Thèse de Doctorat

Université de Limoges

Université de Limoges

ED 653 – Sciences et Ingénierie des Systèmes, Mathématiques, Informatique (SISMI)
Faculté des Sciences et Techniques – Institut de Recherche XLIM

Thèse

pour obtenir le grade de

Docteur de l'Université de Limoges

Discipline: Mathématiques Appliquées

Présentée et soutenue par

Manh Hung LE

Le 20 Décembre 2023

Études mathématiques et numériques de la complémentarité aux valeurs propres et des problèmes d'accélération dans l'optimisation du premier ordre

---

Thèse dirigée par Prof. Samir ADLY

Jury :

Président du jury:
Mme. Marianne AKIAN, Directeur De Recherche – INRIA Saclay

Rapporteurs:
M. Dong Yen NGUYEN, Full Professor – Vietnam Academy of Science and Technology
M. Sorin-Mihai GRAD, Professeur – Institut Polytechnique de Paris

Examinateurs:
Mme. Marianne AKIAN, Directeur De Recherche – INRIA Saclay
Mme. Elisabeth KÖBIS, Associate Professor – Norwegian University of Science and Technology
M. Mounir HADDOU, Professeur Des Universités – Université de Rennes
M. Dominique ORBAN, Full Professor – Ecole Polytechnique de Montréal
M. Francisco SILVA, Maître de Conférences – Université de Limoges

Directeurs:
M. Samir ADLY, Professeur Des Universités – Université de Limoges

To my mother and father, whom I endlessly love

The study of Mathematics, like the Nile, begins in minuteness but ends in magnificence.

Charles Caleb Colton

Acknowledgments

I begin by expressing my deepest gratitude to my thesis supervisor, Professor Samir Adly. His invaluable guidance, unwavering support, and insightful advice have been indispensable throughout the entirety of this research endeavor. Without his exceptional mentorship, this thesis would not have been possible. I am profoundly honored to have had the opportunity to conduct mathematical research under his esteemed supervision.

I also wish to extend my heartfelt appreciation to the late Professor Hedy Attouch, whose profound intellect and innovative ideas have greatly enriched my academic journey. Although he is no longer with us, his contributions to the field of mathematics continue to inspire and guide my work. I am deeply grateful to have had the chance to collaborate with him.

I am grateful to the two referees, Professor Nguyen Dong Yen and Professor Sorin-Mihai Grad who generously dedicated their time and expertise to read and provide feedback on my thesis. Their insightful comments and constructive criticism have significantly strengthened the quality of this work. I extend my sincere thanks to the other members of my PhD jury, Professor Marianne AKIAN, Professor Elisabeth KÖBIS, Professor Mounir HADDOU, Professor Dominique ORBAN, and Professor Francisco SILVA, for their valuable feedback and contributions during the defense of my thesis. Their expertise and insights have been instrumental in shaping the direction of my research.

The research for this dissertation was conducted at the XLIM laboratory, Faculty of Sciences and Technologies of the University of Limoges. I am deeply thankful for the outstanding support and assistance provided by the staff members of XLIM. Their expertise and dedication have significantly contributed to the success of this project. Additionally, I am grateful to my colleagues at XLIM for their insightful discussions and valuable suggestions, which have enhanced the quality of my research.

I am indebted to the University of Limoges for generously funding this thesis. Their financial support has been instrumental in the realization of this project, and I am sincerely grateful for their investment in my academic pursuits.

I would like to express my heartfelt appreciation to my Vietnamese friends, whose friendship, encouragement, and assistance have been a source of inspiration and strength throughout this journey. The memories of our shared experiences and joyous gatherings will forever hold a special place in my heart.

Lastly, I extend my deepest gratitude to my parents for their unwavering love, encouragement, and unwavering belief in my abilities. Their endless support and sacrifices have been

the cornerstone of my academic journey, and I am eternally grateful for their guidance and encouragement during both the triumphs and challenges of this endeavor.

Limoges, December 2023

Manh Hung LE

Contents

1 Introduction

Contents

1.1 Introduction en Français

Le contenu de cette thèse est divisé en deux parties. La première partie traite du sujet des problèmes de complémentarité des valeurs propres de Pareto et de leurs problèmes inverses correspondants. La seconde partie est consacrée à l'accélération des méthodes d'optimisation du premier-ordre en analysant les dynamiques inertielles associées.

1.1.1 Problèmes de complémentarité des valeurs propres de Pareto

La première étape fondamentale vers la résolution d'un large éventail de problèmes en finance, en médecine et dans de nombreuses autres disciplines consiste à formuler un modèle mathématique approprié. L'intérêt se porte alors sur la conception et l'étude d'algorithmes numériques efficaces pour traiter le modèle mathématique en question, ce qui joue un rôle central dans les mathématiques appliquées. Souvent, ces modèles peuvent être considérés comme des problèmes d'optimisation, l'objectif étant d'optimiser un ensemble de paramètres d'intérêt pratique. Les problèmes de complémentarité des valeurs propres, en particulier, sont l'un des types de modèles les plus fréquemment utilisés pour formuler une variété de problèmes dans les domaines de l'ingénierie, de l'économie et des sciences.

Les problèmes de complémentarité constituent un outil important et efficace pour relever un large éventail de défis d'optimisation numérique. Par conséquent, une multitude d'algorithmes ont été proposés et examinés dans la littérature pour traiter ces problèmes de manière efficace. Les problèmes de complémentarité aux valeurs propres (EiCP), également connus sous le nom de problèmes de valeurs propres contraints par un cône, représentent une sous-classe dans le domaine des problèmes de complémentarité. Ils étendent les problèmes classiques de valeurs propres, lorsque le cône coincide avec l'espace tout entier, un domaine d'intérêt significatif avec des applications en physique et en ingénierie. La genèse des EiCP remonte à leur apparition initiale dans l'examen des états d'équilibre statique dans les systèmes mécaniques contenant un nombre fini de degrés de libertés et soumis à un contact unilatéral avec frottement. Les problèmes de complémentarité aux valeurs propres ont été largement explorés dans la littérature à la fois d'un point de vues théorique et numérique. Les applications de l'EiCP couvrent un large éventail de domaines, notamment les analyses dynamiques de systèmes mécaniques structurels, les systèmes

vibro-acoustiques, les simulations de circuits électriques non-réguliers, le traitement du signal, la dynamique des fluides, ainsi que les problèmes de contact en mécanique. D'un point de vue mathématique, la résolution de l'EiCP consiste à trouver un nombre réel λ et un vecteur non nul correspondant $x \in \mathbb{R}^n \setminus \{0\}$ tels que la condition suivante soit satisfaite

$$K \ni x \perp (\lambda x - Ax) \in K^*, \tag{1.1}$$

où K est un cône convexe fermé dans $\mathbb{R}^n$, $\perp$ indique l'orthogonalité dans $\mathbb{R}^n$, K^* représente le cône dual positif associé à K, qui est défini par

$$K^* = \{y \in \mathbb{R}^n : \langle y, x\rangle \geq 0, \quad \forall x \in K\}.$$

Dans (1.1), $A \in M_n(\mathbb{R})$ est une matrice donnée $n \times n$ (pas nécessairement symétrique). Le scalaire λ et le vecteur x sont respectivement appelés valeur propre et vecteur propre de (1.1). Il est clair que lorsque K coïncide avec l'espace tout entier, (1.1) coincide avec le problème classique aux valeurs propres (connu en algèbre linéaire). Une situation importante correspond à l'orthant positif $K = \mathbb{R}^n_+$. Dans ce cas, (1.1) est appelé problème de complémentarité aux valeurs propres de Pareto (ou simplement problème aux valeurs propres de Pareto en abrégé).

Contribution

Dans cette thèse, nous limiterons notre étude au cas où $K = \mathbb{R}^n_+$ est l'orthant positif, d'où le problème de complémentarité aux valeurs propres de Pareto. Bien que l'analyse spectrale théorique de l'EiCP ait été bien développée (voir [124]), la recherche d'algorithmes efficaces pour la résolution de l'EiCP est absolument nécessaire. Nous considérons l'approche consistant à formuler le problème aux valeurs propres de Pareto comme un système d'équations non linéaires, dans l'objectif d'utilisation des méthodes de points intérieurs pour la résolution numérique de ces systèmes. Les méthodes de points intérieurs sont connues pour être l'une des méthodes les plus efficaces en optimisation numérique depuis le travail fondateur de Karmarkar [94] pour la programmation linéaire. De plus, les méthodes de points intérieurs en général et les méthodes primales-duales en particulier peuvent-être étendues pour traiter les problèmes d'optimisation non linéaires, et en particulier les problèmes de complémentarité non linéaires, voir par exemple les références [127]. La base de la plupart des implémentations des méthodes primales-duales est fournie par l'algorithme du prédicteur-correcteur de Mehrotra, qui sera adapté dans cette thèse au contexte de l'EiCP. La méthode des points intérieurs non paramétriques NPIPM présentée dans [148] sera adaptée à cette classe de problèmes. L'idée de base de la NPIPM est de faire du paramètre de relaxation, qui est souvent mis à jour de manière ad hoc dans les méthodes de points

intérieurs, une variable en introduisant une équation appropriée. Nous comparons ces deux méthodes avec une méthode existante appelée la méthode de projection sur treillis (LPM), et une méthode de lissage appelée la méthode Soft Max (SM) que nous proposons sur la base d'un esprit similaire à celui de la NPIPM. Enfin, nous étudions le problème inverse de l'EiCP qui revient à construire une matrice $A \in \mathcal{M}_n(\mathbb{R})$ dans laquelle son ensemble de valeurs propres de Pareto contient un ensemble prescrit de nombres réels distincts; plus précisément, nous adaptons LPM et MPCM au contexte des problèmes inverses aux valeurs propres de Pareto et nous les comparons avec plusieurs méthodes existantes.

1.1.2 Accélération des méthodes d'Optimisation du premier ordre

Pourquoi l'optimisation du premier ordre

Au cours des dernières décennies, la prolifération explosive de l'apprentissage automatique et du big data a engendré un changement de paradigme dans divers domaines scientifiques et industriels. Cette montée en puissance est emblématique d'une transition profonde de la programmation conventionnelle, fondée sur des règles, vers des approches axées sur les données, dans lesquelles les algorithmes discernent des schémas et des relations à partir d'une grande quantité d'informations. Au cœur de cette révolution se trouvent les algorithmes d'optimisation, qui servent de pivot pour affiner les modèles, améliorer la précision des prévisions et accélérer les processus de prise de décision. Ces algorithmes, enracinés dans la théorie de l'optimisation mathématique, ajustent méticuleusement les paramètres du modèle pour minimiser ou maximiser une fonction objective, garantissant ainsi l'utilisation la plus efficace des ressources disponibles. Grâce à un raffinement itératif, les techniques d'optimisation ont catalysé des percées dans diverses applications, allant du traitement du langage naturel et de la vision par ordinateur aux systèmes de recommandation et aux véhicules autonomes. Leur rôle critique dans l'exploitation de la puissance de l'apprentissage automatique et du big data est palpable, éclairant une trajectoire vers des systèmes de plus en plus sophistiqués et performants dans l'ère florissante de l'intelligence artificielle.

Étant donné le rôle central des algorithmes d'optimisation dans le paysage contemporain de l'apprentissage automatique et de l'analyse des données massives, il est impératif de souligner l'importance de la conception et de la mise en œuvre de méthodologies d'optimisation hautement efficaces. Alors que les ensembles de données continuent de croître en taille et en complexité, et que les ressources informatiques deviennent progressivement puissantes, la demande de stratégies d'optimisation capables de naviguer sur ce formidable terrain devient de plus en plus prononcée. Les algorithmes d'optimisation de

premier ordre, qui consistent en des algorithmes d'optimisation utilisant uniquement des informations de premier ordre de la fonction objective f, à savoir ∇f, occupent une position de première importance dans le domaine de l'optimisation en raison de leur efficacité et de leur évolutivité dans le traitement d'ensembles de données à grande échelle et d'espaces de paramètres à haute dimension. Ces algorithmes fonctionnent en utilisant l'information du gradient, qui indique la direction de la montée ou de la descente la plus raide d'une fonction. Cette caractéristique les rend légers sur le plan informatique et bien adaptés aux scénarios dans lesquels la mémoire et les ressources de traitement sont limitées. En outre, les algorithmes du premier ordre présentent des propriétés de convergence favorables, convergeant souvent vers un minimum local en un nombre raisonnable d'itérations. Leur simplicité et leur facilité de mise en œuvre en font un choix attrayant pour un large éventail d'applications, allant de l'apprentissage de modèles complexes d'apprentissage automatique à la résolution de problèmes d'optimisation convexe à grande échelle répandus dans divers domaines tels que le traitement des signaux, la reconstruction d'images et la finance. En outre, les algorithmes du premier ordre constituent une base solide pour des techniques d'optimisation plus sophistiquées, servant de blocs de construction pour des approches hybrides qui combinent les forces de différents paradigmes d'optimisation.

Faits historiques sur les méthodes du premier ordre

Après l'introduction de la méthode de descente du gradient (GDM) au milieu du XIXe siècle, le paysage de l'optimisation numérique s'est transformé avec l'introduction de la méthode de la boule pesante par Polyak en 1964. L'incorporation par Polyak d'un terme appelé momentum à la méthode de descente du gradient (GDM) a considérablement amélioré le taux de convergence de l'algorithme dans le cas fortement convexe. La représentation continue de la méthode de la boule pesante est l'équation différentielle du second ordre suivante

$$\text{(HBF)} \qquad \ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0.$$

Le coefficient γ devant $\dot{x}(t)$ est appelé frottement visqueux. (HBF) assure une convergence exponentielle de $f(x(t))$ vers $\min_{\mathcal{H}} f$ pour une fonction lisse fortement convexe f. Le taux de convergence de (HBF) pour les fonctions convexes générales est de $\mathcal{O}(1/t)$, ce qui n'est pas plus rapide que la méthode de la plus forte pente ou la descente de gradient.

