case 1: T is strictly positive Let $\Delta \coloneqq \{X \in \mathcal{S}_n^+ \mid \operatorname{trace} X = 1\}$. Observe that Δ is compact and convex. Let also $\widehat{T}(X) := \operatorname{trace}(T(X))^{-1} T(X)$.

The map K is continuous on $\mathcal{S}_n^+$, so we have $\alpha I \leqslant K(X) \leqslant \beta I$ for all $X \in \Delta$ for some non-negative α, β. Since T is strictly positive, the map K is strictly positive, hence α is positive. It follows that $\widehat{T}$ maps Δ to a bounded subset of Δ in the Hilbert metric: $\widehat{T}(\Delta) \subseteq B_H(I, R) \cap \Delta$, with $R = \log(N\beta/\alpha)$.

In particular, since the map T is continuous on the interior of $\mathcal{S}_n^+$, $\widehat{T}$ sends continuously $B_T(e, R)$ to itself. We apply Brouwer's theorem to the map $\widehat{T}$ and obtain an eigenvector Q.

Case 2: T is primitive Similarly to the previous proof, we shall show that there is a subset of the interior of Δ that is compact, convex and invariant by $\widehat{T}$. Since the map T is not convex, it is not clear that the set $\widehat{T}^q(\Delta)$ is convex. Instead, we use a "convexified" version of this set. Recall that the map T is positively-homogeneous, so that for any cone $\mathcal{C} \subset \mathcal{S}_n^+$ the following equality holds

$$(\operatorname{cone} \circ T)\big(\mathcal{C}\big) \cap \Delta = (\operatorname{conv} \circ \widehat{T})\big(\mathcal{C}\big)\,.$$

Here $\operatorname{conv}(\mathcal{X})$ denotes the convex hull of the set $\mathcal{X}$ and $\operatorname{cone}(\mathcal{X})$ is the *conic hull* of $\mathcal{X}$. The map $\mathcal{X} \mapsto (\operatorname{cone} \circ T)(\mathcal{X})$ is multivalued and maps a set $\mathcal{X}$ to the cone generated by $T(\mathcal{X})$. In this spirit, we establish the following technical lemma.

Lemma 7.10. *For all $y \in (\operatorname{cone} \circ T)^q\big(\mathcal{S}_n^+\big)$, there is $x \in \mathcal{S}_n^+$ such that*

$$N^{-q} K^q(x) \preccurlyeq y \preccurlyeq K^q(x)\,.$$

Proof. We work by induction on q. The case $q = 0$ is trivial. Assume now that the property holds for some q. Let $y \in (\operatorname{cone} \circ T)^{q+1}(\mathcal{S}_n^+)$, meaning that there are some $z_k \in (\operatorname{cone} \circ T)^q(\mathcal{S}_n^+)$ and $\lambda_k \in \mathbb{R}^+$ such that $y = \sum_k \lambda_k T(z_k)$. By Corollary 7.8, we have

$$N^{-1} \sum_k \lambda_k K(z_k) \preccurlyeq y \preccurlyeq \sum_k \lambda_k K(z_k)\,.$$

Moreover, there are by assumption some $x_k \in \mathcal{S}_n^+$ such that

$$N^{-q} K^q(x_k) \preccurlyeq z_k \preccurlyeq K^q(x_k)\,.$$

The map K is linear and monotone, so

$$N^{-q-1} \sum_k \lambda_k K^{q+1}(x_k) \preccurlyeq y \preccurlyeq \sum_k \lambda_k K^{q+1}(x_k)\,.$$

The set $K^{q+1}(\mathcal{S}_n^+)$ is convex, thus $\sum_k \lambda_k K^{q+1}(x_k) \in K^{q+1}(\mathcal{S}_n^+)$, which concludes the proof. □

Let now q such that K^q (or equivalently T^q) is strictly positive. Following the same argument as in the proof when T is strictly positive, there are positive α, β such that $\alpha I \preccurlyeq K^q(X) \preccurlyeq \beta I$ for all $X \in \Delta$.

Let $\mathcal{P}$ denote the set $(\operatorname{cone} \circ T)^q(\mathcal{S}_n^+) \cap \Delta$. Using Lemma 7.10, we deduce from the previous inequalities that for all $X \in \mathcal{P}$, we have

$$\alpha N^{-q} I \preccurlyeq X \preccurlyeq \beta I\,,$$

so the set $\mathcal{P}$ is a convex bounded subset of the interior of $\mathcal{S}_n^+ \cap \Delta$. Let $\overline{\mathcal{P}}$ denote the closure in the topology induced by the Hilbert metric of $\mathcal{P}$ (Recall that the topology of the Hilbert metric is the same as the Euclidean topology on the interior of Δ because the cone $\mathcal{S}_n^+$ is finite-dimensional, see [LN12a, AGN15]). We have

$$\widehat{T}(\mathcal{P}) \subset (\mathrm{cone} \circ T)(\mathcal{P}) \cap \Delta \subset (\mathrm{cone} \circ T)^{q+1}(\mathcal{S}_n^+) \cap \Delta \subset \mathcal{P}\,,$$

since all operators are monotone as set-valued maps. Moreover, since the map $\widehat{T}$ is continuous on the interior of Δ, we have $\widehat{T}(\overline{\mathcal{P}}) \subset \overline{\widehat{T}(\mathcal{P})}$, so that $\widehat{T}(\overline{\mathcal{P}}) \subset \overline{\mathcal{P}}$. Thus, the map $\widehat{T}$ sends the bounded closed (hence compact) convex set $\overline{\mathcal{P}}$ continuously into itself. We apply Brouwer's theorem to the map $\widehat{T}$ and obtain an eigenvector Q.

Case 3: T is irreducible Note that the matrix Q is an eigenvector of the map T with eigenvalue λ if and only if it is an eigenvector of the map $T_1 \coloneqq X \mapsto T(X) + X$ with eigenvalue $\lambda + 1$.

Moreover, since $0 \preccurlyeq X \preccurlyeq T_1(X)$, the face $F(X)$ must be a subset of the face $F\big(T_1(X)\big)$. If the matrix X is nonzero and not invertible, this inclusion must be strict. Indeed, by the inequality $T(X) \preccurlyeq T_1(X)$, the fact that the equality $F(X) = F\big(T_1(X)\big)$ implies that the map T leaves the non-trivial face $F(X)$ invariant, which contradicts the irreducibility assumption. Hence we must have[1]

$$0 < \mathrm{rk}\, X < n \implies \mathrm{rk}\, X < \mathrm{rk}\, T_1(X)\,.$$

It follows that the n-th iterate of the map T_1 maps nonzero positive semidefinite matrices to positive definite matrices, i.e. the map T_1 is primitive. The existence of an eigenvector follows from Case 2. □

7.2.4 Obstacles for simpler proofs

Several basic methods allow one to prove non-linear extensions of the Perron-Frobenius theorem. These involve contraction properties with respect to Hilbert's projective metric, Brouwer fixed point theorem, or monotonicity properties, see [LN12b]. These approaches fail in the case of the specialized tropical Kraus map T described in this chapter.

Indeed, the map T is not always contracting: applying Theorem 4.1 to tropical Kraus maps only yields an upper bound on the Lipschitz constant in the Riemann metric of $\sqrt{p}$ and we show in Section 7.2.5.a an example where T is expansive on an open set.

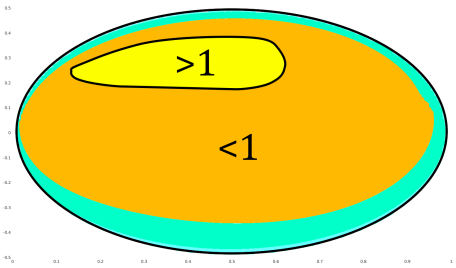
Similarly, one can observe that the map T may not be monotone, which is not surprising since the invariant join is not monotone either by Proposition 3.20.

Finally, the invariant join does not have a continuous extension to the closure on the cone of positive definite matrices, hence it is also not surprising that the tropical Kraus map T is not continuous either on the closed cone $\mathcal{S}_n^+$.

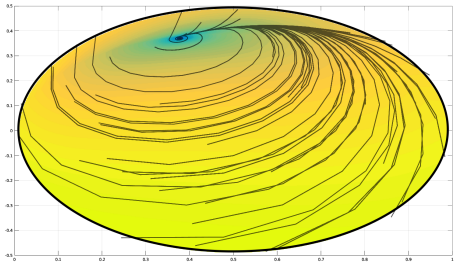
7.2.5 On the convergence towards a non-linear eigenvector

We present two special cases which highlight the difficulties that are encountered in an attempt to prove convergence of the scheme towards a non-linear eigenvector. We exhibit in the

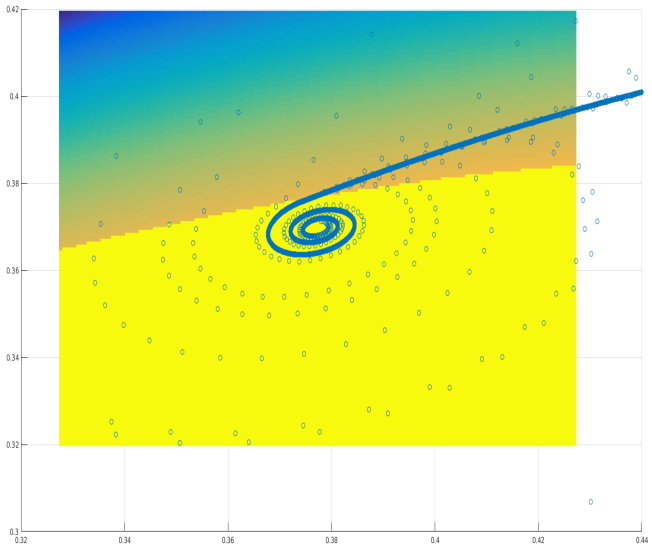
[1] Recall that a matrix Y belongs to the face $F(Z)$ if and only if the image of Y is contained in the image of Z.



(a) Value of the local Lipschitz constant



(b) Traces of iterations from random points



(c) Close-up look at the convergence towards the eigenvector

Figure 7.1: Lipschitz constant, fixed point and trajectory of iterated tropical Kraus maps on the slice Δ

following two examples in their simplest form: the tropical Kraus maps send $\mathcal{S}_2^+$ into $\mathcal{S}_2^+$ and take the form

$$T\colon X \mapsto AXA^T \sqcup BXB^T\,,$$

where A, B are 2×2 real matrices. We study the values of the map $\widehat{T}$ on $\Delta := \{X \in \mathcal{S}_n^+ :\ \mathrm{trace}(X) = 1\}$.