Les travaux fondamentaux de Nesterov en 1983 ont abouti à la méthode du gradient accéléré de Nesterov (NAG). S'appuyant sur les fondements posés par la méthode de la boule pesante, (NAG) a introduit un terme de correction du momentum qui la distingue de ses prédécesseurs. Cette innovation a permis une percée dans les taux de convergence, permettant une convergence significativement plus rapide par rapport aux méthodes

précédentes. En fait, (NAG) a atteint le taux de convergence optimal de la valeur objective parmi les méthodes de premier ordre. Une question naturelle à poser à ce stade est de savoir quel est le système dynamique correspondant à (NAG) ? Su, Boyd et Candès [143] ont plus tard répondu à cette question en remplaçant γ dans (HBF) par un coefficient d'amortissement visqueux évanescent, désigné par $\gamma(t) = \alpha/t$, où α est un paramètre positif. Cela a apporté un complément substantiel au domaine. Le système dynamique en question est connu sous le nom de dynamique de Su-Boyd-Candès et est donné par

$$(\mathrm{AVD})_\alpha \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0.$$

Dans ce cas continu, on a la convergence des valeurs $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}(1/t^2)$ pour toute trajectoire $x(t)$ de $(\mathrm{AVD})_\alpha$ avec $\alpha \geq 3$. Le coefficient d'amortissement visqueux $\frac{\alpha}{t}$ tend vers zéro lorsque le temps t s'approche de l'infini, d'où la terminologie "amortissement évanescent". Les propriétés de convergence de la dynamique $(\mathrm{AVD})_\alpha$ ont fait l'objet de nombreuses études récentes, voir [22, 24, 30–32, 35, 36, 41, 43, 109, 143]. Le cas où le paramètre $\alpha = 3$ est crucial car il correspond à l'algorithme historique de Nesterov. À l'exception du cas unidimensionnel, où la convergence des trajectoires a été démontrée [36], la question de savoir si les trajectoires convergent dans ce cas est encore une question ouverte. Dans l'article Attouch-Chbani-Peypouquet-Redont [35], il a été démontré que chaque trajectoire converge faiblement vers un minimiseur de f pour des valeurs $\alpha > 3$. Le résultat discret correspondant a été obtenu par Chambolle-Dossal [65]. De plus, il a été prouvé dans [41] et [109] que pour $\alpha > 3$, le taux de convergence asymptotique des valeurs est en fait $o(1/t^2)$. Apidopoulos-Aujol-Dossal [24] et Attouch-Chbani-Riahi [36] ont étudié la situation sous-critique où $\alpha < 3$ et ont montré que le taux de convergence des valeurs objectives est $\mathcal{O}(t^{-\frac{2\alpha}{3}})$. Ces taux sont optimaux, ce qui signifie qu'ils peuvent être atteints ou approchés de manière arbitraire.

En 2009, Beck et Teboulle ont présenté l'algorithme FISTA (Fast Iterative Shrinkage-Thresholding Algorithm), apportant une contribution significative au domaine de l'optimisation convexe. Cette innovation représentait une fusion sophistiquée des techniques de gradient proximal avec la méthode du gradient accéléré de Nesterov. Le résultat est un algorithme très efficace capable de résoudre rapidement un large spectre de problèmes d'optimisation convexe ayant une structure additive, où la fonction-objectif est la somme d'une fonction lisse et d'une fonction non lisse. Cet algorithme trouve des applications dans diverses disciplines scientifiques, notamment le traitement du signal, l'apprentissage statistique, la reconstruction d'images et la modélisation parcimonieuse.

Ces dernières années, l'incorporation du terme d'amortissement piloté par le Hessien, qui implique le Hessien $\nabla^2 f$ de la fonction-objectif f, dans les systèmes dynamiques a reçu beaucoup d'attention. L'amortissement piloté par le hessien a un lien naturel avec la propriété de frottement en mécanique et en physique, voir [82]. Il permet de

contrôler et d'atténuer les effets d'oscillations qui se produisent naturellement avec les systèmes inertiels. Plusieurs travaux sur ce sujet existent dans la littérature, nous pouvons citer pare exemple Attouch-Peypouquet-Redont [42], Attouch-Chbani-Fadili-Riahi [34], et Shi-Du-Jordan-Su [138].

Contribution

Nous considérons le système dynamique non régulier suivant

$$\ddot{x}(t) + \gamma\dot{x}(t) + \partial\varphi\Big(\dot{x}(t) + \beta\nabla f(x(t))\Big) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) \ni 0,$$

qui englobe plusieurs termes différents, notamment le frottement sec (qui correspond à φ), l'amortissement visqueux et l'amortissement piloté par le Hessien. Nous dérivons de cette dynamique, par le biais d'une discrétisation temporelle, des algorithmes d'optimisation correspondants. Nous analysons les propriétés de convergence de ces algorithmes et menons ensuite des expériences numériques pour illustrer leur efficacité. En outre, nous étudions également une inclusion d'évolution doublement non linéaire de la forme

$$\text{(DRYAD)} \quad \gamma\Big(\dot{x}(t) + \beta\nabla f(x(t))\Big) + \partial\varphi\Big(\dot{x}(t) + \beta\nabla f(x(t))\Big) + \nabla f(x(t)) \ni 0, \quad t \in [t_0, \infty).$$

Nous procédons à l'accélération de la convergence de cette dynamique via les techniques de mise à l'échelle du temps et de calcul de la moyenne développées par Attouch, Bot et Nguyen [28] pour obtenir la dynamique du second ordre suivante avec des taux de convergence optimaux

$$\ddot{z}(s) + \frac{\alpha}{s}\dot{z}(s) + \frac{\gamma\beta+1}{\gamma}\nabla f\Big(z(s) + \frac{s}{\alpha-1}\dot{z}(s)\Big) + \frac{1}{\gamma}\nabla\varphi_{\frac{s}{\gamma(\alpha-1)}}\Big(-\frac{s}{\gamma(\alpha-1)}\nabla f\Big(z(s) + \frac{s}{\alpha-1}\dot{z}(s)\Big)\Big) = 0.$$

En outre, une approche duale de (DRYAD) sera examinée, dans laquelle la variable fonctionnelle du système dynamique résultant est le gradient de la fonction objective ∇f.

1.1.3 Plan de la thèse

La thèse se compose de 7 chapitres et sera organisée comme suit. Le chapitre 2 est consacré au contexte mathématique. Le chapitre 3 traite de la résolution du problème de complémentarité des valeurs propres de Pareto, où le problème en question est considéré comme un système d'équations non linéaires. Nous étudions ensuite l'utilisation de deux méthodes de points intérieurs, à savoir NPIPM et MPCM. Une méthode de lissage appelée

SM, dont l'idée s'inspire de la NPIPM, est proposée. Une méthode existante, appelée LPM, est proposée avec la SM comme contrepartie de comparaison aux deux méthodes de points intérieurs données. Le chapitre 4 traite du problème de complémentarité de l'inverse des valeurs propres de Pareto, l'objectif étant de construire une matrice A ayant pour valeurs propres de Pareto un ensemble de réels distincts donnés. Le chapitre 5 traite des algorithmes d'optimisation inertielle du premier ordre avec des effets de seuil associés au frottement sec. Dans le chapitre 6, nous étudions une équation d'évolution doublement non linéaire et son dual, après quoi les techniques de mise à l'échelle du temps et de calcul de la moyenne, développées par Attouch, Bot et Nguyen [28], seront adoptées pour obtenir une dynamique inertielle correspondante avec des taux de convergence accélérés. Le chapitre 7 présente les conclusions et les perspectives.

1.2 Introduction in English

The content of this thesis is divided into 2 parts. The first part deals with the subject of Pareto eigenvalue complementarity problems and their corresponding inverse problems. Then, from the standpoint of non regular dynamical systems, we concentrate our research on the topic of first order optimization algorithms and dynamics.

1.2.1 Pareto eigenvalue complementarity problems

The first fundamental step towards solving a wide range of problems in finance, medicine, and many other disciplines is formulating an appropriate mathematical model. Interest would be then directed to designing and studying effective numerical algorithms to tackle the mathematical model in question, which plays a central role in applied mathematics. Oftentimes, these models can be cast under optimization problems where the aim is to optimize an array of parameters of interest for practical purposes. Eigenvalue complementarity problems, in particular, are one of the most frequently used types of models utilized to formulate a variety of problems in engineering, economics and sciences.

Complementarity problems are a valuable and efficient tool for addressing a diverse range of numerical optimization challenges. Consequently, a multitude of algorithms have been put forth and scrutinized to handle these problems effectively. Eigenvalue Complementarity Problems (EiCP), also known as cone-constrained eigenvalue problems, represent a subclass within the realm of complementarity problems. They expand upon classical eigenvalue problems, an area of significant interest with applications in physics and engineering. The genesis of EiCP can be traced back to their initial appearance in the examination of static equilibrium states in finite-dimensional mechanical systems with unilateral frictional contact. Subsequently, they have been extensively explored both theoretically and numerically. The applications of EiCP span a wide array of fields,

encompassing dynamic analyses of structural mechanical systems, vibro-acoustic systems, electrical circuit simulations, signal processing, fluid dynamics, as well as contact problems in mechanics. Mathematically speaking, solving EiCP consists in finding a real number λ and a corresponding nonzero vector $x \in \mathbb{R}^n$ such that the following condition holds

$$K \ni x \perp (\lambda x - Ax) \in K^*, \tag{1.2}$$

where K is a closed convex cone in $\mathbb{R}^n$, $\perp$ indicates the orthogonality, K^* stands for its positive dual cone, which is defined by

$$K^* = \{y \in \mathbb{R}^n : \langle y, x\rangle \geq 0 \quad \forall x \in K\}.$$

In (1.2) $A \in M_n(\mathbb{R})$ is a given $n \times n$ matrix (not necessarily symmetric). The scalar λ and vector v are respectively called eigenvalue and eigenvector of (1.2). It is clear that when K coincides with the whole space, (1.2) recovers the classical eigenvalue problem in linear algebra. One important situation corresponds to the nonnegative orthant $K = \mathbb{R}^n_+$. In this case, (1.2) is called the Pareto eigenvalue complementarity problem (or just the Pareto eigenvalue problem for short).

Contribution

In this thesis, we will restrict our study to the case of $K = \mathbb{R}^n_+$ being the nonnegative orthant, hence the Pareto eigenvalue complementarity problem. While the theoretical spectral analysis for EiCP has been well-developed (see [124]), investigation towards designing efficient algorithms for solving EiCP is of absolute necessity. We consider the approach of formulating the Pareto eigenvalue problem as a nonlinear system of equations, and then our objective is to use interior point methods to contribute to the resolution of such systems. Interior point methods are known to be one of the most efficient and ubiquitous methods in numerical optimization since the founding work of Karmarkar [94] for linear programming. Moreover, interior point methods in general and primal-dual methods, in particular, can be extended to tackle nonlinear optimization problems, and in particular nonlinear complementarity problems, see e.g. [127]. The basis for most implementations of the primal-dual methods is provided by the Mehrotra predictor corrector algorithm, which will be adapted in this thesis to the context of EiCP. The Non Parametric Interior Point Method (NPIPM) which was introduced in [148] will be adapted for this class of problems. The basic idea of the NPIPM is to make the relaxation parameter, which is often updated in an ad hoc manner in interior point methods, become a variable by introducing a proper equation. We compare these two methods with an existing method called the Lattice Projection Method (LPM), and a smoothing method called the Soft Max

method (SM) that we propose based on a similar spirit as NPIPM. Finally, we study the inverse problem of EiCP which amounts to constructing a matrix $A \in \mathcal{M}_n(\mathbb{R})$ in which its set of Pareto eigenvalues contains a prescribed set of distinct real numbers; specifically, we adapt LPM and MPCM to the context of inverse Pareto eigenvalue problems and compare them with several existing methods.

1.2.2 First order optimization from the perspective of dynamical systems

Machine learning and why first order optimization

In recent decades, the explosive proliferation of machine learning and big data has engendered a paradigm shift in various scientific and industrial domains. This surge is emblematic of a profound transition from conventional, rule-based programming to data-driven approaches, wherein algorithms discern patterns and relationships from copious amounts of information. Central to this revolution are optimization algorithms, serving as the linchpin for refining models, enhancing predictive accuracy, and expediting decision-making processes. These algorithms, rooted in mathematical optimization theory, meticulously fine-tune model parameters to minimize or maximize an objective function, thus ensuring the most efficient utilization of available resources. Through iterative refinement, optimization techniques have catalyzed breakthroughs in diverse applications, ranging from natural language processing and computer vision to recommender systems and autonomous vehicles. Their critical role in harnessing the power of machine learning and big data is palpable, illuminating a trajectory towards ever more sophisticated and capable systems in the burgeoning era of artificial intelligence.

Given the pivotal role of optimization algorithms in the contemporary landscape of machine learning and big data analytics, it is imperative to underscore the significance of designing and implementing highly efficient optimization methodologies. As datasets continue to burgeon in size and complexity, and computational resources become progressively potent, the demand for optimization strategies that can navigate this formidable terrain becomes increasingly pronounced. First-order optimization algorithms, which consist of optimization algorithms that only utilize first order information of the objective function f, namely ∇f, hold a position of paramount importance in the realm of optimization due to their efficiency and scalability in handling large scale datasets and high dimensional parameter spaces. These algorithms operate by utilizing gradient information, which indicates the direction of the steepest ascent or descent of a function. This characteristic renders them computationally lightweight and well-suited for scenarios where memory and

processing resources are constrained. Additionally, first-order algorithms exhibit favorable convergence properties, often converging to a local minimum in a reasonable number of iterations. Their simplicity and ease of implementation make them an attractive choice for a wide array of applications, ranging from training complex machine learning models to solving large-scale convex optimization problems prevalent in various domains such as signal processing, image reconstruction, and finance. Moreover, first-order algorithms provide a solid foundation for more sophisticated optimization techniques, serving as building blocks for hybrid approaches that combine the strengths of different optimization paradigms.

Historical facts on first order methods

Beginning with the inception of the Gradient Descent Method (GDM) in the mid-19th century, the optimization landscape saw a transformative shift with the introduction of the Heavy Ball Method by Polyak in 1964. Polyak's incorporation of momentum terms to (GDM) significantly enhanced the convergence rate of the algorithm in the strongly convex case. The continuous representation of the heavy ball method is the following second order differential equation

$$\text{(HBF)} \qquad \ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0.$$

The coefficient γ in front of $\dot{x}(t)$ is said to correspond to the viscous damping. (HBF) ensures exponential convergence of $f(x(t))$ to $\min_{\mathcal{H}} f$ for a smooth strongly convex function f. The convergence rate of (HBF) for general convex functions is $\mathcal{O}(1/t)$, which isn't faster than the steepest descent approach.

Nesterov's seminal work in 1983 yielded the Nesterov Accelerated Gradient (NAG) method. Building on the foundation laid by the Heavy Ball Method, (NAG) introduced a momentum correction term that distinguishes it from its predecessors. This innovation resulted in a breakthrough in convergence rates, enabling significantly faster convergence compared to previous methods. In fact, (NAG) achieved the optimal convergence rate in the objective value amongst first order methods. A natural question to be asked at this point is what is the corresponding dynamical system for (NAG)? Su, Boyd, and Candès [143] later gave an answer to this question by replacing γ in (HBF) by a vanishing viscous damping coefficient, denoted by $\gamma(t) = \alpha/t$, where α is a positive parameter. This has made a substantial addition to the field. The dynamical system in question is known as the Su-Boyd-Candès dynamic and is given by

$$\text{(AVD)}_{\alpha} \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0.$$

We have the inversely quadratic convergence rate of the values $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}(1/t^2)$ for any trajectory $x(t)$ of $(\mathrm{AVD})_\alpha$ with $\alpha \geq 3$. The viscous damping coefficient $\frac{\alpha}{t}$ vanishes (tends to zero) as time t approaches infinity, hence the terminology Asymptotic Vanishing Damping. The convergence properties of the dynamic $(\mathrm{AVD})_\alpha$ have been the subject of many recent studies, see [24,30–32,35,36,41,43,44,109,143]. The case where the parameter $\alpha = 3$ is crucial since it matches Nesterov's historical algorithm. With the exception of the one dimensional case, where convergence of the trajectories has been demonstrated [36], the question of whether the trajectories converge in this case is still unanswered. According to Attouch-Chbani-Peypouquet-Redont [35], each trajectory weakly converges to a minimizer of f for values $\alpha > 3$. The corresponding algorithmic result was obtained by Chambolle-Dossal [65]. Furthermore, it has been proved in [41] and [109] that for $\alpha > 3$, the asymptotic convergence rate of the values is actually $o(1/t^2)$. Apidopoulos-Aujol-Dossal [24] and Attouch-Chbani-Riahi [36] investigated the subcritical situation where $\alpha < 3$ and showed that the convergence rate of the objective values is $\mathcal{O}(t^{-\frac{2\alpha}{3}})$. These rates are optimal, which means they can be reached or approached arbitrarily closely.

In 2009, Beck and Teboulle introduced the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [52], making a significant contribution to the domain of convex optimization. This innovation represented a sophisticated amalgamation of proximal gradient techniques with Nesterov's accelerated gradient method. The result was a highly efficient algorithm capable of swiftly solving a broad spectrum of convex optimization problems having the additive structure, where the objective function is the sum of a smooth function and a nonsmooth function. This algorithm finds applications in various scientific disciplines, including signal processing, statistical learning, image reconstruction, and sparse modeling.

In recent years, the trend of incorporating the Hessian driven damping term, which involves the Hessian $\nabla^2 f$ of the objective function f, to dynamical systems has been receiving a great deal of attention. The Hessian driven damping has a natural connection with the strong damping property in mechanics and physics, see [82]. It helps to control and attenuate the oscillation effects that occur naturally with inertial systems. Several works on this topic include Attouch-Peypouquet-Redont [42], Attouch-Chbani-Fadili-Riahi [34], and Shi-Du-Jordan-Su [138].

Contribution

We will be considering the following non regular dynamical system

$$\ddot{x}(t) + \gamma\dot{x}(t) + \partial\varphi\Big(\dot{x}(t) + \beta\nabla f(x(t))\Big) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) \ni 0,$$

that encompasses several different terms including dry friction (which corresponds to φ), viscous damping, and Hessian driven damping. We derive from this dynamic, through temporal discretization, corresponding optimization algorithms. We analyze the convergence properties of the algorithms and conduct numerical experiments afterward to illustrate their efficiency. Additionally, we also study a doubly nonlinear evolution inclusion of the form

$$\text{(DRYAD)} \quad \gamma\Big(\dot{x}(t)+\beta\nabla f(x(t))\Big)+\partial\varphi\Big(\dot{x}(t)+\beta\nabla f(x(t))\Big)+\nabla f(x(t))\ni 0, \quad t\in[t_0,\infty).$$

We proceed to accelerate the convergence of this dynamic via the time scaling and averaging techniques developed by Attouch, Bot, and Nguyen [28] to attain the following second order dynamic with optimal convergence rates

$$\ddot{z}(s)+\frac{\alpha}{s}\dot{z}(s)+\frac{\gamma\beta+1}{\gamma}\nabla f\Big(z(s)+\frac{s}{\alpha-1}\dot{z}(s)\Big) + \frac{1}{\gamma}\nabla\varphi_{\frac{s}{\gamma(\alpha-1)}}\Big(-\frac{s}{\gamma(\alpha-1)}\nabla f\Big(z(s)+\frac{s}{\alpha-1}\dot{z}(s)\Big)\Big)=0.$$

Additionally, a dual approach to (DRYAD) will be examined where the resulting dynamical system's functional variable is the gradient of the objective function ∇f. By doing so, we gain a greater understanding of the behavior of ∇f.

1.2.3 Outline of the thesis

The dissertation consists of 7 chapters and will be organized as follows. Chapter 2 is devoted to the mathematical background. Chapter 3 deals with the resolution of the Pareto eigenvalue complementarity problem where the problem at hand is cast under a system of nonlinear equations. We then study the use of two interior point methods, namely NPIPM and MPCM. A smoothing method called SM with the idea inspired by NPIPM is proposed. Together with SM as comparison counterparts to the two given interior point methods is an existing method called LPM. Chapter 4 deals with the inverse Pareto eigenvalue complementarity problem where the aim is to construct a matrix A attaining a set of given distinct reals as Pareto eigenvalues. Chapter 5 deals with first order inertial optimization algorithms with threshold effects associated with dry friction. In Chapter 6, we investigate into a doubly nonlinear evolution equation and its dual after which the time scaling and averaging techniques, developed by Attouch, Bot, and Nguyen [28], will be adopted to attain corresponding inertial dynamics with accelerated convergence rates. Chapter 7 provides conclusions and perspectives.

2

Mathematical background

Contents

We will be presenting in this section the fundamentals of Hilbert spaces, and convex analysis. To fix the idea, we will only be working with vector spaces over the real numbers. A reference for this includes [51].

2.1 Hilbert spaces

In optimization, the inner product plays a crucial role in defining concepts like gradient, orthogonality, and convexity, providing the mathematical foundation for various algorithms and optimization techniques.

Definition 2.1 *Given a vector space X, a function $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ is called an inner product on X if it satisfies the following three conditions*

(i) Linearity: $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle \quad \forall x, y \in X$, *and* $a, b \in \mathbb{R}$,

(ii) Symmetry: for every $x, y \in X$, $\langle x, y \rangle = \langle y, x \rangle$,

(iii) Positive definiteness: if x is not zero then $\langle x, x \rangle > 0$.

We say that the norm $\| \cdot \|$ on X is induced from the inner product $\langle \cdot, \cdot \rangle$ on X if $\|x\| = \langle x, x \rangle^2$ for every $x \in X$.

Definition 2.2 *Given a vector space X, equipped with an inner product $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$. X is said to be a Hilbert space if it is complete with respect to the norm induced by $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$.*

From now on, unless stated otherwise, when considering a Hilbert space we will always use the notation $\| \cdot \|$ and $\langle \cdot, \cdot \rangle$ for its inner product and the corresponding norm, respectively. We list several useful equalities, which can be checked directly using the definitions of norms and inner products.

Proposition 2.1 *Given a Hilbert space $\mathcal{H}$, we have the following:*

(i) $\|x+y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y\rangle$ *for all* $x, y \in \mathcal{H}$,

(ii) $\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2)$,

(iii) $\langle x, y\rangle = \frac{1}{4}(\|x+y\|^2 + \|x-y\|^2)$.

We present the Cauchy Schwartz inequality which plays a crucial role in Hilbert spaces. It compares the inner product of 2 arbitrary elements of a Hilbert space with the product of their norms.

Theorem 2.1 (Cauchy Schwartz inequality) *Let* $\mathcal{H}$ *be a Hilbert space, the following inequality holds true for every* $x, y \in \mathcal{H}$

$$|\langle x, y\rangle| \leq \|x\|\|y\|.$$

Proof. For every real number α, we have according the the definition of inner products

$$\langle x - \alpha y, x - \alpha y\rangle = \|x\|^2 - 2\alpha\langle x, y\rangle + \alpha^2\|y\|^2.$$

Hence, the determinant of this quadratic function (with respect to α) has to be non-positive, which translates to

$$|\langle x, y\rangle|^2 \leq \|x\|^2\|y\|^2.$$

This completes the proof. ∎

Note that $\langle \cdot, \cdot \rangle$ is a bilinear operator, so according to the Cauchy Schwartz inequality, we have that $\langle \cdot, \cdot \rangle$ is a continuous bilinear mapping.

Theorem 2.2 (Riesz representation theorem) *If* $T : \mathcal{H} \longrightarrow \mathbb{R}$ *is a bounded/ continuous linear mapping on a Hilbert space* $\mathcal{H}$*, there exists some* $v \in \mathcal{H}$ *such that for every* $u \in \mathcal{H}$ *we have*

$$T(u) = \langle u, v\rangle.$$

Moreover, $\sup_{\|u\|=1} T(u) = \|v\|$.

The Riesz representation theorem answers the question of what the analogy of the gradient of functionals on Euclidean spaces is for functionals on Hilbert spaces. The result is captured in the following corollary.

Corollary 2.1 (Gradient of smooth functionals on Hilbert spaces) *Suppose* $f : \mathcal{H} \longrightarrow \mathbb{R}$ *is Fréchet differentiable at* $x \in \mathcal{H}$*. Denote by* $f'(x)$ *the Fréchet derivative of* f *at* x*. By*

definition, $f'(x)$ is a bounded linear mapping from $\mathcal{H}$ into the reals. Applying the Riesz representation theorem, there exists some element in $\mathcal{H}$, denoted by $\nabla f(x) \in \mathcal{H}$ such that

$$f'(x)(y) = \langle \nabla f(x), y \rangle.$$

We call $\nabla f(x)$ the gradient of f at x.

Definition 2.3 *Let $\mathcal{H}$ be a Hilbert space and (x_k) be a sequence in $\mathcal{H}$. Then (x_k) is said to converge to x if and only if for every $x \in \mathcal{H}$*

$$\langle x_k, y \rangle \longrightarrow \langle x, y \rangle \text{ as } t \longrightarrow \infty.$$

Theorem 2.3 *Any Hilbert space is a reflexive Banach space. As a result, Kakutani's theorem yields that any bounded sequence in a Hilbert space $\mathcal{H}$ has a subsequence converging weakly to an element in $\mathcal{H}$.*

It is straightforward from the above theorem that we have the following corollary, which is a useful remark to show the weak convergence of a bounded sequence in Hilbert space.

Corollary 2.2 *Let $\mathcal{H}$ be a Hilbert space and (x_k) be a bounded sequence in $\mathcal{H}$. If the set of weak accumulation points of (x_k) has a cardinality of at most 1, then (x_k) converges weakly to some element in $\mathcal{H}$.*

In fact, using this corollary we can prove a more verification-friendly result to show the weak convergence of a bounded sequence in Hilbert spaces, namely the Opial's lemma.

Lemma 2.1 (Opial's lemma) *Let S be a nonempty set of a Hilbert space $\mathcal{H}$. Suppose that $(x_k)_k$ is a sequence in $\mathcal{H}$ which satisfies*

- *$\lim_{k\to\infty} \|x_k - p\|$ exists for all $p \in S$.*
- *For each subsequence $(x_{k_l})_l$ of $(x_k)_k$ that converges weakly to x, we have $x \in S$.*

Then, there exists $x \in S$ such that $(x_k)_k$ converges weakly to x.

Proof. As mentioned right before the statement of the lemma, it is sufficient to show that if x and y are two weak cluster points of (x_k), meaning there are two subsequences (x_{m_k}) and (x_{n_k}) such that (x_{m_k}) and (x_{n_k}) converges weakly to x and y respectively, then $x = y$.
It is apparent that we have for all $n \in \mathbb{N}$ (or from the first item of Proposition 2.1)

$$\|x_n - x\|^2 = \|x_n - y\|^2 + \|x - y\|^2 + 2\langle x_n - x, x_n - y \rangle. \tag{2.1}$$

The first assumption of the lemma follows that $\|x_n - x\|^2$ and $\|x_n - y\|^2$ are strongly convergent, say to a and b, respectively. Now, replacing x_n in 2.1 respectively by (x_{m_k}) and (x_{n_k}) and pass to the limit, we obtain

$$a = b + \|x - y\|^2,$$
$$a = b - \|x - y\|^2.$$

This means that $\|x - y\|^2$=0, hence $x = y$. The proof is completed. ∎

2.2 Convex analysis

To start off, we introduce the concept of convex sets.

Definition 2.4 *A subset C of a Hilbert space is said to be convex if and only if for every x and y in C and $\lambda \in [0, 1]$ we have*

$$\lambda x + (1 - \lambda)y \in C.$$

Particularly, $\mathcal{H}$ and $\varnothing$ are convex.

A convex combination of a collection $\{x_1, x_2, \cdots, x_n\}$ is an element defined by

$$x = \sum_{i=1}^{n} \alpha_i x_i,$$

where $\alpha_i \geq 0$ such that $\sum_{i=1}^{n} \alpha_i = 1$.

The set of all convex combinations of elements of a set C is called the convex hull of C, denoted by Co(S). The following provides a characterization of convex sets through convex hulls.

Proposition 2.2 *A subset C of a Hilbert space $\mathcal{H}$ is convex if and only if it coincides with its convex hull.*

Let us enumerate some of the basic properties of convex sets.