In the first case, we exhibit a tropical Kraus map which is locally expansive at its only non-linear eigenvector. In the second example, the tropical Kraus map does not have a unique eigenvector.

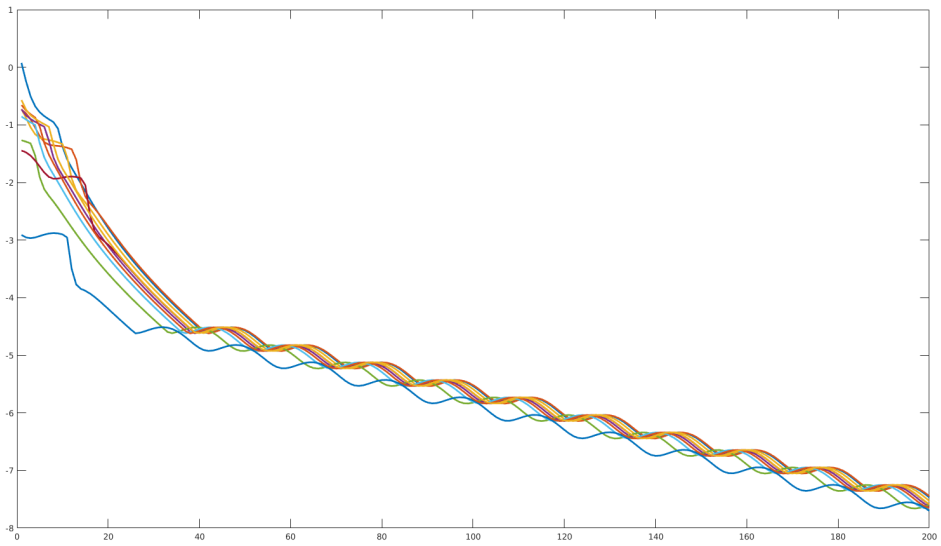


Figure 7.2: The two regimes of convergence

7.2.5.a Locally expansive tropical Kraus maps We consider the tropical Kraus map T_1 with

$$A = \begin{pmatrix} -0.4285 & -0.2825 \\ 0.0400 & 0.4006 \end{pmatrix} \qquad B = \begin{pmatrix} -2.1017 & 0.4002 \\ -0.6605 & -1.3176 \end{pmatrix}.$$

This map is strictly expansive in the Riemann metric on an open subset of Δ as shown in Figure 7.1a. It can be observed on Figure 7.1b that this does not impair the convergence of the iterative scheme. Moreover, the limit point of all sequences is actually inside the "expansivity set": the directional derivative at the limit point takes values between 0.94 and 1.03. Figure 7.1 shows in more detail the position of this limit point.

We observe several regimes in the convergence towards the fixed point. Although the iterations lead to a neighborhood of the fixed point quite fast, a much slower regime is observed within this neighborhood. This change of regime is shown in Figure 7.2 where we have plotted the (logarithm of the) distance to the fixed point in Riemann's metric with respect to the number of iterations: 40 iterations are sufficient to reach an error of 10^{-4}, yet 4 times more are required to reach an error of 10^{-8}.

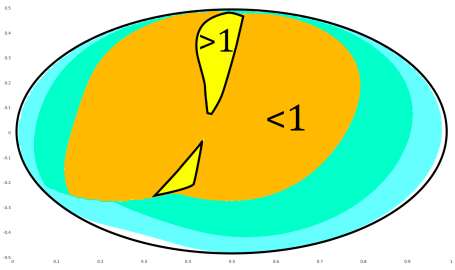
This example illustrates that, although the tropical Kraus map is expansive with respect to the Riemann metric near its unique fixed point, it is still possible that this map is a contraction with respect to other metrics. Indeed, on the set Δ the map $\widehat{T_1}$ can be identified to a map from the unit disk in $\mathbb{R}^2$ to $\mathbb{R}^2$ by

$$\iota \colon \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right) \mapsto \left(\begin{smallmatrix} x & y \\ y & 1-x \end{smallmatrix}\right),$$

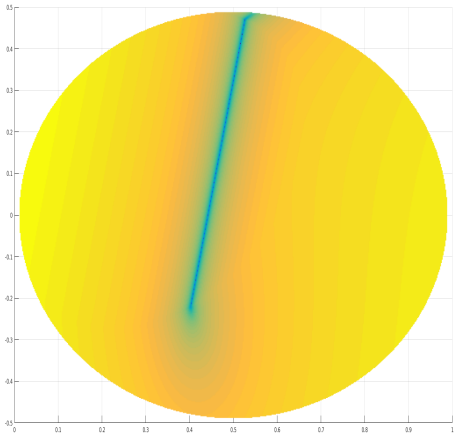
and the differential of the map $\iota^{-1} \circ \widehat{T_1} \circ \iota$ at this fixed point is given by the matrix

$$\left(\begin{smallmatrix} 1.0177 & -0.2809 \\ 0.1355 & 0.9093 \end{smallmatrix}\right)$$

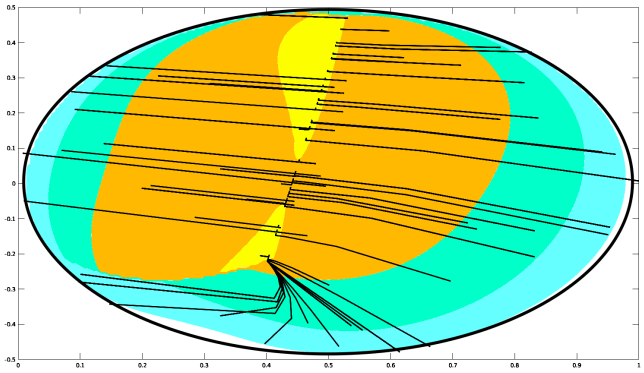
whose eigenvalues are $0.9635 \pm 0.1874i$. Their modulus is $0.9816 < 1$, hence the fixed point is attractive.



(a) Value of the local Lipschitz constant



(b) Plot of the map $X \mapsto d_R\big(X, \widehat{T}(X)\big)$ showing the position of the fixed points



(c) Trajectories converging towards the fixed points

Figure 7.3: Lipschitz constant, fixed point and trajectory of iterated tropical Kraus maps on the slice Δ

7.2.5.b Multiplicity of eigenvectors We consider the tropical Kraus map T_2 with

$$A = \begin{pmatrix} -0.4285 & -0.2825 \\ 0.0400 & 0.4006 \end{pmatrix} \qquad B = \begin{pmatrix} -2.1017 & 0.4002 \\ -0.6605 & -1.3176 \end{pmatrix}.$$

This map has infinitely many nonlinear eigenvectors, and the Krasnoselskii-Mann iterations do not converge towards a single eigenvector. All eigenvectors are located in the interior of the cone, associated with the same non-linear eigenvalue: one can check numerically that the matrices $X(s)$ yield such eigenvectors for the eigenvalue 1.9428 with

$$X(s) = \begin{pmatrix} s & 5.5656s - 2.4548 \\ 5.5656s - 2.4548 & 1 - s \end{pmatrix} \qquad 0.41 \leqslant s \leqslant 0.52\,.$$

These matrices satisfy $AX(s)A^T \preccurlyeq BX(s)B^T$, so $T_2\big[X(s)\big] = BX(s)B^T$. This entails that the joint spectral radius $\rho = \rho(\{A_2, B_2\})$ is the spectral radius of B and that $X(s)$ is a

common quadratic Lyapunov certificate: $AXA^T, BXB^T \preccurlyeq \rho^2 X$. In this example, the maps T and $X \mapsto BXB^T$ coincide on a proper subset of Δ with non-empty interior. We deduce in particular that the map T is non-expansive on this subset, which contains the former matrices.

7.3 Perturbations methods to ensure convergence

We show in this section that the perturbation methods introduced in Section 6.4.1 converge provided their parameter ε is large enough. We deal with the additive approach in Section 7.3.1 and with the multiplicative perturbation in Section 7.3.2.

We consider the tropical Kraus map T defined on $\mathcal{S}_n^+$ by

$$T(X) \coloneqq \bigsqcup_{1 \leqslant i \leqslant p} A_i X A_i^T .$$

We define the *projective Riemann metric* between $X, Y \in \Delta$ as the infimum of the associated distance between elements on the rays passing through X and Y by:

$$d_R^p(X, Y) \coloneqq \inf_{\lambda,\mu>0} d_R(\lambda X, \mu Y) .$$

Since the Riemann metric in invariant by multiplication of its arguments by a positive scalar, one can choose $\mu = \lambda^{-1}$ in the minimization above.

We can define the "projective Thompson metric d_T^p" in the same way. By Equation (4.18), this projective metric coincides with the (double of the) Hilbert metric d_H.

A key remark in the following is that when the invariant join is evaluated sequentially, the Lipschitz constant of the map T in the Hilbert metric and in the projective Riemann metric are finite.

Corollary 7.11. *They are respectively bounded by* $\mathrm{Lip}_H[T]$ *and* $\mathrm{Lip}_R^p[T]$ *with*

$$\log(n)/\pi - 1 \leqslant \mathrm{Lip}_H[T] \leqslant (2 + \frac{4}{\pi} \log n)^{p-1} \qquad 1 \leqslant \mathrm{Lip}_R^p[T] \leqslant \sqrt{p} . \tag{7.7}$$

Proof. These bounds are obtained as corollaries of Theorems 4.1 and 4.3. By Theorem 4.1, the invariant join is Lipschitz with respect to the product Riemann metric on $\mathcal{S}_n^{++} \times \mathcal{S}_n^{++}$, i.e.

$$d_R(X_1 \sqcup X_2, Y_1 \sqcup Y_2) \leqslant \left[d_R(X_1, Y_1)^2 + d_R(X_2, Y_2)^2 \right]^{1/2} .$$

When the invariant join is evaluated sequentially, the invariance of the Riemann metric by congruences implies that

$$d_R\big(\bigsqcup_{1 \leqslant i \leqslant p} A_i X A_i^T, \bigsqcup_{1 \leqslant i \leqslant p} A_i Y A_i^T \big) \leqslant \sqrt{p}\, d_R(X, Y) .$$

We deduce that the Lipschitz constant of the operator T in the Riemann metric does not exceed $\sqrt{p}$. Similarly, the Lipschitz constant in Hilbert's metric is bounded by $(2+\frac{4}{\pi}\log n)^{p-1}$ as a consequence of Theorem 4.3. □

The bound on $\mathrm{Lip}_H[T]$ is very conservative, especially if $p > 2$, but we do expect it to depend both on n and p. This result is mostly of theoretical interest. Observe that the bound on $\mathrm{Lip}_R^p[T]$ is independent of the dimension n.

7.3.1 Additive damping approach

We first consider an additive variant of the tropical Kraus map T, like in Section 6.4, where we introduce a small positive perturbation ε. Let g_ε denote the map $g_\varepsilon \colon X \mapsto X + \varepsilon \operatorname{trace}(X) I_n$. Then the variant T^ε is defined by $T^\varepsilon(X) \coloneqq g_\varepsilon \circ T(X)$:

$$T^\varepsilon(X) \coloneqq T(X) + \varepsilon \operatorname{trace}\Big[T(X)\Big] I_n \,. \tag{7.8}$$

This leads to a new power iteration

$$X^{k+1} = \frac{1}{\sum_w \langle I_n, T^\varepsilon_w(X^k)\rangle} T^\varepsilon(X^k) \,. \tag{7.9}$$

The map g_ε is linear and maps the cone $\mathcal{S}_n^+$ into itself. By the Birkhoff-Hopf contraction theorem, the map g_ε is a contraction if the cone $g_\varepsilon(\mathcal{S}_n^+)$ has a finite diameter in Hilbert's metric:

Theorem 7.12 (Birkhoff-Hopf, see [Bir57, Nus87])**.** *A linear map f such that $f(\mathcal{S}_n^+) \subseteq \mathcal{S}_n^+$ is a contraction in* Hilbert's metric *if and only if the image of the cone $\mathcal{S}_n^+$ has a finite diameter δ in this metric. Its contraction coefficient is bounded by* $\operatorname{Lip}_H f \leqslant \tanh(\delta/4)$.

We show that the diameter of $g_\varepsilon(\mathcal{S}_n^+)$ in Hilbert's metric is finite as soon as $\varepsilon > 0$ and deduce the contraction rate of the map g_ε in this metric:

Lemma 7.13. *For all $x, y \in \mathcal{S}_n^+$, we have $d_H\big[g_\varepsilon(x), g_\varepsilon(y)\big] \leqslant 2 \log \frac{1+\varepsilon}{\varepsilon}$. Hence g_ε is a contraction in the Hilbert metric:*

$$d_H\big[g_\varepsilon(x), g_\varepsilon(y)\big] \leqslant \frac{1}{1+2\varepsilon} d_H(x, y) \,.$$

Proof. It is sufficient to bound the distance between two rank-one 2×2 matrices. Without loss of generality, we choose $x = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and $y = \left(\begin{smallmatrix} \cos^2(t) & \cos(t)\sin(t) \\ \cos(t)\sin(t) & \sin^2(t) \end{smallmatrix}\right)$. We have

$$\frac{(1+\varepsilon)^2}{\varepsilon} g_\varepsilon(y) - g_\varepsilon(x) = \begin{pmatrix} \frac{1+\varepsilon}{\varepsilon} - \varepsilon - \cos^2(t) & \sin(t)\cos(t) \\ \sin(t)\cos(t) & 1 - \sin^2(t) \end{pmatrix} .$$

The inequality $(1+\varepsilon)^2/\varepsilon - \varepsilon \geqslant 1$ holds trivially, hence $\frac{(1+\varepsilon)^2}{\varepsilon} g_\varepsilon(y) - g_\varepsilon(x) \succcurlyeq 0$. The same inequality holds if x and y are swapped. We deduce that

$$d_H\big[g_\varepsilon(x), g_\varepsilon(y)\big] = \inf\big\{ \log(\mu/\lambda) \colon \lambda^{-1} g_\varepsilon(y) \preccurlyeq g_\varepsilon(y) \preccurlyeq \mu g_\varepsilon(x) \big\} \leqslant 2 \log \frac{1+\varepsilon}{\varepsilon} \,.$$

Using the formula $\tanh(u) = \big(1 - \exp(-2u)\big)\big(1 + \exp(-2u)\big)^{-1}$ and Theorem 7.12, we get $\operatorname{Lip}_H g_\varepsilon \leqslant \frac{1}{1+2\varepsilon}$. □

Hence the modified iteration scheme using T^ε converges if ε is large enough:

Theorem 7.14. *The modified iteration starting at $X^0 \in \mathcal{S}_n^+$ defined by*

$$X^{k+1} = \frac{T(X^k) + \varepsilon \operatorname{trace}(X^k) I_n}{\operatorname{trace}\big[T(X^k) + \varepsilon \operatorname{trace}(X^k) I_n\big]} \,,$$

converges if $\varepsilon > \min\big[(n\sqrt{p} - 1)/2 \,,\, (2 + 4\log(n)/\pi)^{p-1}/2 - 1/2)\big]$.

Proof. The Hilbert metric is equal to the double of the projective Thompson metric. Moreover, following classical norm comparison results, we have

$$d_T(x,y) \leqslant d_R(x,y) \leqslant n\, d_T(x,y)\,.$$

Hence the Lipschitz constant of a map in the projective Riemann metric does not exceed n times its Lipschitz constant in the Hilbert metric. Combined with Lemma 7.13, we obtain that g_ε is $n/(1+2\varepsilon)$-Lipschitz in the projective Riemann metric. By Equation (7.7), the map T is at most $\sqrt{p}$-Lipschitz. Thus the iteration is contractive in the projective Riemann metric if $n\sqrt{p} < 1+2\varepsilon$. The reasoning for Hilbert's metric is more direct, since we only combine Lemma 7.13 and Theorem 4.3. □

Although the bound in the Hilbert metric is very coarse, when $p=2$ (2 switching modes), it provides more reasonable estimates for moderate values of n: $\varepsilon > (1+4\log(n)/\pi)/2$. When $p>2$, it is preferable to consider the estimation from the projective Riemann metric.

We obtain as a corollary of Theorem 7.14 that the iterative schemes in Equation (6.14) and Equation (7.9) converge for a larger enough ε:

Corollary 7.15. *The iterative scheme in Equation* (7.9) *starting at* $(X_w^0)_w \in (\mathcal{S}_n^+)^{\mathcal{W}}$ *converges if* $\varepsilon > (n\sqrt{p}|\mathcal{W}|^{3/2}-1)/2$.

Proof. In the lifting procedure described in Section 7.1.1, the dimension of the matrices grows from n to $n|\mathcal{W}|$ and the number of matrices in the collection $\mathcal{A}$ grows from p to $p|\mathcal{W}|$. □

These estimates on ε are very conservative: in practice, we use values of ε less than 10^{-2}. This is due to the coarse estimation of the contraction rate in Riemann's projective metric of the map g_ε in the proof of Theorem 7.14, which seems to indicates that the map g_ε is a contraction in this metric only if ε is large. We observe experimentally that this is not true, and we conjecture that it is a contraction as soon as $\varepsilon > 0$:

Conjecture 7.16. *The map* g_ε *defined on* $\mathcal{S}_n^+$ *is a contraction in the Riemann projective metric:*

$$d_R^p\big[g_\varepsilon(X), g_\varepsilon(Y)\big] \leqslant \frac{1}{1+\alpha_n\varepsilon} d_R^p(X,Y) \quad \textit{for some } \alpha_n \geqslant n^{1/2}\,.$$

Let us give some elements supporting the conjecture. Locally, the Riemann projective metric is the standard deviation σ of the spectrum of the modified matrix $X^{-1}H$: $d_R^p(X, X+H) = \inf_{\mu\in\mathbb{R}} \|X^{-1}H - \mu I_n\|_F = \sigma\big[\operatorname{Sp} X^{-1}H\big]$. When $\varepsilon \ll 1$, we linearize to get

$$\big[g_\varepsilon(X)\big]^{-1} g_\varepsilon(H) = X^{-1}H + \varepsilon\big(\operatorname{trace}(H)X^{-1} - \operatorname{trace}(X)HX^{-2}\big)$$

When $X = I_n$, we have $\big[g_\varepsilon(X)\big]^{-1} g_\varepsilon(H) = (1-n\varepsilon)H + \varepsilon \operatorname{trace}(H) I_n$. Classical formulas for standard variation give a local contraction rate of $1-n\varepsilon \sim (1+n\varepsilon)^{-1}$. This shows that $\alpha_n \leqslant n$. However, equality does not hold. Indeed, one can check that $\alpha_n < n$, using the example

$$X = \begin{pmatrix} 7.2776 & -0.3214 & 0.2509 \\ -0.3214 & 6.3741 & -0.0739 \\ 0.2509 & -0.0739 & 3.7221 \end{pmatrix} \qquad H = \begin{pmatrix} 4.8879 & -3.1582 & 2.1688 \\ -3.1582 & -6.0836 & -1.6669 \\ 2.1688 & -1.6669 & -1.0169 \end{pmatrix}.$$

We deduce, conditionally to Conjecture 7.16, a better estimation on ε:

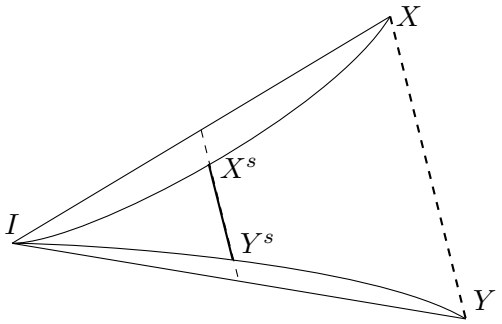


Figure 7.4: The non-positive curvature property of Thompson's metric on the space of positive definite matrices.

Theorem 7.17. *Assuming that Conjecture 7.16 holds, the modified multiplicative iteration defined on* $\mathcal{S}_n^+$ *by*

$$X^{k+1} = \frac{1}{\operatorname{trace}\left[T^\varepsilon(X^k)\right]} T^\varepsilon(X^k)\,,$$

converges if $\varepsilon > (\sqrt{p} - 1)/n$.

Consequently, the iterations defined in Equation (6.14) *converge if* $\varepsilon \geqslant p^{1/2} n^{-1}$.

7.3.2 Convergence analysis of the multiplicative power iteration

We now consider a multiplicative variant of the tropical Kraus map T as used in Equation (6.15) for $0 \leqslant \varepsilon < 1$ by

$$T^\varepsilon(X) \coloneqq \left[T(X)\right]^{1-\varepsilon}$$

The reason for considering the multiplicative power iteration is that, when Y is positive definite and $0 < s < 1$, the map $Y \mapsto Y^s$ is a contraction with respect to Thompson's (part) metric d_T [Nus88] and Riemann's metric d_R [Bha07, Theorem 6.1.12].