Proposition 2.3 *The following properties hold true*

(i) The sum of two convex sets is a convex set.

(ii) The product of a convex set with a real number remains convex.

(iii) The intersection of an arbitrary family of convex sets is a convex set.

(iv) The convex hull Co(S) is the smallest convex sets containing S.

(v) The closure and interior of a convex set is a convex set.

Definition 2.5 *Let C be a subset of a Hilbert space $\mathcal{H}$. We call C a cone if $\lambda x \in C$ for all $x \in \mathcal{H}$ and $\lambda \geq 0$.*

Definition 2.6 *Let C be a convex subset of a Hilbert space $\mathcal{H}$. The positive dual cone of C is defined by*

$$C^* = \{y \in \mathcal{H} : \langle y, x\rangle \geq 0 \quad \forall x \in C\}.$$

Definition 2.7 *Let C be a convex subset of a Hilbert space $\mathcal{H}$. The normal cone to C at $x \in C$ is defined by*

$$N_C(x) = \{v \in \mathcal{H} : \langle v, y - x\rangle \leq 0 \text{ for all } y \in C\}.$$

Let us now move on to some notions concerning functions.
Definition 2.8 *An extended real-valued function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ defined on a Hilbert space $\mathcal{H}$ is caller proper if its domain, defined by*

$$\mathrm{dom}(f) = \{x \in \mathcal{H} : f(x) < +\infty\},$$

is non empty.

Definition 2.9 *Let $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be an extended real-valued function defined on a Hilbert space $\mathcal{H}$. The function f is said to be convex if and only if for all $x, y \in \mathrm{dom}(f)$ and $\lambda \in [0, 1]$ we have*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

If the above inequality holds strictly for $\lambda \in (0, 1)$, meaning that for all $x, y \in \mathrm{dom}(f)$ and $\lambda \in (0, 1)$, we have

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y),$$

we call the function f to be strictly convex.

Proposition 2.4 *Let $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be an extended real-valued function defined on a Hilbert space $\mathcal{H}$. If f is convex then for any $\alpha \in \mathbb{R}$ the sublevel set*

$$\{x \in \mathcal{H} : f(x) \leq \alpha\},$$

is also convex.

It's worth mentioning the first-order characterization of convex functions which is a fundamental property that distinguishes convex functions from non convex ones

Definition 2.10 *A smooth function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ defined on a Hilbert space $\mathcal{H}$ is convex if and only if for all $x, y \in \mathrm{dom}(f)$, one of the following is satisfied*

(i) $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$.

(ii) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$.

Essentially, the first item says that if the function's graph lies above its tangent lines at all points in its domain, then it is convex. Conversely, if this condition is not satisfied for even a single pair of points, the function is not convex. The second item, on the other hand, characterizes the monotonicity of (smooth) convex functions.

Definition 2.11 *An extended real-valued function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ defined on a Hilbert space $\mathcal{H}$ is called lower semicontinuous at a point $x \in \mathrm{dom}(f)$ if for every sequence (x_k) converging to x we have*

$$f(x) \leq \liminf_{n \longrightarrow \infty} f(x_k).$$

We call f to be lower semicontinuous on $\mathcal{H}$ if it is lower semicontinuous at every point in $\mathrm{dom}(f)$. In this case, we also refer to f as a closed function. The interpretation of this is due to the following proposition

Proposition 2.5 *Given a function $f : \mathcal{H} \longrightarrow \mathbb{R} \cup \{+\infty\}$, f is lower semicontinuous on $\mathcal{H}$ if and only if for every $\alpha \in \mathbb{R}$ the sublevel set*

$$\{x \in \mathcal{H} : f(x) \leq \alpha\},$$

is closed in $\mathcal{H}$.

We have similar definitions for the concept of weakly lower semicontinuity.

Definition 2.12 *An extended real-valued function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ defined on a Hilbert space $\mathcal{H}$ is called weakly lower semicontinuous at a point $x \in \mathrm{dom}(f)$ if for every sequence (x_k) converging weakly to x we have*

$$f(x) \leq \liminf_{n \longrightarrow \infty} f(x_k).$$

We call f to be weakly lower semicontinuous on $\mathcal{H}$ if it is weakly lower semicontinuous at every point in $\mathrm{dom}(f)$. In this case, we also refer to f as a weakly closed function. The interpretation of this is due to the following proposition

Proposition 2.6 *Given a function $f : \mathcal{H} \longrightarrow \mathbb{R} \cup \{+\infty\}$, f is lower semicontinuous if and only if for every $\alpha \in \mathbb{R}$ the sublevel set*

$$\{x \in \mathcal{H} : f(x) \leq \alpha\},$$

is weakly closed in $\mathcal{H}$.

We have the following interesting result connecting the two concepts in the case of f being convex.

Proposition 2.7 *When $f : \mathcal{H} \longrightarrow \mathbb{R} \cup \{+\infty\}$ is convex, f is lower semicontinuous if and only if it is weakly lower semicontinuous.*

Proof. Taking advantage of the characterization of these two concepts through sublevel sets, it is necessary and sufficient to prove that for any $\alpha \in \mathbb{R}$ the closedness and weak closedness of the sublevel set

$$\{x \in \mathcal{H} : f(x) \leq \alpha\},$$

are equivalent. Since f is convex, the sublevel set is also convex. The result follows from the Mazur's theorem. ■

Definition 2.13 *We call a function $f : \mathcal{H} \longrightarrow \mathbb{R} \cup \{+\infty\}$ strongly convex with constant m if*

$$g(x) = f(x) - \frac{m}{2}\|x\|^2,$$

defines a convex function g.

Analogously to convex functions, first order characterizations can be obtained for strong convexity

Proposition 2.8 *Suppose $f : \mathcal{H} \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a smooth function. Then the following statements are equivalent*

(i) *f is strongly convex with constant m.*

(ii) *For any $x, y \in \mathrm{dom}(f)$, $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq m\|x - y\|^2$.*

(iii) *For any $x, y \in \mathrm{dom}(f)$, $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{m}{2}\|x - y\|^2$.*

We should also mention this useful estimation when it comes to strongly convex functions

Proposition 2.9 *If $f : \mathcal{H} \longrightarrow \mathbb{R}$ is a strongly convex function with constant m, then f has a unique global minimizer x^* satisfying*

$$\frac{m}{2}\|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2m}\|\nabla f(x)\|^2 \quad \forall x \in \mathrm{dom}(f).$$

One of the most important quantitative assumptions in convex optimization and analysis is the Lipschitz continuity of a function. This property essentially guarantees that a function's output doesn't change too quickly in response to small changes in its input. Specifically, if a function is Lipschitz continuous, there exists a non-negative constant (referred to as the Lipschitz constant) that bounds how much the function's value can differ for any two points in its domain, relative to the distance between those points. To put it rigorously, we have the following definition

Definition 2.14 *Consider a smooth function $f : \mathcal{H} \longrightarrow \mathbb{R}$ defined on a Hilbert space $\mathcal{H}$. f is said to be Lipschitz continuous with constant L if and only if for all $x, y \in \mathcal{H}$*

$$\|f(x) - f(y)\| \leq L\|x - y\|.$$

A crucial implication from the Lipschitz continuity is the gradient descent lemma, which establishes a quadratic upper bound on how much the function value can increase when moving from x to y, taking into account both the gradient and the Lipschitz continuity. It provides a theoretical foundation for the convergence proofs of various convex optimization algorithms.

Lemma 2.2 (Gradient descent lemma) *Consider a smooth function (of class C^1) $f : \mathcal{H} \longrightarrow \mathbb{R}$ defined on a Hilbert space $\mathcal{H}$. Suppose that f is Lipschitz continuous with constant L, then we have*

$$f(x) \leq f(y) + \langle \nabla f(y)x - y\rangle + \frac{L}{2}\|x - y\|^2 \quad \forall x, y \in \mathcal{H}.$$

In fact, when f is convex, the Lipschitz continuity assumption is equivalent to the gradient descent lemma.

Now, let us introduce the notion of subgradients in the context of convex analysis. Subgradients are a fundamental concept in convex analysis, providing a generalization of gradients for non differentiable convex functions.

Definition 2.15 *For a convex function $f : \mathcal{H} \longrightarrow \mathbb{R}$ defined on a Hilbert space $\mathcal{H}$, a vector g is considered a subgradient of f at a point x if, for all points $y \in \mathcal{H}$,*

$$f(x) \geq f(y) + \langle g, x - y\rangle.$$

The set consisting of all such g is called the subdifferential of f at x, denoted by $\partial f(x)$

Essentially, a subgradient g provides a linear approximation of f that lies below the function's graph. Subgradients are particularly crucial in situations where functions are

not differentiable at certain points. For example, in cases involving non-smooth convex functions or functions defined on non-smooth domains, traditional gradients may not exist. Subgradients, however, offer a means to generalize the concept of derivative and extend it to these non-differentiable settings. Moreover, subgradients are indispensable in the study of duality theory.

As a generalization of the gradient, the subdifferential also enjoys the monotonicity property
Proposition 2.10 *Consider a convex function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ defined on a Hilbert space $\mathcal{H}$, for all $x, y \in \mathrm{dom}(f)$, $u \in \partial f(x)$, and $v \in \partial f(y)$, we have $\langle u - v, x - y\rangle \geq 0$.*

The following proposition gives a necessary and sufficient condition for minimizers of a convex function via subdifferential.
Proposition 2.11 *Consider a closed convex function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ defined on a Hilbert space $\mathcal{H}$, x^* is a minimizer of f if and only if $0 \in \partial f(x^*)$.*

The Fenchel conjugate is an important concept in convex analysis and optimization theory. It plays a crucial role in duality theory, which is fundamental in understanding and solving optimization problems.
Definition 2.16 *Given a function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ defined on a Hilbert space $\mathcal{H}$, its Frenchel conjugate $f^* : \mathcal{H} \to [-\infty, +\infty]$ is defined by*

$$f^*(v) = \sup\{\langle v, x\rangle - f(x) : x \in \mathcal{H}\},$$

and the biconjugate of f is $f^{**} = (f^*)^*$.

Proposition 2.12 *Given a proper function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ defined on a Hilbert space $\mathcal{H}$, then we have the following properties*

(i) The Fenchel conjugate f^* is convex and lower semicontinuous on $\mathcal{H}$.
(ii) $f(x) + f^*(v) \geq \langle v, x\rangle$ for all $x, v \in \mathcal{H}$.
(iii) $f^{**}(x) \leq f(x)$ for all $x \in \mathcal{H}$.
(iv) $f^{**}(x) = f(x)$ for all $x \in \mathcal{H}$ if and only if f is closed and convex.

The following result gives a connection between subgradients and Fenchel conjugates of convex functions.
Proposition 2.13 *Given a convex function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ defined on a Hilbert space $\mathcal{H}$, then for any $x \in \mathrm{dom}(f)$ we have that $v \in \partial f(x)$ if and only if*

$$f(x) + f^*(v) = \langle v, x\rangle.$$

3

Interior point methods for solving Pareto eigenvalue complementarity problems

Contents

This chapter covers the material discussed in the published paper [13], which was produced in collaboration with S. Adly and M. Haddou

3.1 Introduction

The area of complementarity problems (CP) has received great attention over the last few decades due to their various applications in engineering, economics, and sciences. Since the pioneering work by Lemke and Howson, who showed that computing a Nash equilibrium point of a bimatrix game can be modeled as a linear complementarity problem [99], the theory of CP has become a useful and effective tool for studying a wide class of problems in numerical optimization. As a result, a variety of algorithms have been proposed and analyzed in order to deal efficiently with these problems, see the thorough survey [73] and references therein. On the other hand, Eigenvalue Complementarity Problems (EiCP) (also known as cone-constrained eigenvalue problems) form a particular subclass of complementarity problems that extend the classical (linear algebra) eigenvalue problems. Solving classical eigenvalue problems is also a topic of great interest and finds its various applications in physics and engineering, see [79,146]. EiCP appeared for the first time in the study of static equilibrium states of finite dimensional mechanical systems with unilateral frictional contact [122], and since then it has been widely studied both theoretically and numerically. On this subject, we refer to [15–17, 75, 88, 90–92, 96, 102, 116, 129] and references therein. Applications of EiCP were found in many fields such as the dynamic analysis of structural mechanical systems, vibro-acoustic systems, electrical circuit simulation, signal processing, fluid dynamics, and contact problems in mechanics (see for instance [105–108, 123]). Mathematically speaking, solving EiCP consists in finding a real number $\lambda \in \mathbb{R}$ and a corresponding nonzero vector $x \in \mathbb{R}^n \setminus \{0\}$ such

that the following condition holds

$$K \ni x \perp (\lambda x - Ax) \in K^*, \tag{3.1}$$

where K is a closed convex cone in $\mathbb{R}^n$, $\perp$ indicates the orthogonality, and K^* stands for its positive dual cone, which is defined by

$$K^* = \{y \in \mathbb{R}^n : \langle y, x\rangle \geq 0 \quad \forall x \in K\}.$$

In (3.1) $A \in \mathcal{M}_n(\mathbb{R})$ is a given $n \times n$ matrix (not necessarily symmetric).
Such scalar λ and vector x are respectively called eigenvalue and eigenvector of (3.1). It is clear that when K coincides with the whole space, (3.1) recovers the classical eigenvalue problem in linear algebra. One important situation corresponds to the nonnegative orthant $K = \mathbb{R}^n_+$. In this case, (3.1) is called the Pareto eigenvalue complementarity problem (or just the Pareto eigenvalue problem for short). It is shown in [124, 136] that

$$3(2^{n-1} - 1) \leq \max_{A \in \mathcal{M}_n(\mathbb{R})} \text{card}[\sigma(A)] \leq n2^{n-1} - (n-1),$$

where $\sigma(A)$ denotes the Pareto spectrum of A containing all eigenvalues of the Pareto eigenvalue problem corresponding to A, and $\text{card}[\sigma(A)]$ denotes the cardinality of $\sigma(A)$. This means that the number of Pareto eigenvalues grows exponentially with the dimension n of the matrix A. Therefore, finding all Pareto eigenvalues of a large or even medium-sized problem is not an easy task, especially in the context of iterative methods. For instance, a matrix of order 25 may have more than 3 million Pareto eigenvalues, which is notably huge.