It is known (*ibid.*) that a geodesic for Thompson's and Riemann's metric linking I and a positive matrix X is given by the curve sending $t \in [0,1]$ to $t \mapsto X^t$. By geodesic, we mean that the equality holds in the triangular inequality $d_m(I, X^t) \leqslant d_m(I, X^s) + d_m(X^s, X^t)$ for all $0 < s < t$ and $m \in \{T, R\}$.

The geodesics between I and two positive matrices X and Y have the following property, which is known in metric geometry as *non-positive curvature in the sense of Busemann* for the Thompson metric,

$$d_T(X^t, Y^t) \leqslant t\, d_T(X, Y)\,. \tag{7.10}$$

This can be deduced either from [Bha03] or from classical log-majorization inequalities for matrix eigenvalues [Zha02], see [GV12] for details. This inequality means that the triangles are thin, it is illustrated Figure 7.4. We warn the reader that non positive curvature in the sense of Busemann is a milder condition than other non-positive curvature conditions more commonly used like being CAT(0), see [Pap05] for background. The same inequality holds for the Riemann metric $d_R(X^t, Y^t) \leqslant t\, d_R(X, Y)$. In this setting, the underlying notion of non-positive curvature is CAT(0), see [Bha07, Theorem 6.1.9].

Finally, the same inequality holds for the projective Riemann metric and Hilbert's metric. For the projective Riemann metric, if $s \neq 0$, we have

$$d_R^p(X^s, Y^s) \leqslant d_R\big[\lambda X^s, \mu Y^s\big] \leqslant s\, d_R[\lambda^{1/s} X, \mu^{1/s} Y]\,,$$

thus by the change of variable $x' \coloneqq x^{1/s}$, taking the infimum over λ', μ' yields the inequality

$$d_R^p(X^s, Y^s) \leqslant s\, d_R^p(X, Y)\,. \tag{7.11}$$

The proof in Hilbert's metric is done identically.

The next result shows that the power algorithm does converge for a large enough ε. In the present experiments, we use a much smaller value of ε.

Theorem 7.18. *The modified iteration starting at $X^0 \in \mathcal{S}_n^+$ defined by*

$$X^{k+1} = \frac{\big[T(X)\big]^{1-\varepsilon}}{\operatorname{trace}\Big[\big[T(X)\big]^{1-\varepsilon}\Big]}$$

converges if $\varepsilon > 1 - 1\sqrt{p}$.

Remark 7.3. The limit X_∞ can be approximated with an accuracy η in

$$p^* \coloneqq \Big\lceil \frac{\log \eta - \log d_m^p(X_0, X_\infty)}{\log(1-\varepsilon) + \log(\sqrt{p})} \Big\rceil\,,$$

iterations, leading to $O(n^3 N p^*)$ arithmetic operations.

7.4 Experimental results

7.4.1 Implementation issues

We describe in this section the resolution to several issues that arise in the implementation of the iterative scheme.

In practice, we use values for ε in the range $10^{-4} - 10^{-2}$ in Equation (7.8). This additional parameter allows us to obtain, in a finite number of iterations, a solution (X, ρ) that satisfies $\rho^2 X_j \succcurlyeq A_\sigma^T X_i A_\sigma$ for all admissible (i, σ, j). Moreover, this parameter absorbs numerical imprecisions that may appear during the computation and ensures that the matrices X_j are positive definite, so the assumptions of Theorem 7.1 and Theorem 7.2 are satisfied.

We choose $\mho$ to be the trace minimizing selection denoted by $\sqcup_{\mathrm{tr}}$ in the computations. Then, when the set Σ contains only two elements, $T_j(X)$ can be computed analytically thanks to Theorem 2.3. When Σ has more than two elements, instead of computing the true minimal upper bound $\sqcup_{\mathrm{tr}} \mathcal{Q}$, we compute an approximation by sequential evaluation: $Q_1 \sqcup_{\mathrm{tr}} (Q_2 \sqcup_{\mathrm{tr}} (\cdots \sqcup_{\mathrm{tr}} Q_p))$. Although the original evaluation can be performed by solving a semidefinite program of a reasonable size, we observe that the sequential evaluation still converges and produces very good eigenvectors.

Finally, as pointed out in [GMQ11], the propagation of the Riccati operator $\mathrm{ricc}_{\tau,\sigma}$ on a single quadratic form P_0 is computed analytically by $\mathrm{ricc}_{\tau,\sigma} = Y(\tau)X(\tau)^{-1}$, with $\mathcal{M}^\sigma = \left(\begin{smallmatrix} -A^\sigma & -Q^\sigma \\ D^\sigma & A^\sigma \end{smallmatrix}\right)$ and $\big(X(\tau)\,;\, Y(\tau)\big)^T = \exp\big(\mathcal{M}^\sigma \tau\big)\big(I_n\,;\, P_0\big)^T$.

7.4.2 Application to the joint spectral radius

Given that the approximation of the joint spectral radius $\rho(\mathcal{A})$ depends on the graph $\mathcal{G}$ that underlies the analysis, we denote by $\widehat{\rho}(\mathcal{A}, \mathcal{G})$ the approximation obtained as (the square root of) an eigenvalue of a tropical Kraus map and by $\rho(\mathcal{A}, \mathcal{G})$ the one obtained by solving the LMI (7.6).

The map $\cdot$ sending $\mathcal{W} \times \Sigma$ to $\mathcal{W}$ defined in Section 7.1.1 can be interpreted as a path-complete graph. For this reason, our method, when applied to the joint spectral radius, is a relaxation of the path-complete Lyapunov function framework, and thus we always have

$$\rho(\mathcal{A}) \leqslant \rho(\mathcal{A}, \mathcal{G}) \leqslant \widehat{\rho}(\mathcal{A}, \mathcal{G}) \,.$$

However, we shall see that the tropical method is much more tractable, so we may use a bigger graph and sometimes get a better approximation than by solving LMIs, for a similar time or computational budget.

We compare the performance of our algorithm with the path-complete graph Lyapunov method, in terms of computation time and accuracy of the approximation of the joint spectral radius, measured by

$$\delta \coloneqq (\widehat{\rho}(\mathcal{A}, \mathcal{G}') - \rho(\mathcal{A}, \mathcal{G})/\rho(\mathcal{A}, \mathcal{G}) \,.$$

All the experiments were implemented in Matlab, running on one core of a 2.2 GHz Intel Core i7 with 8 GB RAM. The semidefinite programs were solved using YALMIP (R20160930), calling SeDuMi 1.3.

7.4.2.a Accuracy of the approximation We generate 600 pairs $\mathcal{A} = \{A_1, A_2\}$ of random 6×6 matrices. For each of these pairs, we compare the approximation of the joint spectral radius obtained by the LMI method on the graph D_3 (involving 8 positive semidefinite matrices) and by the tropical Kraus method on the graph D_6 (involving 64 positive semidefinite matrices). On these examples, we report that the tropical method obtains a similar approximation of the joint spectral radius, within a margin of 2.5%, and outperforms the LMI-method on 25% of these examples. Moreover, whereas the LMI-method requires between 3s and 5s to obtain this approximation, the tropical method consistently returns an approximation in 1s.

7.4.2.b Scalability - dimension We generate random pairs of $n \times n$ matrices, for n ranging from 5 to 500. We use again the De Bruijn graph D_3 in the LMI-method and the graph D_6 in the tropical Kraus method. We show in Table 7.1 the mean computation time required to obtain an overapproximation and the mean relative accuracy of the tropical method with respect to the LMI-method, when it applies. First, one can observe the major speedup provided by the tropical method, from 4 times faster when $n = 5$ to 80 times faster for $n = 40$.

Also note that the tropical method is using 8 times more quadratic forms in its analysis and remains much faster than the LMI-method. Thus, given a fixed time budget, the tropical method enjoys more flexibility regarding the size of the graph that is used in the analysis.

Moreover, observe that the LMI-method cannot provide estimates on the joint spectral radius for values of n greater than 45, whereas the tropical method easily reaches values of n greater than 100.

Finally, the accuracy of the tropical approximation remains within a 1.5% margin of the one obtained by the LMI-method.

Table 7.1: Comparison of the methods with respect to the size of the dimension of the matrices.

Dimension n	CPU time (tropical)	CPU time (LMI)	Upper bound on $\rho(\mathcal{A})$ (tropical)	Upper bound on $\rho(\mathcal{A})$ (LMI)	Accuracy
5	0.9 s	3.1 s	2.767	2.7627	0.1 %
10	1.5 s	4.2 s	3.797	3.7426	1.4 %
20	3.5 s	31 s	5.4093	5.3891	0.4 %
30	7.9 s	3min	6.2038	6.1942	0.2 %
40	13.7 s	18min	7.3402	7.3363	0.05 %
45	18.1 s	–	7.687	–	–
50	25.2 s	–	8.1591	–	–
100	1min	–	11.487	–	–
500	8min	–	25.44	–	–

7.4.2.c Scalability - graphs We now analyze the influence of the order of the De Bruijn graph D_d used in the analysis on the computation of the upper bound on the joint spectral radius obtained by both methods. We use the matrices $A_1 = \begin{pmatrix} -1 & 1 & -1 \\ -1 & -1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$ and $A_2 = \begin{pmatrix} -1 & 1 & -1 \\ -1 & -1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$, introduced in [GZ14]. Their joint spectral radius is $\rho(\mathcal{A}) = 1.78893$.

We show in Table 7.2 the upper bound on the joint spectral radius and the computation time with respect to the length order d of the De Bruijn graphs.

7.4.3 A faster curse of dimensionality attenuation scheme

We now apply the iteration scheme described in Section 7.1.4 to the approximation of the value function V. In all examples, we measure the quality of the approximation of the value function as in [McE09, GMQ11] with the H-infinity back-substitution error $\max_{x^T x \leqslant 1} |H(x, \nabla V(x))|$ on the subspace spanned by the canonical vectors e_1 and e_2.

The first example is Example 1 in [McE07] and we use the instance of [GMQ11] in the second example. Examples 3 and 4 are randomly generated examples that satisfy the technical assumptions in [McE07].

Table 7.3 depicts the results of the computations. In particular, we give the back-substitution error at the beginning of the computation, when the value function is approximated by a single quadratic form ($Q(x) = 0.1|x|^2$ in all cases) and the final back-substitution error when the scheme has converged.

Table 7.2: Comparison of the methods w.r.t. the size of the graph D_d.

Order d	2	4	6	8	10
Size of $\mathcal{W}$	8	32	128	512	2048
CPU time (tropical)	0.03s	0.07s	0.4s	2.0s	9.0s
CPU time (LMI)	1.9s	4.0s	24s	1min	10min
Upper bound on $\rho(\mathcal{A})$ (tropical)	1.842	1.821	1.804	1.800	1.801
Upper bound on $\rho(\mathcal{A})$ (LMI)	1.8216	1.7974	1.7957	1.7922	1.7905
Accuracy	1.1 %	1.3 %	0.4 %	0.4 %	0.6 %

Table 7.3: Numerical benchmarks of the tropical Kraus method applied to McEneaney's switched linear quadratic problem

Example	1	2	2	3	4
Dimension	2	6	6	20	20
Size of Σ	3	6	6	2	4
τ	0.05 s	0.2	0.1	0.1	0.1
Size of $\mathcal{W}$	81	216	1296	128	256
Initial error	0.78	1.12	1.12	4.2	4.79
Final error	0.047	0.071	0.090	0.0006	0.17
Iterations	194	115	200	55	288
CPU time	8 s	41 s	5 min	5 s	2.5 min

CHAPTER 8

Implementations of the algorithms

8.1 Presentation of MEGA

MEGA (short for *Minimal Ellipsoids Geometric Analyzer*) is a static analyzer for linear and affine programs implementing the algorithms introduced in Chapters 6 and 7. The software weighs approximately 1600 lines of OCaml and 800 lines of Matlab and consist in several parts.

First, a parser reads a text file containing the program to be analyzed and transforms it into an internal representation. This is a common step regardless of the linear or affine character of the program. Then the analyzes diverge:

- For linear programs, the algorithm in Chapter 7 has been implemented using the LaCaml library [Mot16] to perform linear algebra operations in OCaml. We have also implemented the method of [AJPR14] for comparison purposes, using the OSDP library [GR17]. Optionally, we produce a piece of Matlab code that instantiates the two former algorithm on the desired program for further analysis within Matlab. In both cases, we return an upper bound on the joint spectral radius, the computation time and the quadratic forms whose supremum produces the approximate Barabanov norm. In the affine case, we also return a png image of the invariant and the image of the invariant by the abstract operators corresponding to each branch.

- For affine programs, we directly instantiate[1] a piece of Matlab code that is then executed to produce the collection of ellipsoids whose reunion produces an invariant as in Chapter 6. We also implement the big-BMI method for comparison purposes.

[1] As of October 2017, the pure OCaml implementation using OSDP is not stable, thus the detour through Matlab is necessary.

The programs that are fed as input are written in a C-style form, with minor adaptations for non-deterministic assignments and parallel assignments. Two typical programs are shown in Programs 12 and 13 and exhibit this syntax.

The non-deterministic assignment of a variable x to a value in the interval $[a, b]$ is written $x \leftarrow [a, b]$. The assignment operator $\coloneqq$ works as follows. If the assignment is part of a "if-then" or "switch-case" structure, it is considered a parallel vector assignment, not a sequential assignment. Otherwise, it acts like a standard assignment operator, and it is usually used to set constants for the program.

A program is comprised of *variables*, *floats*, *expressions* and *statements*. Expressions are recursively defined as variables, floating point numbers as constants and sum and products of expressions. In other words:

$$\text{type } expr \;=\; \text{Variable } v \;\mid\; \text{Float } f \;\mid\; expr + expr \;\mid\; expr * expr\,.$$

A statement is recursively defined as an deterministic or non-deterministic assignment of an expression to a variable, two statements separated by a semi-colon, an if-then-else structure discriminating between the execution of two consecutive statements based on the conjunction of several expression comparisons, a switch-case between several statements based on the value of a variable or a looping process based on the comparison of two expressions. In other words:

$$\begin{aligned}\text{type } stm \;=\; &\text{Variable } v \coloneqq expr \;\mid\; \text{Variable } v \leftarrow [\text{Float } low, \text{Float } up] \;\mid\\ &stm\,;\, stm \;\mid\; \text{if } expr \leqslant expr \text{ then } stm \text{ else } stm \text{ end} \;\mid\\ &\text{switch(Variable } u\text{): case i : } stm_i \text{ end} \;\mid\; \text{while } expr \leqslant expr \text{ do } stm \text{ end}\end{aligned}$$

We show in Programs 12 and 13 two illustrative programs written in this syntax.

Program 12: Non-deterministic switched linear program

```
x_1 ← [−1, 1];
x_2 ← [−1, 1];
h := 0.02; p := 1; q := 5; r := 0.1; s := 10;
while 0 ⩽ 1 do
    switch u :
        case 0 :
            x := x + p ∗ h ∗ v;
            v := (−1) ∗ p ∗ h ∗ x + (1 + (−1) ∗ p ∗ h) ∗ v;
        end
        case 1 :
            x := x + q ∗ h ∗ v;
            v := (−1) ∗ r ∗ r ∗ q ∗ h ∗ x + (1 + (−1) ∗ r ∗ q ∗ h) ∗ v;
        end
        case 2 :
            x := x + s ∗ h ∗ v;
            v := (−1) ∗ r ∗ s ∗ h ∗ x + (1 + (−1) ∗ r ∗ s ∗ h) ∗ v;
        end
end
```

Program 13: Switched affine program with guards

$x \leftarrow [-0.2, 0.2]$;
$y \leftarrow [-0.2, 0.2]$;
while $0 \leqslant 1$ **do**
  if $y \geqslant 0$ **then**
    $x := 0.6835 * x + (-0.7468) * y + 0.5432$;
    $y := 0.1867 * x + 0.6835 * y + (-0.0724)$;
  else
    $x := 0.7185 * x + (-0.3925) * y$;
    $y := 0.3925 * x + 0.7185 * y$;
  end
end

8.2 Using MEGA

We give some information on how to use MEGA. A compressed file containing its source files is available at `www.cmap.polytechnique.fr/~stott/mega` with instructions on how to install the required dependencies. The executable is *mega* and can take several arguments:

-d int	Int is the depth of the De Bruijn automaton
-i "??.txt"	Name of the input text file containing the program
-o "??.txt"	Name of the ouput file for the Matlab code
-cmp	Activate the comparison with the big-lmi/big-bmi methods
-q	Moves Matlab output from terminal to LOG file
-eps float	Float is the precision on the joint spectral radius upper bound
-rand int	Randomly generated linear program: 2 modes in dimension int

The hosted commpressed file contains the two test programs of Programs 12 and 13. Here are some executions of mega and the obtained results.

```
$ ./mega -i "bench2.txt" -d 3
   Total Time:                    74.42 s
```

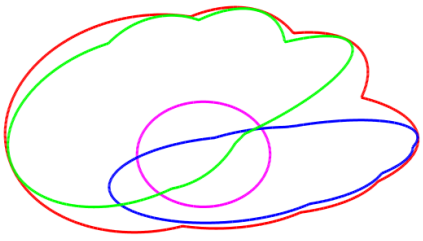


Figure 8.1: MEGA on example in Program 12

```
$ ./mega -i "bench1.txt" -d 5 -cmp
  Depth of the automaton: 5
  Abstraction Time:            6.70e-05 s
  Linear program structure: true

  Finished in 84 iterations.
  Tropical computation Time:  1.83 s
  Upper bound on the jsr: 1.021
  Computation result is sound: true

  SDP computation Time:        2.69 s
  Upper bound on the jsr: 1.013

  Precision: 0.75%
  Speed-up : 1.47
```

Figure 8.2: MEGA on example in Program 12

```
$ ./mega -rand 10 -d 3 -cmp
  Depth of the automaton: 3
  Abstraction Time:            1.33e-04 s
  Random instance of linear switched program

  Finished in 27 iterations.
  Tropical computation Time:  0.12 s
  Upper bound on the jsr: 4.434
  Computation result is sound: true

  SDP computation Time:        16.29 s
  Upper bound on the jsr: 4.431

  Precision: 0.08%
  Speed-up : 135.88
```

Figure 8.3: MEGA on a randomly generated example

CHAPTER 9

Conclusion and perspectives

We now briefly summarize our contributions and point out some open problems.

We have established in Chapters 1 to 3 several theoretical results on minimal upper bound selections in the cone of positive semidefinite matrices and in more general cones. We have shown that such a minimal upper bound can be selected by means of a vector belonging to the interior of the dual cone. Moreover, we have extended the definition of the Löwner ellipsoid to more general cone as the unique minimal upper that selects itself (via a canonical involution between a cone and its dual). There is still much to be understood regarding minimal upper bound selections, even in the special case of positive semidefinite matrices. Indeed, very few properties of the selection map $c \mapsto \Phi(\mathcal{A}, c)$ defined in Section 3.1.3 are known. Our main conjecture regarding the latter map (Conjecture 3.7) states that it is contractive with respect to the Thompson metric when restricted to positive definite matrices.

A key object in the study of minimal upper bounds is the notion of tangency faces or tangency subspaces. We expect this notion should also play a major role in the study minimal upper bounds of geometric objects that do not derive from cones, like zonotopes [GPV12] and uncentered ellipsoids. In the latter case, we point out that it is not possible to reduce to the diagonal case by simultaneous reduction under congruences, because the underlying order relation is defined on *indefinite matrices* and relies on the S-Lemma [BTN01]:

$$\mathcal{E}(I,0) \subseteq \mathcal{E}(Q,q) \iff \exists \lambda > 0 \colon \begin{pmatrix} I & 0_{n,1} \\ 0_{1,n} & -1 \end{pmatrix} \preccurlyeq \lambda \begin{pmatrix} Q - qq^T & -q \\ -q^T & -1 \end{pmatrix}.$$

We believe that some aspects of the analysis may be preserved by considering anti-Jordan reduction results [Uhl76, GLO05].

We have computed in Chapter 4 lower and upper Lipschitz bounds for the invariant selection on positive semidefinite matrices in the Thompson and Hilbert metrics, as well as shown that it is non-expansive in the product space for the Riemann metric when dealing

with 2 matrices. We have conjectured that this property holds regardless of the number of arguments of the invariant join. Moreover, since we have used other selections in practical applications, it is desirable to be able to obtain estimations for other selections that the invariant join.

We have introduced in Chapters 6 and 7 a new method to approximate the value function of optimal control problems for switched systems. This includes the computation of quadratic invariants of switched affine systems, the computation of the joint spectral radius and a class of linear quadratic control problems with switches considered by McEneaney. In the first case, it enabled the study of affine systems that include guards and external inputs. This was not possible with earlier methods based on the solution of a large semidefinite program and required instead the solution of a non-convex bilinear matrix inequality. The benefit of our approach is the major gain in scalability, but it comes at the cost of a loss in precision: since we replace LMI/BMI formulations by a specific selection of minimal upper bounds of ellipsoids, we induce a relaxation gap. In other words, the scheme currently allows one to compute quickly a coarse approximation of an invariant set.