While the theoretical spectral analysis for EiCP has been well-developed (see [124]), investigation towards designing efficient algorithms for solving EiCP is of absolute necessity. There are several interesting approaches for solving EiCP in the literature. Let us briefly summarize some of the existing methods.

• The Semismooth Newton Method (SNM), studied in [17], is specially tailored for dealing with the Pareto eigenvalue problem

$$x \geq 0_n, \quad \lambda x - Ax \geq 0_n, \quad \langle x, \lambda x - Ax\rangle = 0, \tag{3.2}$$

which is one of the most interesting examples of cone-constrained eigenvalue problems. The symbol 0_n refers to the n-dimensional zero vector and $x \geq 0_n$ indicates that each component

of x is nonnegative. The idea proposed in [17] is to convert (3.2) into a system of equations

$$\begin{aligned} U_\varphi(x,y) &= 0_n, && (3.3)\\ Ax - \lambda x + y &= 0_n, && (3.4)\\ \langle 1_n, x\rangle - 1 &= 0, && (3.5) \end{aligned}$$

and then apply a nonsmooth Newton type algorithm to the (semismooth) resulting system. Here U_φ is the vector function corresponding to some complementarity function φ

$$U_\varphi(x,y) = \begin{bmatrix} \varphi(x_1,y_1) \\ \varphi(x_2,y_2) \\ \vdots \\ \varphi(x_n,y_n) \end{bmatrix},$$

where $\varphi : \mathbb{R}^2 \to \mathbb{R}$ stands for any function that satisfies $\varphi(a,b) = 0 \iff a \geq 0, \quad b \geq 0, \quad ab = 0$.
We refer to [17] for more details.

• The Lattice Projection Method (LPM) proposed in [15] is another semismooth approach for solving Pareto eigenvalue problems. It is different from SNM in the sense that LPM does not use any complementarity function. Its principle is based on the observation that for every $\lambda > 0$

$$0_n \leq x \perp \lambda x - Ax \geq 0_n \iff (\mathrm{P}_{\mathbb{R}^n_+} \circ A)(x) = \lambda x,$$

where $\mathrm{P}_{\mathbb{R}^n_+}$ stands for the projection operator onto $\mathbb{R}^n_+$. Therefore, the system to solve in this case can be rewritten as

$$\begin{aligned} \max(\tilde{y}, 0_n) - \lambda x &= 0_n, \\ Ax - \tilde{y} &= 0, \\ \langle 1_n, x\rangle - 1 &= 0, \end{aligned} \qquad (3.6)$$

where the max function is carried out componentwisely. Finally, a nonsmooth Newton type algorithm is used to solve (3.6). In [15], the authors have shown that LPM is a more efficient and robust method for solving Pareto eigenvalue problems than SNM.
There are several other approaches to tackle EiCP problems. Indeed, one can see EiCP as a global optimization problem and then use Branch-and-Bound techniques [91] or some other global optimization methods. One can also use smoothing techniques by interpreting EiCP as a system of nonlinear complementarity equations, see, for instance, [81,103,141,150,151].

In this chapter, we consider the approach of formulating the Pareto eigenvalue problem as a nonlinear system of equations, and then our purpose is to use interior point methods to solve such systems. Interior point methods are known to be one of the most efficient and ubiquitous methods in numerical optimization since the founding work of Karmarkar [94] for linear programming. Moreover, interior point methods in general and primal dual methods in particular can be extended to tackle nonlinear optimization problems, and in particular nonlinear complementarity problems, see e.g. [127]. The basis for most implementations of the primal-dual methods is provided by the Mehrotra predictor corrector algorithm, which is considered in this chapter in the context of EiCP. The Non Parametric Interior Point Method (NPIPM) which was introduced in [148] will be adapted to our context. The basic idea of the NPIPM is to make the relaxation parameter, which is often updated in an ad hoc manner in interior point methods, become a variable by introducing a proper equation. We compare these two methods with two other ones namely the Soft Max method (SM) and the Lattice Projection Method (LPM).

This chapter is organized as follows: In Section 3.2, we introduce two considered interior point methods including NPIPM and the Mehrotra Predictor Corrector Method (MPCM). The smoothing method SM is introduced in Section 3.3. In these sections, along with the methods' formulation, we also provide conditions under which the nonsingularity of the Jacobian at a solution is ensured. Section 3.4 is devoted to some numerical tests for solving three Pareto eigenvalue problems corresponding to three given matrices of order 3, 4, and 5, which are known to have the maximum number of Pareto eigenvalues (9, 23, and 56 Pareto eigenvalues respectively). This test game constitutes a first comparison of the four methods, namely MPCM, NPIPM, LPM, and SM. After that, we compare the four methods by using the performance profiles [72] on a set of data taken from random generators and the MatrixMarket. The average computing time, the average number of iterations, the percentage of failure and the maximum number of eigenvalues found by each solver are used as performance measures to compare these algorithms. Section 3.5 provides an application of NPIPM, MPCM and LPM to a geomechanical fracture problem with data provided by IFP Energies Nouvelles. In this application, we consider the closed convex cone $K = \mathbb{R}^m_+ \times \mathbb{R}^{n-m}$ in the EiCP. In this latter case, the problem is known as the partially cone-constrained eigenvalue problem. Finally, we show in Section 3.6 that MPCM and NPIPM can be extended to deal with more general cone-constrained eigenvalue problems, including the quadratic pencil eigenvalue complementarity problem. We end this chapter with some concluding remarks in Section 3.7.

3.2 Interior point methods for eigenvalue complementarity problems

3.2.1 Non Parametric Interior Point Method (NPIPM)

The problem (3.1) can be represented by the following system of equations

$$Ax - \lambda x + y = 0_n, \tag{3.7}$$

$$\langle 1_n, x\rangle - 1 = 0, \tag{3.8}$$

$$x \bullet y = 0, \tag{3.9}$$

$$x \in K, y \in K^*, \tag{3.10}$$

where $\bullet$ denotes the Hadarmard product meaning that $x \bullet y = (x_1y_1, ..., x_ny_n)^T$. Now, unambiguously, we will use the notation xy instead of $x \bullet y$.

The Non Parametric Interior Point Method (NPIPM) was first introduced in the thesis [148]. Inspired by the classical IPM for optimization, the first step towards NPIPM is to introduce the relaxation parameter $\mu > 0$ and then to replace the equation (3.9) by $xy = \mu 1_n$. We now add the following equation which distinguishes NPIPM from the classical IPM

$$\frac{1}{2}\left(\|\mathrm{P}_{K^*}(-x)\|^2 + \|\mathrm{P}_K(-y)\|^2\right) + \mu^2 + \epsilon\mu = 0, \tag{3.11}$$

where $\epsilon > 0$ is a fixed positive real number.
One of the reasons behind considering equation (3.11) is that with $\mu \geq 0$, this equation is equivalent to condition (3.10). Indeed, since K is a closed convex cone, it holds that $K = K^{**}$. It is straightforward to check that

$$\begin{cases} \mathrm{P}_{K^*}(-x) = 0, \\ \mathrm{P}_K(-y) = 0. \end{cases} \iff \begin{cases} x \in K, \\ y \in K^*. \end{cases}$$

Notice that when we introduce equation (3.11), μ becomes a variable, so the situation is quite different from classical IPMs where μ is a parameter. This additional equation in the NPIPM's algorithm will make μ be driven to zero automatically using the Newton method. At this point, NPIPM is somehow more advantageous than classical IPMs in the sense that there is no need to find a good strategy to drive μ to zero, which can vary from one problem to another.

Set $X = (x, y, \lambda)$, and denote by L the following system of equations

$$L(X, \mu) = \begin{bmatrix} Ax - \lambda x + y \\ \langle \mathbf{1}_n, x \rangle - 1 \\ xy - \mu 1_n \end{bmatrix},$$

and

$$G_K(X, \mu) = \begin{bmatrix} L(X, \mu) \\ \frac{1}{2}\left(\|\mathrm{P}_{K^*}(-x)\|^2 + \|\mathrm{P}_K(-y)\|^2\right) + \mu^2 + \epsilon\mu \end{bmatrix}.$$

In the particular case where K is the nonnegative orthant $\mathbb{R}^n_+$, which is a self dual closed convex cone, we can easily rewrite G_K (and for the ease of notation $G_{\mathbb{R}^n_+} = G$) as follows

$$G := G_{\mathbb{R}^n_+}(X, \mu) = \begin{bmatrix} L(x, y, \lambda, \mu) \\ \frac{1}{2}\sum_{i=1}^{n}(\min\{x_i, 0\}^2 + \min\{y_i, 0\}^2) + \mu^2 + \epsilon\mu \end{bmatrix}.$$

Accordingly, its Jacobian matrix has the form

$$J_G(X, \mu) = \begin{bmatrix} \frac{\partial L}{\partial X} & \frac{\partial L}{\partial \mu} \\ M & 2\mu + \epsilon \end{bmatrix}. \tag{3.12}$$

where

$$M = [\min\{x_1, 0\}, ..., \min\{x_n, 0\}, \min\{y_1, 0\}, ..., \min\{y_n, 0\}]. \tag{3.13}$$

When $x, y \geq 0$ (in the componentwise sense), the determinant $\det(J_G(X, \mu)) = (2\mu + \epsilon)\det(\frac{\partial L}{\partial X})$. One can notice that the presence of the term $\epsilon\mu$ in the equation (3.11) prevents $J_G(X, \mu)$ from being ill-conditioned near a solution, in which case μ may get too small.

NPIPM Algorithm

1. Initialization: Select $X^0 = (x^0, y^0, \lambda^0)$ such that $x^0 \in \mathrm{int}(K), y^0 \in \mathrm{int}(K^*), \lambda^0 \in \mathbb{R}$ and $\mu^0 > 0$; Set $k = 0$.
2. Unless the stopping criterion is satisfied, do the following
3. Compute the Newton direction by solving the following linear system

$$J_G(X^k, \mu^k)d^k = -G(X^k, \mu^k) \quad \text{with} \quad d^k = \begin{bmatrix} d_X^k \\ d_\mu^k \end{bmatrix} \text{ and } d_X^k = \begin{bmatrix} d_x^k \\ d_y^k \\ d_\lambda^k \end{bmatrix}. \tag{3.14}$$

4. Find a stepsize $\alpha^k \in (0,1]$ as large as possible such that

$$x^k + \alpha^k d_x^k \in \text{int}(K), \tag{3.15}$$

$$y^k + \alpha^k d_y^k \in \text{int}(K^*). \tag{3.16}$$

5. Update $X^{k+1} = X^k + \alpha^k d^k$ and set $k = k+1$.

Basically, the NPIPM algorithm employs a (damped) Newton method to solve the system $G(X,\mu) = 0$ where stepsizes are chosen such that during the iteration, x^k and y^k respectively lie in the interior of K and K^*.

In what follows, we consider the nonnegative orthant case $K = \mathbb{R}^n_+$ for our theoretical results.

Proposition 3.1 *Assume that the μ^k generated by the NPIPM algorithm is well-defined and that the sequence of stepsizes satisfies* $\liminf \alpha^k > 0$. *Then $(\mu^k)_k$ is a positive decreasing sequence and converges to 0 as k goes to $+\infty$.*

Proof. With $(x,y,\mu) \in \text{int}(\mathbb{R}^n_+) \times \text{int}(\mathbb{R}^n_+) \times \text{int}(\mathbb{R}_+)$, the linear system, for which the Newton direction is satisfied, gives

$$\begin{bmatrix} \frac{\partial L}{\partial X} & \frac{\partial L}{\partial \mu} \\ 0_{1\times(2n+1)} & 2\mu+\epsilon \end{bmatrix} \begin{bmatrix} d_X \\ d_\mu \end{bmatrix} = - \begin{bmatrix} L(x,y,\lambda,\mu) \\ \mu^2 + \epsilon\mu \end{bmatrix}.$$

This implies that $(2\mu+\epsilon)d_\mu = -(\mu^2+\epsilon\mu)$. In other words, we have $d_\mu = -\dfrac{\mu^2+\epsilon\mu}{2\mu+\epsilon}$. Therefore, with $\alpha \in (0,1]$ being the stepsize, we have

$$\begin{aligned} \mu^+ &:= \mu + \alpha d_\mu \\ &= \mu - \alpha \frac{\mu^2+\epsilon\mu}{2\mu+\epsilon} \\ &= \frac{(2-\alpha)\mu^2 + (1-\alpha)\epsilon\mu}{2\mu+\epsilon} > 0 \ \text{ (since } \mu > 0 \text{ and } \alpha \in (0,1]). \end{aligned}$$

Moreover, it is clear that $\mu^+ < \mu$. Thus, μ^k is a positive and decreasing sequence and therefore has a limit which is denoted by μ^*. We have also shown that

$$\mu^{k+1} = \mu^k - \alpha^k \frac{(\mu^k)^2 + \epsilon\mu^k}{2\mu^k + \epsilon}.$$

Letting $k \to \infty$ yields

$$\alpha^k \frac{(\mu^k)^2 + \epsilon\mu^k}{2\mu^k + \epsilon} \to 0.$$

Since $\liminf \alpha^k > 0$ and $\mu^k > 0$, it follows that $\frac{(\mu^*)^2 + \epsilon\mu^*}{2\mu^* + \epsilon} = 0$ and consequently, $\mu^* = 0$. ■

Remark 3.1 The assumption on $\liminf \alpha^k$ might seem too restrictive at first glance. However, when the algorithm converges to a nondegenerate solution (i.e., the Jacobian matrix at this solution is nonsingular), we observe practically superlinear or quadratic convergence and at the few last iterations we can choose stepsizes α^k to be near 1 (for more details, see e.g. [57, Theorem 6.9]).

Remark 3.2 Note that in the algorithm NPIPM, it is possible to use a line search such as the Armijo line search after Step 4. In order not to distort the comparison with the other methods like LPM, we have opted not to use any line search technique in this context. Furthermore, extensive preliminary numerical experiments show that NPIPM with or without line search has equivalent performances.

The following lemmas will be useful. The first one is inspired by the Schur complement result.

Lemma 3.1 *Suppose A, B, C, and D are matrices of dimension $n \times n, n \times m, m \times n$, and $m \times m$, respectively. If D is nonsingular, we can easily check the following decomposition*

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I_n & B \\ 0 & D \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & 0 \\ D^{-1}C & I_m \end{bmatrix},$$

and therefore we have $\det\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right) = \det(D)\det(A - BD^{-1}C)$.

The proof of Lemma 3.1 is straightforward and will be omitted.

Lemma 3.2 *Given an $n \times n$ matrix A and $x \in \mathbb{R}^n \setminus \{0\}$, set*

$$S = A^T A + 1_n 1_n^T - \frac{1}{\|x\|^2} A^T x x^T A.$$

Then, one has the following equivalent statements

(a) $M = \begin{bmatrix} A & -x \\ 1_n^T & 0 \end{bmatrix}$ *is nonsingular.*

(b) S is nonsingular.

(c) S is (symmetric) positive definite.

(d) For all $y \in \mathbb{R}^n \setminus \{0\}$, if $Ay \in \operatorname{span}(x)$, then $\sum_{i=1}^n y_i \neq 0$.

Proof. A direct calculation gives

$$M^T M = \begin{bmatrix} E & F \\ F^T & g \end{bmatrix},$$

where $E = A^T A + 1_n 1_n^T, F = -A^T x$ and $g = \|x\|^2 > 0$.

We have

$$\begin{aligned} M \text{ is nonsingular} &\iff M^T M \text{ is nonsingular} \\ &\iff E - F g^{-1} F^T \text{ is nonsingular (due to Lemma 3.1)} \\ &\iff S \text{ is nonsingular.} \end{aligned}$$

The equivalence between (b) and (c) is due to the fact that S is positive semidefinite. Indeed, let $y \in \mathbb{R}^n \setminus \{0\}$, we have

$$y^T S y = \underbrace{\|Ay\|^2 - \frac{|x^T Ay|^2}{\|x\|^2}}_{m_y \geq 0} + \underbrace{\left(\sum_{i=1}^{n} y_i\right)^2}_{n_y \geq 0},$$

where $m_y \geq 0$ due to the Cauchy-Schwarz inequality.
For the last equivalence,

$$\begin{aligned} S \text{ is positive definite} &\iff \forall y \in \mathbb{R}^n \setminus \{0\}, \ m_y + n_y > 0 \\ &\iff \forall y \in \mathbb{R}^n \setminus \{0\}, \ m_y \neq 0 \text{ or } n_y \neq 0 \\ &\iff \forall y \in \mathbb{R}^n \setminus \{0\}, \ \text{if } m_y = 0, \text{ then } n_y \neq 0. \end{aligned}$$

Note that $m_y = 0$ if and only if Ay and x are linearly dependent, which can be translated into $Ay \in \text{span}(x)$. The proof is thereby completed. ∎

The next proposition provides a characterization of the nonsingularity of the Jacobian matrix J_G at a solution $(X, 0)$ of the system $G(X, \mu) = 0$, where $X = (x, y, \lambda)$. Before giving the statement of this proposition, we would like to introduce some notations. For a given matrix D of size $n \times n$, I and J being subsets of $\{1, 2, \ldots, n\}$, D_{IJ} denotes the submatrix created by rows and columns of D with indices in I and J respectively. Similarly, if $x \in \mathbb{R}^n$, x_I represents the vector containing components of x with indices in I.

Proposition 3.2 *Denote $X = (x, y, \lambda)$. Assume $(X, 0)$ is a solution of $G(X, \mu) = 0$. Then, $J_G(X, 0)$ is nonsingular if and only if (x, y) satisfies the strict complementarity conditions, i.e., $x_i + y_i > 0$ for every $i = 1, 2, ..., n$, and the principle submatrices $\tilde{A}_{\alpha\alpha}$ of $\tilde{A} = A - \lambda I_n$, where $\alpha = \{1 \leq l \leq n : x_l \neq 0\}$, satisfy*

$$\text{for all } y \in \mathbb{R}^{|\alpha|} \setminus \{0\}, \text{ if } \tilde{A}_{\alpha\alpha} y \in \text{span}(x_\alpha), \text{ then } \sum_{i=1}^{|\alpha|} y_i \neq 0. \tag{3.17}$$

Proof. Since at the solution we have $x \geq 0$ and $y \geq 0$, the Jacobian matrix $J_G(X)$ has the form

$$J_G(X, 0) = \begin{bmatrix} \frac{\partial L}{\partial X} & \frac{\partial L}{\partial \mu} \\ 0_{1\times(2n+1)} & \epsilon \end{bmatrix}.$$

It is necessary and sufficient to show the nonsingularity of the first block $\frac{\partial L}{\partial X}$ of $J_G(X, 0)$. A simple computation yields

$$\frac{\partial L}{\partial X} = \begin{bmatrix} \tilde{A} & I_n & -x \\ 1_n^T & 0_{1\times n} & 0 \\ \text{diag}(y) & \text{diag}(x) & 0_{n\times 1} \end{bmatrix},$$

where $\tilde{A} = A - \lambda I_n$.
We can see that if the strict complementarity does not hold, then there exists $i \in \{1, 2, \ldots, n\}$ such that $x_i = y_i = 0$. This, according to the form of $\frac{\partial L}{\partial X}$ shown above, leads to $\frac{\partial L}{\partial X}$ being singular. Through this observation, it is seen that the strict complementarity assumption is not discardable.
To complete the proof, we show that $J_G(X, 0)$ is nonsingular if and only if the condition (3.17) holds. If $y = 0$, then the strict complementarity condition implies $x_i > 0$ for every $i = 1, .., n$. Using the Laplace expansion along the last n rows of $\frac{\partial L}{\partial X}$ gives

$$\begin{aligned} \left|\det\left(\frac{\partial L}{\partial X}\right)\right| &= \left|\prod_{l=1}^{n} x_l \det\left(\begin{bmatrix} \tilde{A} & -x \\ 1_n^T & 0 \end{bmatrix}\right)\right| \\ &\neq 0 \quad \text{(due to Lemma 3.2)}. \end{aligned}$$

If $y \neq 0$, we assume it has k nonzero components $y_{j_1}, .., y_{j_k}$, where $1 \leq j_l \leq n$ for all $1 \leq l \leq k$. Due to the strict complementarity assumption, the vector x has $n - k$ nonzero components which we will denote by $x_{i_1}, .., x_{i_{n-k}}$, where $1 \leq i_l \leq n$ for all $1 \leq l \leq n-k$. Set

$$\alpha = \{i_1, i_2, .., i_{n-k}\} \text{ and } \beta = \{j_1, j_2, .., j_k\}.$$

Since exchanging rows (resp. columns) of a matrix does not change its rank, with the strict complementarity assumption we can do so for the matrix $\frac{\partial L}{\partial X}$ in a proper way so that we obtain the following matrix with the same rank as $\frac{\partial L}{\partial X}$

$$H = \left[\begin{array}{c|c|c} \begin{bmatrix} \tilde{A}_{\beta\beta} & \tilde{A}_{\beta\alpha} \\ \tilde{A}_{\alpha\beta} & \tilde{A}_{\alpha\alpha} \end{bmatrix} & \begin{bmatrix} I_{k\times k} & 0_{k\times(n-k)} \\ 0_{(n-k)\times k} & I_{(n-k)\times(n-k)} \end{bmatrix} & \begin{bmatrix} 0_{k\times 1} \\ -x_\alpha \end{bmatrix} \\ \hline 1_n^T & 0_{1\times n} & 0 \\ \hline \begin{bmatrix} y_{j_1} & & & & & & \\ & y_{j_2} & & & & & \\ & & \ddots & & & & \\ & & & y_{j_k} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix} & \begin{bmatrix} 0 & & & & & & \\ & 0 & & & & & \\ & & \ddots & & & & \\ & & & 0 & & & \\ & & & & x_{i_1} & & \\ & & & & & \ddots & \\ & & & & & & x_{i_{n-k}} \end{bmatrix} & 0_{n\times 1} \end{array}\right],$$

where $x_\alpha = [x_{i_1}, .., x_{i_{n-k}}]^T$.
Applying the Laplace expansion along the last $n-k$ rows of the H and Lemma 3.2, we get successively

$$|\det(H)| = \left|\prod_{l=1}^{n-k} x_{i_l} \det\left(\begin{bmatrix} \tilde{A}_{\beta\beta} & \tilde{A}_{\beta\alpha} & I_{k\times k} & 0_{k\times 1} \\ \tilde{A}_{\alpha\beta} & \tilde{A}_{\alpha\alpha} & 0_{(n-k)\times k} & -x_\alpha \\ 1_k^T & 1_{n-k}^T & 0_{1\times k} & 0 \\ \mathrm{diag}(y_{j_1}, .., y_{j_k}) & 0_{k\times(n-k)} & 0_{k\times k} & 0_{k\times 1} \end{bmatrix}\right)\right| \tag{3.18}$$

$$= \left|\prod_{l=1}^{n-k} x_{i_l} \prod_{l=1}^{k} y_{j_l} \det\left(\begin{bmatrix} \tilde{A}_{\beta\alpha} & I_{k\times k} & 0_{k\times 1} \\ \tilde{A}_{\alpha\alpha} & 0_{(n-k)\times k} & -x_\alpha \\ 1_{n-k}^T & 0_{1\times k} & 0 \end{bmatrix}\right)\right| \tag{3.19}$$

$$= \left|\prod_{l=1}^{n-k} x_{i_l} \prod_{l=1}^{k} y_{j_l} \det\left(\begin{bmatrix} \tilde{A}_{\alpha\alpha} & -x_\alpha \\ 1_{n-k}^T & 0 \end{bmatrix}\right)\right| \tag{3.20}$$

$$\neq 0. \tag{3.21}$$

This completes the proof of Proposition 3.2. ∎

Remark 3.3 We can see that the condition for the Jacobian's nonsingularity shown in Proposition 3.2 is not restrictive to the eigenvalue complementarity problems. The fact that the strict complementarity assumption cannot be eliminated has made our proof become simple since only basic linear algebra is used. We note that for problems where the Jacobian matrix can be nonsingular without the strict complementarity conditions, we

may need further assumptions such as the P-matrix property. For more details, we refer to Theorem 2.8 in [93].
Now we give an example in dimension 2 to illustrate Proposition 3.2.
Example 3.1 Consider the following matrix

$$A = \begin{bmatrix} 3 & -4/3 \\ 3 & -1 \end{bmatrix}.$$

It can be checked that $\lambda_1 = 1$ is a Pareto eigenvalue of A, which is also a double standard eigenvalue of A. This follows from $\ker(A - \lambda_1 I_2)^2 = \mathbb{R}^2$. So the assumptions of Proposition 3.2 cannot be satisfied and therefore the Jacoibian $J_G(X_1, 0)$ is singular.
On the other hand, another solution of this problem is $\lambda_2 = -1$ with the corresponding eigenvector $x_2 = [0 \quad 1]^T$ and dual vector $y_2 = [4/3 \quad 0]^T$. This solution satisfies the strict complementarity, and moreover with

$$\alpha = \{2\} \text{ and } \tilde{A} = \begin{bmatrix} 4 & -4/3 \\ 3 & 0 \end{bmatrix},$$

we have

$$S_\alpha = \tilde{A}_{\alpha\alpha}^T \tilde{A}_{\alpha\alpha} + 1_{|\alpha|} 1_{|\alpha|}^T - \frac{1}{\|x_\alpha\|^2} \tilde{A}_{\alpha\alpha}^T x_\alpha x_\alpha^T \tilde{A}_{\alpha\alpha} = 1 \neq 0,$$

We can see that this solution satisfies all the assumptions of Proposition 3.2. Therefore, the Jacobian $J_G(X_2, 0)$ is nonsingular, where $X_2 = (x_2, y_2, \lambda_2)$.

3.2.2 Mehrotra Predictor Corrector Method (MPCM)

MPCM was first proposed in 1989 by Sanjay Mehrotra [110], as a variant of the primal-dual interior point method for optimization problems. Most of today's interior-point general-purpose software for linear and nonlinear programming are based on predictor-corrector algorithms like the one of Mehrotra. We give now a description of the method when applied to eigenvalue complementarity problems. For this purpose, let us set

$$F(X) = \begin{bmatrix} Ax - \lambda x + y \\ \langle 1_n, x \rangle - 1 \\ xy \end{bmatrix}, \quad \text{where} \quad X = (x, y, \lambda). \tag{3.22}$$

The Jacobian matrix of F has the form

$$J_F(X) = \begin{bmatrix} A - \lambda I_n & I_n & -x \\ 1_n^T & 0_{1\times n} & 0 \\ \mathrm{diag}(y) & \mathrm{diag}(x) & 0_{n\times 1} \end{bmatrix}.$$

MPCM Algorithm

1. Choose an initial point such that $x^0 \in \mathrm{int}(K), y^0 \in \mathrm{int}(K^*)$, $\lambda^0 \in \mathbb{R}$ and let $k = 0$.
2. Compute the affine scaling (predictor) direction d_a^k, which is given by solving the linear system $J_F(X^k)d_a^k = -F(X^k)$, and then compute a stepsize $\alpha_a^k \in (0, 1]$ that ensures

$$x^k + \alpha_a^k dx_a^k \in \mathrm{int}(K), \tag{3.23}$$
$$y^k + \alpha_a^k dy_a^k \in \mathrm{int}(K^*), \tag{3.24}$$

   where

$$d_a^k = \begin{bmatrix} dx_a^k \\ dy_a^k \\ d\lambda_a^k \end{bmatrix}.$$

3. Use the information from the predictor step to compute the corrector direction by solving the following linear system

$$J_F(X^k)d_c^k = -F(X^k) + B^k \text{ with } B^k = \begin{pmatrix} 0_{(n+1)\times 1} \\ \mu^k 1_n - dx_a^k dy_a^k \end{pmatrix}, \tag{3.25}$$

   where $\mu^k = \gamma^k\sigma^k$ with $\gamma^k = \frac{1}{n}\langle x^k, y^k\rangle$ and $\sigma^k = \left(\frac{r_a^k}{r^k}\right)^3$ is the adaptively chosen centering parameter, where

$$r^k = \gamma^k, \tag{3.26}$$
$$r_a^k = \tfrac{1}{n}\langle x^k + \alpha_a^k dx_a^k, y^k + \alpha_a^k dy_a^k\rangle. \tag{3.27}$$

4. Find a step size $\alpha_c^k \in (0, 1]$ such that

$$x^k + \alpha_c^k dx_c^k \in \mathrm{int}(K), \tag{3.28}$$
$$y^k + \alpha_c^k dy_c^k \in \mathrm{int}(K^*), \tag{3.29}$$

then compute the next iterate $X^{k+1} = X^k + \alpha_c^k d_c^k$ and update $k = k + 1$.