In the setting of the joint spectral radius and McEneaney's linear quadratic control problem, we have introduced the notion of *tropical Kraus maps* as a tool to approximate the value function of an optimal control problem. In this case, our scheme belongs to the family of max-plus methods as it approximates the value function by a supremum of quadratic forms. It completely avoids the recourse to semidefinite programming (which was the bottleneck of earlier max-plus methods) by a reduction to a non-linear eigenproblem. This leads to a major speedup, allowing us to obtain approximate solutions of instances in dimension up to 500 in the case of the joint spectral radius, and 20 for McEneaney's problem, hardly accessible by other methods. There is again a trade-off in precision due to the specific selection process. In other words, the scheme currently allows one to compute quickly a coarse approximation of the solution of a Hamilton-Jacobi PDE.

The most promising improvement of the scheme may be to adapt dynamically the selection of a minimal upper bound, which will reduce the relaxation gap, and might also improve the convergence. This may be achieved by adding the dual tangency variables Λ into the computation and deriving an update mechanism for these variables from the optimality conditions in the "big-LMI". These dual variables can then be used to select a good minimal upper bound as in Section 3.2.3:

$$X^{(n+1)} \leftarrow \Phi\Big(\big\{A_k^T X^{(n)} A_k\big\}_k, \sum_k \Lambda_k^{(n)}\Big) \qquad \Lambda^{(n+1)} \leftarrow \Psi\Big(X^{(n)}, \Lambda^{(n)}\Big)\,,$$

where the map Ψ remains to be defined.

The structure of eigenvalues and eigenvectors of tropical Kraus maps is yet to be fully explored, since it is, for instance, not yet known whether the spectrum of a tropical Kraus map is finite. The convergence of the Krasnoselkii-Mann iteration of tropical Kraus maps also remains an open problem in the absence of a perturbation parameter, or an arbitrarily small perturbation. We point out that although the present work has used tropical Kraus maps in the setting of quadratic forms, it may be possible to extend their use in applications dealing with different geometric objects, such as polyhedra.

Finally, the use of tropical Kraus maps may not be the only alternative to interior-point methods to solve the large scale LMIs that we are dealing with here. The good structure of the LMIs may enable an efficient analysis by means of first order methods, such as gradient coordinate descent [FR15].

APPENDIX A

Elements of Semidefinite Programming

A linear matrix inequality (LMI for short) refers to a constraint of the form

$$A_0 + \sum_{k=1}^{d} x_k A_k \succcurlyeq 0\,, \tag{A.1}$$

where $x \in \mathbb{R}^d$ is the variable, $(A_k)_{1 \leqslant k \leqslant d}$ are given symmetric $n \times n$ matrices and $\preccurlyeq$ is the Löwner order. In other words, given a symmetric matrix $A(x_1, \ldots, x_d)$ whose entries depend in an affine way on $x \in \mathbb{R}^d$, the constraint "$A(x_1, \ldots, x_d)$ is positive semidefinite" is an LMI.

Several LMIs can be combined into a single LMI, since

$$A(x) \succcurlyeq 0 \wedge B(x) \succcurlyeq 0 \iff \begin{pmatrix} A(x) & 0 \\ 0 & B(x) \end{pmatrix} \succcurlyeq 0\,.$$

A constraint of the form $X \succcurlyeq A_0$ is an LMI in the variable $X \in \mathcal{S}_n$, since

$$X \succcurlyeq A_0 \iff \sum_{i \leqslant j} x_{ij} E^s_{i,j} \succcurlyeq A_0$$

in the variables $X = (x_{ij})_{1 \leqslant i \leqslant j \leqslant n}$, where $E^s_{i,j}$ denotes the matrix with zeroes everywhere except for a 1 in the (i,j)-th and (j,i)-th entry.

Finally, the combination of the LMIs $A(x) \preccurlyeq 0$ and $A(x) \succcurlyeq 0$ allows one to consider equality constraints $A(x) = 0$.

The problem of minimizing a convex function in the variable x that satisfies the LMI in Equation (A.1) is called a semidefinite program (SDP). We refer to [BEFB94] for introductive background on these programs.

Semidefinite programs can be solved in "polynomial time" in the following approximate sense (semidefinite feasibility is not known to be polynomial time in the Turing model of computation). Given an accuracy parameter $\varepsilon > 0$, one can obtain, in particular by interior point methods (the most efficient in practice), a ε-approximate solution of a SDP in a number of arithmetic operations which is polynomial in n, d, $\log \varepsilon$, and $\log(R/r)$, assuming that the set $\mathcal{F}$ of vectors which satisfy (A.1) is such that $B(a, r) \subset \mathcal{F} \subset B(a, R)$ for some point $a \in \mathbb{R}^n$, where $B(a, r)$ denotes the Euclidean ball of center a and radius r, see [dKV16]. We warn the reader, however, that the exponent of the polynomial is relatively high. (see Section 6.3.4 for details). Hence, it is essential for scalability purposes to limit as far as possible the growth of the dimension n and of the number of variables d, which is one of our main goals in this thesis.

Bibliography

[ACH+95] R. Alur, C. Courcoubetis, N. Halbwachs, T.A. Henzinger, P.-H. Ho, X. Nicollin, A. Olivero, J. Sifakis, and S. Yovine. The algorithmic analysis of hybrid systems. *Theoretical Computer Science*, 138(1):3 – 34, 1995. Hybrid Systems.

[ACN92] J. R. Angelos, C. C. Cowen, and S. K. Narayan. Triangular truncation and finding the norm of a hadamard multiplier. *Linear Algebra and its Applications*, 170:117 – 135, 1992.

[AFGM+15] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash. Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys Tutorials*, 17(4):2347–2376, Fourthquarter 2015.

[AG15] A. Adjé and P.-L. Garoche. Automatic synthesis of piecewise linear quadratic invariants for programs. In *Proceedings of VMCAI*, pages 99–116, 2015.

[AGG12] A. Adjé, S. Gaubert, and E. Goubault. Coupling policy iteration with semi-definite relaxation to compute accurate numerical invariants in static analysis. *Logical Methods in Computer Science*, 8(1), 2012.

[AGG+15] Xavier Allamigeon, Stéphane Gaubert, Eric Goubault, Sylvie Putot, and Nikolas Stott. A scalable algebraic method to infer quadratic invariants of switched systems. In *2015 International Conference on Embedded Software, EMSOFT 2015, Amsterdam, Netherlands, October 4-9, 2015*, pages 75–84, 2015.

[AGG+17] X. Allamigeon, S. Gaubert, E. Goubault, S. Putot, and N. Stott. A fast method to compute disjunctive quadratic invariants of numerical programs. In *2017 International Conference on Embedded Software, EMSOFT 2017, Seoul, South Korea, October 15-20, 2017*, 2017. To appear.

[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.

[AGN15] M. Akian, S. Gaubert, and R. Nussbaum. Uniqueness of the fixed point of nonexpansive semidifferentiable maps. *Trans. of AMS*, 2015. To appear.

[AGS+16] X. Allamigeon, S. Gaubert, N. Stott, E. Goubault, and S. Putot. A scalable algebraic method to infer quadratic invariants of switched systems. *ACM Trans. Embedded Comput. Syst.*, 15(4):69:1–69:20, 2016.

[AIM10] L. Atzori, A. Iera, and G. Morabito. The internet of things: A survey. *Computer Networks*, 54(15):2787 – 2805, 2010.

[AJPR14] A. A. Ahmadi, R. M. Jungers, P. A. Parrilo, and M. Roozbehani. Joint spectral radius and path-complete graph lyapunov functions. *SIAM J. Control and Optimization*, 52(1):687–717, 2014.

[And99] T. Ando. Problem of infimum in the positive cone. In Themistocles M. Rassias and Hari M. Srivastava, editors, *Analytic and Geometric Inequalities and Applications*, pages 1–12. Springer Netherlands, Dordrecht, 1999.

[Ang13] J. Angulo. Supremum/infimum and nonlinear averaging of positive definite symmetric matrices. In *Matrix Information Geometry*, pages 3–24. Springer, 2013.

[AQV+98] M. Akian, J.-P. Quadrat, M. Viot, J. M. Taylor, and M. Atiyah. *Duality between probability and optimization*, page 331–353. Publications of the Newton Institute. Cambridge University Press, 1998.

[Bal97] K. Ball. An elementary introduction to modern convex geometry. In *Flavors of geometry*, volume 31 of *Math. Sci. Res. Inst. Publ.*, pages 1–58. Cambridge Univ. Press, 1997.

[Bar] A. Barvinok. *A Course in Convexity*. Graduate studies in mathematics. American Mathematical Soc.

[Bar81] G. P. Barker. Theory of cones. *Linear Algebra and its Applications*, 39:263 – 291, 1981.

[Bar88] N. E. Barabanov. Lyapunov indicator for discrete inclusions, I–III. *Autom. Remote Control*, 49:152–157, 1988.

[BBP+07] B. Burgeth, A. Bruhn, N. Papenberg, M. Welk, and J. Weickert. Mathematical morphology for matrix fields induced by the loewner ordering in higher dimensions. *Signal Processing*, 87:277–290, 2007.

[BEFB94] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA, June 1994.

[BGT81] R. G. Bland, D. Goldfarb, and M. J. Todd. The ellipsoid method: A survey. *Operations Research*, 29(6):1039–1091, 1981.

[Bha97] R. Bhatia. *Operator Monotone and Operator Convex Functions*, pages 112–151. Springer New York, New York, NY, 1997.

[Bha00] Rajendra Bhatia. Pinching, trimming, truncating, and averaging of matrices. *The American Mathematical Monthly*, 107(7):602–608, 2000.

[Bha03] R. Bhatia. On the exponential metric increasing property. *Linear Algebra and its Applications*, 375:211–220, 2003.

[Bha07] R. Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

[Bir57] G. Birkhoff. Extensions of jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

[BK13] B. Burgeth and A. Kleefeld. *Morphology for Color Images via Loewner Order for Matrix Fields*, pages 243–254. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[BT99] V. Blondel and J. N. Tsitsiklis. Complexity of stability and controllability of elementary hybrid systems. *Automatica*, 35(3):479 – 489, 1999.