Remark 3.4 The Jacobian matrix J_F associated with MPCM is just the matrix $\frac{\partial L}{\partial X}$ in the previous section. Hence, the nonsingularity condition for MCPM in the case $K = \mathbb{R}^n_+$ is the same as that of NPIPM.

Remark 3.5 Despite its efficiency in practice, there is no convergence result available yet for the Mehrotra predictor corrector method even in a general context of nonlinear programming. Here, we do not give any result on the convergence of MPCM when applied to solve Pareto eigenvalue problems.

3.3 Smoothing Method

For comparison purposes, we propose in this section a smoothing method, called the Soft Max method (SM) for solving Pareto eigenvalue problems. It is known that several smoothing techniques can be used to address nonlinear complementarity problems where we would substitute nonsmooth equations with differentiable approximations. The first step to be done towards SM is to observe that the condition $K \ni x \perp (\lambda x - Ax) \in K^*$ can be presented as follows:

For all $\rho > 0$, we have

$$K \ni x \perp (\lambda x - Ax) \in K^* \Longleftrightarrow x = \mathrm{P}_K(x - \rho y) \quad \text{with} \quad y = \lambda x - Ax. \tag{3.30}$$

Indeed, since K is a closed convex cone, one has for $y = \lambda x - Ax$:

$$\begin{aligned} K \ni x \perp (\lambda x - Ax) \in K^* &\Longleftrightarrow -y \in \mathrm{N}_K(x) \\ &\Longleftrightarrow -\rho y \in \mathrm{N}_K(x) \\ &\Longleftrightarrow x - \rho y \in x + \mathrm{N}_K(x) \\ &\Longleftrightarrow x = \mathrm{P}_K(x - \rho y), \end{aligned}$$

where $\mathrm{N}_K(x)$ stands for the nornal cone to K at x.

Consider $K = \mathbb{R}^n_+$, in this case (3.30) becomes

$$(x \geq 0, \lambda x - Ax \geq 0, \langle x, \lambda x - Ax \rangle = 0) \Longleftrightarrow \begin{cases} y = \lambda x - Ax, \\ x = \max(0, x - \rho y). \end{cases}$$

Unfortunately, the max function is not differentiable, it, however, can be smoothed in the following way

$$\max(t,s) \sim f_\mu(s,t) = \mu \ln(e^{s/\mu} + e^{t/\mu}) \text{ when } \mu \to 0.$$

More precisely, we have the following proposition:
Proposition 3.3 $|f_\mu(s,t) - \max(s,t)| \leq \mu \ln(2)\ \forall \mu > 0, s \in \mathbb{R}, t \in \mathbb{R}$.

Proof. Without loss of generality, we assume that $s \geq t$. Since $\mathbb{R} \ni x \mapsto e^x$ and $\text{int}(\mathbb{R}_+) \ni x \mapsto \ln(x)$ are increasing, we have

$$\begin{aligned}
|f_\mu(s,t) - \max(s,t)| &= \left|\mu \ln(e^{s/\mu} + e^{t/\mu}) - s\right| \\
&= \mu \left|\ln(e^{s/\mu} + e^{t/\mu}) - \ln(e^{s/\mu})\right| \\
&= \mu \left(\ln(e^{s/\mu} + e^{t/\mu}) - \ln(e^{s/\mu})\right) \\
&\leq \mu \left(\ln(2e^{s/\mu}) - \ln(e^{s/\mu})\right) \\
&= \mu \ln(2).
\end{aligned}$$

The proof is thereby completed. ∎

With this observation, we are led to consider the following (uniform) approximation of the equation $x = \max(0, x - \rho y)$

$$x = \mu \ln(1 + e^{(x-\rho y)/\mu}),$$

or

$$x - \mu \ln(1 + e^{(x-\rho y)/\mu}) = 0,$$

where $\rho > 0$ is given.
Set

$$P(X,\mu) = \begin{bmatrix} Ax - \lambda x + y \\ \langle 1_n, x\rangle - 1 \\ x - \mu \ln(1 + e^{(x-\rho y)/\mu}) \end{bmatrix}.$$

As the spirit of a smoothing method, we apply the Newton method to this system so that μ will ultimately be driven to 0. One common option is to consider μ as a parameter while applying the Newton method to $P(X,\mu)$. However, one may face issues with finding good

strategies to update μ after each iteration. In our context, we choose the same strategy as NPIPM, which consists in keeping μ as a variable controlled by the following equation

$$\frac{1}{2}\sum_{i=1}^{n}(\min\{x_i,0\}^2+\min\{y_i,0\}^2)+\mu^2+\epsilon\mu=0,$$

where $\epsilon>0$ is fixed.
Now, the (damped) Newton method can be applied to solve the following system:

$$Q(X,\mu)=\begin{bmatrix} P(X,\mu) \\ \frac{1}{2}\sum_{i=1}^{n}(\min\{x_i,0\}^2+\min\{y_i,0\}^2)+\mu^2+\epsilon\mu \end{bmatrix}=0,$$

where stepsizes will be chosen such that μ will be driven to 0, which means that the sequence μ^k will converge to 0. To this end, the same way of selecting stepsizes as in NPIPM is chosen . As a result, we have a similar result for SM as Proposition 3.1. That is, assume that (X^k,μ^k) is the sequence generated by SM and that $\liminf \alpha^k>0$, where α^k is the sequence of stepsizes. Then, μ^k is a positive decreasing sequence convergent to 0. A simple computation yields the Jacobian matrix of $Q(X^k,\mu^k)$, where $X^k=(x^k,y^k,\lambda^k)$

$$J_Q(X^k,\mu^k)=\begin{bmatrix} A-\lambda^k I_n & I_n & -x^k & 0_{n\times 1} \\ 1_n^T & 0_{1\times n} & 0 & 0 \\ U^k & V^k & 0_{n\times 1} & W^k \\ 0_{1\times n} & 0_{1\times n} & 0 & 2\mu^k+\epsilon \end{bmatrix},$$

where $U=\operatorname{diag}\left(\frac{1}{1+e^{c_1^k}},..,\frac{1}{1+e^{c_n^k}}\right), V=\rho.\operatorname{diag}\left(\frac{e^{c_1^k}}{1+e^{c_1^k}},..,\frac{e^{c_n^k}}{1+e^{c_n^k}}\right)$, and $W^k\in\mathbb{R}^{n\times 1}$ satisfying

$$W_i^k=-\ln\left(1+e^{c_i^k}\right)+c_i^k\frac{e^{c_i^k}}{1+e^{c_i^k}}\quad\forall 1\le i\le n.$$

Here $c_i^k=\dfrac{x_i^k-\rho y_i^k}{\mu^k}$ for all $i=1,2,\ldots,n$.

Assume that (X^k,μ^k) is the sequence generated by SM and that it converges to a solution $(X^*,0)$ of $Q(X,\mu)=0$. We will now provide a condition to ensure the nonsingularity for SM, that is a condition under which $\lim_{k\to\infty} J_Q(X^k,\mu^k)$ is nonsingular.

Proposition 3.4 *Assume that (X^k,μ^k) is the sequence generated by SM and that it converges to a solution $(X^*,0)$ of $Q(X,\mu)=0$, where $X^*=(x^*,y^*,\lambda^*)$. Then, $\lim_{k\to\infty} J_Q(X^k,\mu^k)$*

is nonsingular if (x^*, y^*) *satisfies the strict complementarity, i.e.,* $x_i^* + y_i^* > 0$ *for every* $i = 1, 2, ..., n$, *and the principle submatrices* $\tilde{A}_{\alpha\alpha}$ *of* $\tilde{A} = A - \lambda^* I_n$, *where* $\alpha = \{1 \leq l \leq n : x_l^* \neq 0\}$, *satisfy*

$$\text{for all } y \in \mathbb{R}^{|\alpha|} \setminus \{0\}, \text{ if } \tilde{A}_{\alpha\alpha} y \in \operatorname{span}(x_\alpha^*), \text{ then } \sum_{i=1}^{|\alpha|} y_i \neq 0.$$

Proof. Denote by

$$\alpha = \{1 \leq l \leq n : x_l^* \neq 0\} \text{ and } \beta = \{1 \leq l \leq n : y_l^* \neq 0\}.$$

Due to the strict complementarity assumption and the fact that $\mu^k \to 0$, we have

$$\lim_{k\to\infty} c_i^k = \begin{cases} +\infty & \text{if} \quad i \in \alpha, \\ -\infty & \text{if} \quad i \in \beta, \end{cases}$$

and therefore

$$\lim_{k\to\infty} e^{c_i^k} = \begin{cases} +\infty & \text{if} \quad i \in \alpha, \\ 0 & \text{if} \quad i \in \beta. \end{cases}$$

This leads to $U^* = \lim_{k\to\infty} U^k \in \mathbb{R}^n$, $V^* = \lim_{k\to\infty} V^k \in \mathbb{R}^n$ and that U^* and V^* are the diagonal matrices satisfying

$$U_{ii}^* = \begin{cases} 0 & , \text{if} \quad i \in \alpha, \\ 1 & , \text{if} \quad i \in \beta \end{cases} \quad \text{and} \quad V_{ii}^* = \begin{cases} \rho & , \text{if} \quad i \in \alpha, \\ 0 & , \text{if} \quad i \in \beta. \end{cases}$$

It is clear that the nonsingularity of $\lim_{k\to\infty} J_Q(X^k, \mu^k)$ is equivalent to that of

$$\lim_{k\to\infty} \begin{bmatrix} A - \lambda I_n & I_n & -x^k \\ 1_n^T & 0_{1\times n} & 0 \\ U^k & V^k & 0_{n\times 1} \end{bmatrix} = \begin{bmatrix} A - \lambda I_n & I_n & -x^* \\ 1_n^T & 0_{1\times n} & 0 \\ U^* & V^* & 0_{n\times 1} \end{bmatrix},$$

which can be proved by the same arguments given in the proof of Proposition 3.2. ∎

3.4 Numerical tests

Our first comment concerns the choice of initial points. First we present how initial points for the three methods including MPCM, NPIPM and SM are chosen. A random vector $\xi \in \mathbb{R}^n$ is first chosen with the uniform distribution on $(0, 1]^n$. After that, we set

$$\begin{aligned} x^0 &= \frac{\xi}{\langle 1_n, \xi \rangle}, \\ \lambda^0 &= \frac{\langle x^0, Ax^0 \rangle}{\langle x^0, x^0 \rangle}, \\ \mu^0 &= 10^{-2} \text{ fixed (only for NPIPM and SM).} \end{aligned}$$

For y^0, we initially assign $y^0 = \lambda^0 x^0 - Ax^0$ and then replace any nonpositive component of y^0 by 0.01.

Regarding LPM, after getting a random vector $\xi \in \mathbb{R}^n$ with the uniform distribution on $[0, 1]^n$, we set

$$\begin{aligned} x^0 &= \frac{\xi}{\langle 1_n, \xi \rangle}, \\ \tilde{y}^0 &= Ax^0, \\ \lambda^0 &= \frac{\langle x^0, Ax^0 \rangle}{\langle x^0, x^0 \rangle}, \end{aligned}$$

Our second comment is that we use the built-in backslash function of Matlab for solving the linear system of equations at each iteration. Finally, for all the considered algorithms, a solution $(\lambda, x) \in \mathbb{R} \times \mathbb{R}^n$ is claimed to be found when the following conditions are satisfied

$$\begin{aligned} &\| \min(x, \lambda x - Ax) \|_2 \leq 10^{-8}, \\ &\|x\|_2 > 10^{-6}, \end{aligned}$$

where the min function is carried out componentwisely.

Remark 3.6 In the way of choosing initial points presented above, we first take x^0 so that the condition $\langle 1_n, x^0 \rangle = 1$ is satisfied, and then select λ^0 as if it is a Pareto eigenvalue corresponding to x^0 and, as we can see, a necessary condition for that is $\lambda^0 = \frac{\langle x^0, Ax^0 \rangle}{\langle x^0, x^0 \rangle}$.

Remark 3.7 Hereafter, whenever numerical experiments are conducted, this pattern of choosing initial points will be applied.

3.4.1 Testing on special matrices

The first numerical experiment is given by taking matrices of order 3, 4 and 5 that are known to have 9, 23 and 57 Pareto eigenvalues, respectively.

$$A_1 = \begin{bmatrix} 5 & -8 & 2 \\ -4 & 9 & 1 \\ -6 & -1 & 13 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 132 & -106 & 18 & 81 \\ -92 & 74 & 24 & 101 \\ -2 & -44 & 195 & 7 \\ -21 & -38 & 0 & 230 \end{bmatrix}$$

and

$$A_3 = \begin{bmatrix} 788 & -780 & -256 & 156 & 191 \\ -548 & 862 & -190 & 112 & 143 \\ -456 & -548 & 1308 & 110 & 119 \\ -292 & -374 & -14 & 1402 & 28 \\ -304 & -402 & -66 & 38 & 1522 \end{bmatrix}.$$

We compare MPCM, NPIPM, LPM, and SM by computing the average number of iterations, computing time and percentage of failures.

Remark 3.8 In our case, a failure is declared if the number of iterations exceeded 100 or the Jacobian matrix is ill-conditioned according to Matlab's criterion.

The comparison results are summarized in Table 3.1 where "Iter" denotes the average number of iterations and "Failure (%)" represents the percentage of failures to find a solution of the corresponding EiCP.

We note that 7×10^3 initial points have been used to find all the Pareto eigenvalues of A_3 simultaneously by MPCM, NPIPM and LPM. SM shows its inefficiency in finding many solutions when it only finds around 50 eigenvalues of A_3 despite being run with 10^5 initial points. A quick look at the table clearly reveals that LPM performs best among all the considered solvers and that NPIPM and LPM are the most robust with respect to initial points.