[BT00] V. Blondel and J. N. Tsitsiklis. The boundedness of all products of a pair of matrices is undecidable. *Systems & Control Letters*, 41(2):135 – 140, 2000.

[BTN01] A. Ben-Tal and A. S. Nemirovskiaei. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[Bus50] H. Busemann. The foundations of minkowskian geometry. *Commentarii mathematici Helvetici*, 24:156–187, 1950.

[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.

[CB11] C.-T. Chang and V. D. Blondel. Approximating the joint spectral radius using a genetic algorithm framework. *IFAC Proceedings Volumes*, 44(1):8681 – 8686, 2011. 18th IFAC World Congress.

[CC77a] P. Cousot and R. Cousot. Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *Proceedings of POPL'77*, pages 238–252. ACM, 1977.

[CC77b] P. Cousot and R. Cousot. Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *Proceedings of the 4th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*, POPL '77, 1977.

[CC94] P. Cousot and R. Cousot. Higher-order abstract interpretation (and application to comportment analysis generalizing strictness, termination, projection and per analysis of functional languages). In *Computer Languages, 1994., Proceedings of the 1994 International Conference on*, pages 95–112, May 1994.

[CCF+05] P. Cousot, R. Cousot, J. Feret, L. Mauborgne, A. Miné, D. Monniaux, and X. Rival. The astreé analyzer. In *Proceedings of ESOP'05*, pages 21–30, 2005.

[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.

[CH78] P. Cousot and N. Halbwachs. Automatic discovery of linear restraints among variables of a program. In *Proceedings of POPL'78*, pages 84–96. ACM, 1978.

[CL84] M. G. Crandall and P.-L. Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Math. Comp.*, 43(167):1–19, 1984.

[Cou05] P. Cousot. Proving program invariance and termination by parametric abstraction, lagrangian relaxation and semidefinite programming. In *Proceedings of VMCAI*, volume 3385 of *LNCS*. Springer, 2005.

[DDL06] H. Du, C. Deng, and Q. Li. On the infimum problem of hilbert space effects. *Science in China Series A*, 49(4):545–556, 2006.

[Die11] M. Dierkes. *Formal Analysis of a Triplex Sensor Voter in an Industrial Context*, pages 102–116. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[dKV16] E. de Klerk and F. Vallentin. On the Turing model complexity of interior point methods for semidefinite programming. *SIAM J. Optim.*, 26(3):1944–1961, 2016.

[DLL57] L. Danzer, D. Laugwitz, and H. Lenz. Uber das lownersche ellipsoid und sein analogon unter den einem eikorper einbeschriebenen ellipsoiden. *Arch. Math.*, 8:214–219, 1957.

[DLRS10] Jesus A. De Loera, Jorg Rambau, and Francisco Santos. *Triangulations: Structures for Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[Dra12] N. Dragon. *The Geometry of Special Relativity-a Concise Course*. Springer, 2012.

[ET99] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems*. Society for Industrial and Applied Mathematics, 1999.

[FA08a] E. Feron and F. Alegre. Control software analysis, part I open-loop properties. *CoRR*, abs/0809.4812, 2008.

[FA08b] E. Feron and F. Alegre. Control software analysis, part II: closed-loop analysis. *CoRR*, abs/0812.1986, 2008.

[Far96] D. R. Farenick. Irreducible positive linear maps on operator algebras. *Proc. Amer. Math. Soc.*, 124(11):3381–3390, 1996.

[Fer04] J. Feret. Static analysis of digital filters. In *Proceedings of ESOP'04*, pages 33–48, 2004.

[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.

[FK94] J. Faraut and A. Korányi. *Analysis on Symmetric Cones*. Oxford mathematical monographs. Clarendon Press, 1994.

[FLGD+11] G. Frehse, C. Le Guernic, A. Donzé, S. Cotton, R. Ray, O. Lebeltel, R. Ripado, A. Girard, T. Dang, and O. Maler. Spaceex: Scalable verification of hybrid systems. In Shaz Qadeer Ganesh Gopalakrishnan, editor, *Proc. 23rd International Conference on Computer Aided Verification (CAV)*, LNCS. Springer, 2011.

[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.

[FR15] Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

[Gal12] J. Gallier. Notes on differential geometry and lie groups. *University of Pennsylvannia*, 2012.

[GGP09] K. Ghorbal, E. Goubault, and S. Putot. The zonotope abstract domain taylor1+. In *Proceedings of CAV'09*, pages 627–633, 2009.

[GLO05] L. Gohberg, P. Lancaster, and L Odman. *Canonical Forms*, pages 73–123. Birkhäuser Basel, Basel, 2005.

[GMQ11] S. Gaubert, W. McEneaney, and Z. Qu. Curse of dimensionality reduction in max-plus based approximation methods: Theoretical estimates and improved pruning algorithms. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, pages 1054–1061. IEEE, 2011.

[GNS+13] G. Gange, J. A. Navas, P. Schachte, H. Søndergaard, and P. J. Stuckey. *Abstract Interpretation over Non-lattice Abstract Domains*, pages 6–24. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[Gou13] E. Goubault. Static analysis by abstract interpretation of numerical programs and systems, and FLUCTUAT. In *Proceedings of SAS'13*, pages 1–3, 2013.

[GPV12] E Goubault, S Putot, and F Védrine. *Modular Static Analysis with Zonotopes*, pages 24–40. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[GQ14a] S. Gaubert and Z. Qu. Checking the strict positivity of kraus maps is np-hard. *CoRR*, abs/1402.1429, 2014.

[GQ14b] S. Gaubert and Z. Qu. The contraction rate in thompson's part metric of order-preserving flows on a cone – application to generalized riccati equations. *Journal of Differential Equations*, 256(8):2902 – 2948, 2014.

[GR17] P.-L. Garoche and P. Roux. Osdp library. `https://cavale.enseeiht.fr/osdp/`, 2017.

[Gru11] P. M. Gruber. John and loewner ellipsoids. *Discrete & Computational Geometry*, 46(4):776–788, Dec 2011.

[GS17] S. Gaubert and N. Stott. Tropical Kraus maps for optimal control of switched systems. *ArXiv e-prints*, June 2017.

[GV12] S. Gaubert and G. Vigeral. A maximin characterization of the escape rate of nonexpansive mappings in metrically convex spaces. *Math. Proc. of Cambridge Phil. Soc.*, 152:341–363, 2012.

[GVL13] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, 2013.

[GZ14] N. Guglielmi and M. Zennaro. *Stability of Linear Problems: Joint Spectral Radius of Sets of Matrices*, pages 265–313. Springer International Publishing, Cham, 2014.

[HLLL17] A. Hu, J. Lin, F. Li, and B. Luo. Cyber-physical systems security - A survey. *CoRR*, abs/1701.04525, 2017.

[JF13] M. Stingl J. Fiala, M. Koċvara. Penlab: A matlab solver for nonlinear semidefinite optimization, 2013.

[Jun09] R. Jungers. *The joint spectral radius*, volume 385 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag, Berlin, 2009.

[KA79] F. Kubo and T. Ando. Means of positive linear operators. *Mathematische Annalen*, 246:205–224, 1979.

[Kad51] R. V. Kadison. Order properties of bounded self-adjoint operators. *Proceedings of the American Mathematical Society*, 2(3):505–510, 1951.

[KBDW83] K. Kraus, A. Böhm, J. D. Dollard, and W. H. Wootters, editors. *The first Representation theorem*, pages 42–61. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983.

[KLvG14] A. Kalauch, B. Lemmens, and O. van Gaans. Riesz completions, functional representations, and anti-lattices. *Positivity*, 18(1):201–218, 2014.

[KM16] H. Kaise and W. M. McEneaney. Idempotent expansions for continuous-time stochastic control. *SIAM Journal on Control and Optimization*, 54(1):73–98, 2016.

[Koz10] V. Kozyakin. Iterative building of Barabanov norms and computation of the joint spectral radius for matrix sets. *Discrete Contin. Dyn. Syst., Ser. B*, 14(1):143–158, 2010.

[KR48] M. G. Krein and M. A. Rutman. Linear operators leaving invariant a cone in a Banach space. *Uspehi Matematičeskih Nauk*, 3:3–95, 1948. AMS Translations Number 26.

[KV97] A. B. Kurzhanski and I. Vályi. *Ellipsoidal calculus for estimation and control*. Systems & control. IIASA Boston, 1997.

[KV00] A. B. Kurzhanski and P. Varaiya. *Ellipsoidal Techniques for Reachability Analysis*, pages 202–214. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[KV06] A. A. Kurzhanskiy and P. Varaiya. Ellipsoidal toolbox (et). In *45th IEEE Conference on Decision and Control*, pages 1498–1503. IEEE, 2006.

[LL06] J. Lawson and Y. Lim. The symplectic semigroup and riccati differential equations. *Journal of Dynamical and Control Systems*, 12(1):49–77, 2006.

[LN12a] B. Lemmens and R. Nussbaum. *Nonlinear Perron-Frobenius Theory*. Cambridge Tracts in Mathematics. Cambridge University Press, 2012.

[LN12b] B. Lemmens and R. D. Nussbaum. *Non-linear Perron-Frobenius theory*, volume 189 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 2012.

[Löf04] J. Löfberg. Yalmip : A toolbox for modeling and optimization in matlab. In *In Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[Mas87] V. P Maslov. *Méthodes opératorielles*. Éd. Mir, 1987.

[Mat93] R. Mathias. The hadamard operator norm of a circulant and applications. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1152–1167, 1993.

[Mat07] MathWorks Inc. Polyspace static analyzer, `fr.mathworks.com/products/polyspace/`, 2007.

[McE07] W. M. McEneaney. A curse-of-dimensionality-free numerical method for solution of certain hjb pdes. *SIAM journal on Control and Optimization*, 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 Journal on Control and Optimization*, 48(4):2651–2685, 2009.

[MD15] W. M. McEneaney and P. M. Dower. The principle of least action and fundamental solutions of mass-spring and n-body two-point boundary value problems. *SIAM Journal on Control and Optimization*, 53(5):2898–2933, 2015.

[MG99] T. Moreland and S. Gudder. Infima of hilbert space effects. *Linear Algebra and its Applications*, 286(1):1–17, 1999.

[Min97] H. Minkowski. Allgemeine lehrsätze über die convexen polyeder. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1897:198–220, 1897.

[MK10] W. M. McEneaney and L. J. 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.

[MK87] K. G. Murty and S. N. Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, Jun 1987.

[Mot16] M. Mottl. Lacaml library. `http://mmottl.github.io/lacaml/`, 2016.

[MR05] L. Mauborgne and X. Rival. Trace partitioning in abstract interpretation based static analyzers. In M. Sagiv, editor, *European Symposium on Programming (ESOP'05)*, volume 3444 of *Lecture Notes in Computer Science*, pages 5–20. Springer-Verlag, 2005.

[NBSN13] P. Nilsson, U. Boscain, M. Sigalotti, and J. Newling. Invariant sets of defocused switched systems. In *Conference of Decision and Control*, 2013.

[Nus87] Roger D. Nussbaum. *Iterated Nonlinear Maps and Hilbert's Projective Metric: A Summary*, pages 231–248. Springer Berlin Heidelberg, Berlin, Heidelberg, 1987.

[Nus88] R. D. Nussbaum. Hilbert's projective metric and iterated nonlinear maps. *Mem. Amer. Math. Soc.*, 75(391), 1988.

[NW05] G. Norris and M. Wagner. *Airbus A380: Superjumbo of the 21st Century*. Zenith Press, 2005.

[Pap05] A. Papadopoulos. *Metric spaces, convexity and nonpositive curvature*. European Mathematical Society, 2005.

[PJ08] P.A. Parrilo and A. Jadbabaie. Approximation of the joint spectral radius using sum of squares. *Linear Algebra and its Applications*, 428(10):2385–2402, 2008.

[Pro96] V.Yu. Protasov. The joint spectral radius and invariant sets of linear operators. *Fundam. Prikl. Mat.*, 2(1):205–231, 1996.

[PT05] D. Petz and R. Temesi. Means of positive numbers and matrices. *SIAM Journal on Matrix Analysis and Applications*, 27(3):712–720, 2005.

[Qu13] Z. Qu. *Nonlinear Perron-Frobenius theory and max-plus numerical methods for Hamilton-Jacobi equations*. PhD thesis, Ecole Polytechnique X, 2013.

[Qu14] Z. Qu. Contraction of riccati flows applied to the convergence analysis of a max-plus curse-of-dimensionality–free method. *SIAM Journal on Control and Optimization*, 52(5):2677–2706, 2014.

[Ram97] M. V. Ramana. An exact duality theory for semidefinite programming and its complexity implications. *Mathematical Programming*, 77(1):129–162, Apr 1997.

[RG13] P. Roux and P.-L. Garoche. Integrating policy iterations in abstract interpreters. In *ATVA*, pages 240–254, 2013.

[RJGF12] P. Roux, R. Jobredeaux, P.-L. Garoche, and E. Feron. A generic ellipsoid abstract domain for linear time invariant systems. In *Proceedings of HSCC*, pages 105–114, 2012.

[RMF13] M. Roozbehani, A. Megretski, and E. Feron. Optimization of lyapunov invariants in verification of software systems. *IEEE Trans. Automat. Contr.*, 58(3):696–711, 2013.

[Rou13] P. Roux. *Analyse statique de systèmes de contrôle commande, synthèse d'invariants non linéaires*. PhD thesis, 2013.

[RVS16] P. Roux, Y.-L. Voronin, and S. Sankaranarayanan. *Validating Numerical Semidefinite Programming Solvers for Polynomial Invariants*, pages 424–446. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

[RZ00] S. Reich and A.J. Zaslavski. Convergence of krasnoselskii-mann iterations of nonexpansive operators. *Mathematical and Computer Modelling*, 32(11-13):1423–1431, 2000.

[SGJM10] S. Sridharan, M. Gu, M. R. James, and W. M. McEneaney. Reduced-complexity numerical method for optimal gate synthesis. *Phys. Rev. A*, 82:042319, Oct 2010.

[SH10] H. R. Shaker and J. P. How. Stability analysis for class of switched nonlinear systems. In *American Control Conference (ACC)*, pages 2517–2520, Baltimore, MD, July 2010.

[SSM05] S. Sankaranarayanan, H. B. Sipma, and Z. Manna. Scalable analysis of linear systems using mathematical programming. In *Proceedings of VMCAI'05*, pages 25–41, 2005.

[Sto16] N. Stott. Maximal lower bounds in the Löwner order. 2016.

[SWYS11] J. Shi, J. Wan, H. Yan, and H. Suo. A survey of cyber-physical systems. In *2011 International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–6, Nov 2011.

[Tar55] A. Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific J. Math.*, 5(2):285–309, 1955.

[TD09] S. Tripakis and T. Dang. chapter Modeling, Verification, and Testing Using Timed and Hybrid Automata, pages 383–436. Computational Analysis, Synthesis, & Design Dynamic Systems. CRC Press, Nov 2009. 0.

[TTT03] R. H. Tütüncü, K. C. Toh, and M. J. Todd. Solving semidefinite-quadratic-linear programs using sdpt3. *Mathematical Programming*, 95(2):189–217, 2003.

[Uhl76] F. Uhlig. A canonical form for a pair of real symmetric matrices that generate a nonsingular pencil. *Linear Algebra and its Applications*, 14(3):189 – 209, 1976.

[Wey35] H. Weyl. Elementare theorie der konvexen polyeder. *Commentarii mathematici Helvetici*, 7:290–306, 1934/35.

[Wil93] D. K. Wilde. A library for doing polyhedral operations. Technical report, 1993.

[Zha02] X. Zhan. *Matrix inequalities*, volume 1790 of *Lecture Notes in Mathematics*. Springer, 2002.

[Zie95] G. M. Ziegler. *Polytopes, Polyhedra, and Cones*, pages 27–50. Springer New York, New York, NY, 1995.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Majorants minimaux dans l'ordre de Löwner et application au calcul d'invariants de systèmes commutés

Mots clés : Ordre de Löwner, majorant minimal, invariant, système commuté, rayon spectral commun, contrôle optimal

Résumé : Le calcul d'ensembles invariants est un élément crucial en vérification de programme et en théorie du contrôle, car de tels ensembles certifient l'absence de comportements indésirables. Nous étudions en particulier les systèmes commutés, pour lesquel le calcul d'un ensemble invariant est déjà difficile. Plusieurs approches récentes utilisant des techniques d'optimisation telles la programmation semidéfinie ont été appliquées avec succès au calcul d'invariants quadratiques par morceaux pour des systèmes commutés. En revanche, ces méthodes ne sont pas utilisables en grande dimension car elles nécessitent trop de ressources informatiques. Nous développons dans cette thèse une nouvelle classe d'algorithmes pour calculer des invariants quadratiques par morceaux. Ces algorithmes reposent sur les propriétés géométriques et métriques de l'espace des matrices positive semidéfinies équipées de l'ordre de Löwner. Tout d'abord, nous caractérisons l'ensemble des majorants minimaux dans cet ordre. Nous montrons que l'ensemble des majorants minimaux de deux matrices s'identifie au quotient d'un groupe orthogonal indéfini, donnant ainsi un raffinement "quantitatif" d'un théorème de Kadison. Plus généralement, nous caractérisons les majorants minimaux dans un ordre défini par un cône et nous prouvons qu'il existe pour une grande famille de cônes une sélection de majorant minimal canonique, définie à partir des fonctions génératrices de ces cônes. Ceci généralise la définition de l'ellipsoïde de Löwner. dans le cas du cône des matrices positives semidéfinies, nous montrons que cette sélection canonique satisfait plusieurs inégalités matricielles et nous donnons des estimations de sa constante de Lipschitz par rapport à plusieurs métriques convenables définies sur l'intérieur du cône (métrique Riemannienne, métrique de Thompson). Nous appliquons ensuite ces résultats au calcul d'invariants quadratiques par morceaux. Nous formulons ce dernier comme un problème de point fixe non-linéaire sur un produit de cônes de matrices positives semidéfinies. Ce problème fait intervenir un opérateur qui peut s'interpréter comme l'analogue tropical d'une application de Kraus (un canal quantique) qui apparaît en théorie d'information quantique. Nous obtenons ainsi une classe de schémas itératifs rapides, n'utilisant les inégalités linéaires matricielles, dont nous prouvons la convergence sous quelques restrictions. Nous avons implémenté cette approche en développant l'outil MEGA ("Minimal ellipsoid geometric analyser"). Nos résultats expérimentaux démontrent une amélioration de l'ordre de quelques ordres de grandeur en termes de scalabilité (par exemple, approximation du "rayon spectral joint" en dimension 500). Nous avons aussi appliqué cette méthode à l'approximation de la fonction valeur d'un problème de contrôle optimal commutant entre des modèles linéaires-quadratiques.

Title : Minimal upper bounds in the Löwner order and application to invariant computation of switched systems

Keywords : Löwner order, minimal upper bound, invariant set, switched system, joint spectral radius, optimal control

Abstract : The computation of invariant sets for dynamical systems is a crucial element of program verification and control theory, as such sets certify the absence of unwanted behaviors. We consider in particular switched systems, for which the computation of invariants is already difficult. Recently, several approaches based on optimization techniques such as semi-definite programming have been applied successfully to compute piecewise quadratic invariants of switched systems. However, their high computational cost becomes prohibitive on large instances. In this thesis, we develop a new class of algorithms to compute piecewise quadratic invariants. These algorithms rely on geometrical and metric properties of the space of positive semidefinite matrices equipped with the Löwner order. First, we characterize minimal upper bounds in this order. We show in particular that the set of minimal upper bounds of two matrices can be identified to a quotient of an indefinite orthogonal group, providing a "quantitative" refinement of a theorem of Kadison. More generally, we characterize minimal upper bounds with respect to a cone ordering, and show that for a wide family of cones, there is a canonical selection of a minimal upper bound, defined in terms of the generating function of the cones. This extends the construction of the Löwner ellipsoid. In the case of the cone of positive semidefinite matrices, we show that this canonical selection satisfies several matrix inequalities, and we estimate its Lipschitz constant with respect to convenient invariant metrics defined on the interior of the cone (Riemannian metric, Thompson metric). Then, we apply these results to the computation of piecewise-quadratic invariants. We formulate the latter as a non-linear fixed point problem over a product of spaces of positive definite matrices. This problem involves an operator which may be thought of as the tropical analogue of the Kraus maps (quantum channels) arising in quantum information theory. This leads to a class of fast iterative numerical schemes, avoiding the recourse to linear matrix inequalities, which we show to converge, under some restrictions. We implemented this approach, by developing the tool MEGA ("Minimal ellipsoid geometric analyzer"), and report experimental results, on switched linear and affine systems, showing an improvement of several orders of magnitude in terms of scalability, on some instances (with e.g., approximations of the joint spectral radius in dimension 500). We also applied this method to the approximation of the value function of switched linear quadratic optimal control problems.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France