Methods	A_1		A_2		A_3	
	Iter	Failure (%)	Iter	Failure (%)	Iter	Failure (%)
MPCM	6	8	8	5	7	0.6
NPIPM	9	0	10	0.6	8	0.2
LPM	6	0	7	0	7	0
SM	8	33	8	44	9	46

Table 3.1: Comparison of the 4 solvers on the matrices A_1, A_2 and A_3

Remark 3.9 We compare the number of iterations because the computational effort in each iteration of all solvers is almost the same.

3.4.2 Performance Profiles

In this section, we compare the 4 solvers that have been defined in Section 3.1 and Section 3.2. In order to complete this experiment, we choose the *performance profiles* developed by E. D. Dolan and J. J. Moré [72] as a tool for comparing the solvers. The performance profiles give for each $t \in \mathbb{R}$, the proportion $\rho_s(t)$ of test problems on which each solver under comparison has a performance within the factor t of the best possible ratio.
Average computing time, average number of iterations, percentage of failure and maximum number of eigenvalues found by each solver are used as performance measures to compare these algorithms.
Due to the absence of a library dedicated to EiCP, we have chosen a set P of 40 random matrices for this test. Let S be the set of the four solvers that will be compared. The *performance ratio* is defined by

$$r_{p,s} = \frac{t_{p,s}}{\min\{t_{p,s} : s \in S\}},$$

where $p \in P$, $s \in S$, and $t_{p,s}$ is either

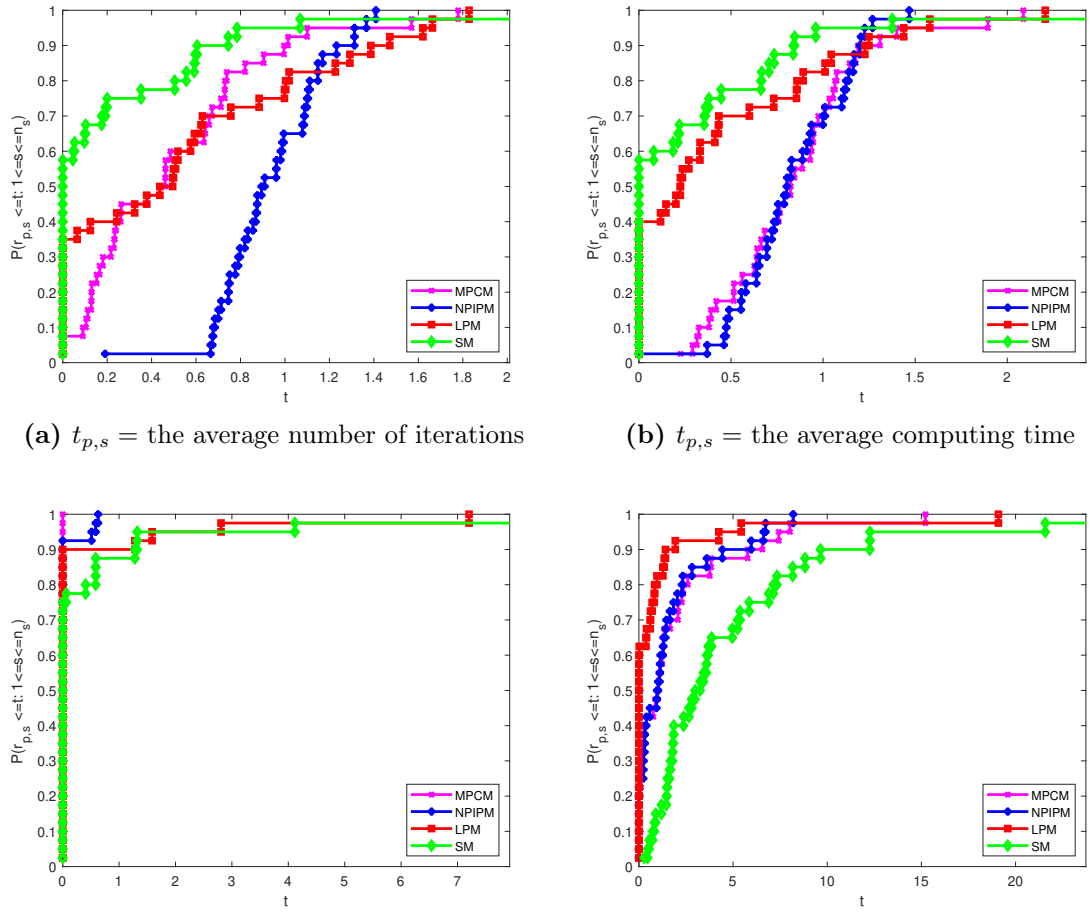
- the average number of iterations required to solve problem p by solver s corresponding to Figure (a), or
- the maximum number of solutions corresponding to Figure (b), or
- the percentage of failure (in the sense of Remark 3.8) corresponding to Figure (c), or
- the average computing time corresponding to Figure (d).

The performance of the solver $s \in S$ is defined by

$$\rho_s(t) = \frac{1}{n_p}\text{card}\{p \in P \ : \ \log_2(r_{p,s}) \leq t\},$$

where, n_p is the number of problems, and t is a real factor. The numerical experiments are conducted on an ordinary computer. All the program codes are written and executed in *Matlab 9.6*.

Figure 1a presents the performance profiles of the four solvers with the criterion: the average number of iterations that each solver takes to find a solution. We observe that SM and LPM have the most number of wins, dominating MPCM and NPIPM. Also, on the considered interval, SM outperforms other solvers, followed by LPM, MPCM and NPIPM respectively.



(a) $t_{p,s}$ = the average number of iterations

(b) $t_{p,s}$ = the average computing time

(c) $t_{p,s}$ = the maximum number of solutions found

(d) $t_{p,s}$ = the percentage of failures

Figure 3.1: The performance profiles of MPCM, NPIPM, LPM and SM.

Figure 1*b* presents the performance profiles of the four solvers corresponding to the average computing time. We can see that relative to this criterion, SM is the best solver, closely followed by LPM, while NPIPM and MPCM have no wins. Another interesting point is that on the interval [0.5, 1.5] the performances of NPIPM and MPCM are quite competitive compared to the others.

Figure 1*c* displays the performance profiles of the four solvers when considering the maximum number of solutions found by each one. MPCM can solve 100 % of the problems with the greatest efficiency and has the most number of wins, followed respectively by NPIPM, LPM and SM. As in the previous test on the three given matrices, SM shows that its ability to find many solutions is the worst among all the methods.

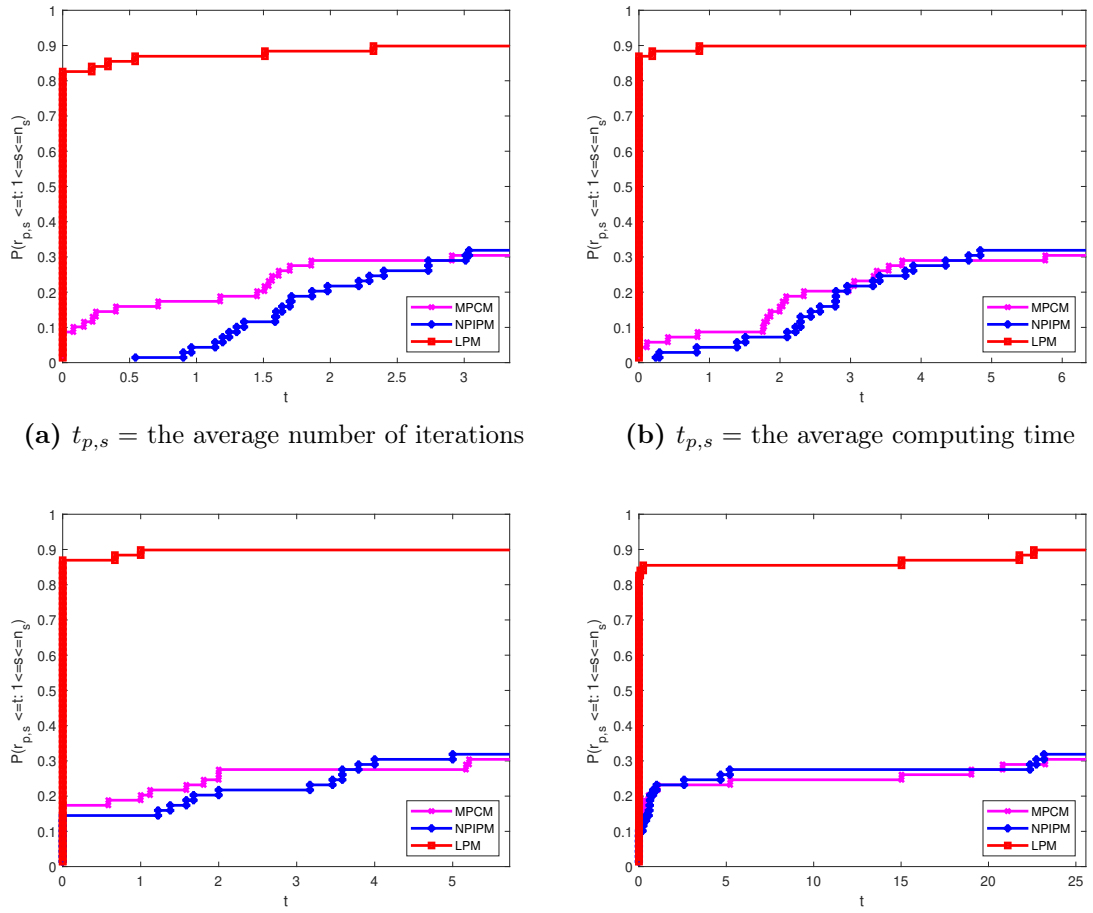
In Figure 1*d*, we depicted the performance profiles of the four solvers for the percentage of

failure. LPM encounters the least number of failures among all the methods while the performances of the MPCM and NPIPM are quite the same in this regard. With respect to this criterion, SM is not the winner on any problem and performs the worst.

In conclusion, SM proved to be the best solver when it comes to the average number of iterations and computing time while the situation with it is completely reversed with respect to other criteria. Concerning the percentage of failure, LPM ranks first. The performances of MPCM and NPIPM are roughly equivalent regarding all criteria except the average number of iterations where MPCM performs better. MPCM can find the most number of solutions among all the methods.

We have made various numerical experiments and realized that SM could only solve problems of small size. Accordingly, we now conduct another comparison only between MPCM, NPIPM and LPM on a set of problems of larger sizes. We have chosen a set of 30 square matrices with an average size of 131, all of them taken from the *Matrix Market* [1].

[1] https://math.nist.gov/MatrixMarket/



(a) $t_{p,s}$ = the average number of iterations

(b) $t_{p,s}$ = the average computing time

(c) $t_{p,s}$ = the maximum number of solutions found

(d) $t_{p,s}$ = the percentage of failures

Figure 3.2: The performance profiles of MPCM, NPIPM and LPM.

Looking at Figure 3.2 where we present the performance profiles of the three solvers MPCM, NPIPM and LPM on a set of problems of larger size, we can conclude that LPM is the best one with the most number of wins regardless of the criterion considered. In terms of the average number of iterations, MPCM wins over NPIPM on the given interval. Regarding the maximum number of solutions and the average computing time, we can see that MPCM performs slightly better than NPIPM. We note that LPM is a robust solver, respectively followed by NPIPM and MPCM.

3.5 Partially constrained eigenvalue problems

In this section, we consider a class of problems called partially constrained eigenvalue problems. As its name suggests, in this class of problems, only a portion of the unknown x is cone-constrained while the remaining components of x are free. Assume the first

m components of x are nonnegative, and the other ones are not restricted. In this case, x can be written with two splitting parts as follows

$$x = \begin{bmatrix} x_c \\ x_f \end{bmatrix},$$

where x_c is the m-dimensional block vector containing the first m components of x and x_f is the remaining part which, of course, belongs to $\mathbb{R}^{n-m}$. As a result, it turns out that we are dealing with a cone-constrained eigenvalue problem constrained by the following convex cone

$$K_{m,n-m} = \mathbb{R}^m_+ \times \mathbb{R}^{n-m}.$$

More precisely, the problem now is to find a non zero $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ such that

$$K_{m,n-m} \ni x \perp (\lambda x - Ax) \in K^*_{m,n-m}. \tag{3.31}$$

Because $K^*_{m,n-m} = \mathbb{R}^n_+ \times \{0\}$, we can express the cone-constrained eigenvalue problem corresponding to the convex cone $K_{m,n-m}$ as follows

$$x_c \geq 0, \qquad x_f \text{ is free}, \quad x \neq 0, \tag{3.32}$$

$$\lambda x_c - Ax_c \geq 0, \qquad \lambda x_f - Ax_f = 0, \tag{3.33}$$

$$\langle x_c, \lambda x_c - A\lambda x_c \rangle = 0. \tag{3.34}$$

In reality, some applications related to solving boundary value problems by boundary integral equation methods can lead to this kind of problem, see [124, Example 1] for an instance in mechanics. We will now present a partially cone-constrained eigenvalue problem arising from a geomechanical fractures problem.
Let $N \geq 1$ be an integer and A be a $3N \times 3N$ matrix with real entries. We are interested

in finding $\lambda \in \mathbb{R}$ such that there exists a $3N$-vector u satisfying

$$\begin{aligned}
(\lambda u - Au)_1 &= 0,\\
(\lambda u - Au)_2 &= 0,\\
0 \leq u_3 \perp (\lambda u - Au)_3 &\geq 0,\\
(\lambda u - Au)_4 &= 0,\\
(\lambda u - Au)_5 &= 0,\\
0 \leq u_6 \perp (\lambda u - Au)_6 &\geq 0,\\
&\vdots\\
(\lambda u - Au)_{3N-2} &= 0,\\
(\lambda u - Au)_{3N-1} &= 0,\\
0 \leq u_{3N} \perp (\lambda u - Au)_{3N} &\geq 0,\\
u \in \mathbb{R}^N \setminus \{0\}&,
\end{aligned} \tag{3.35}$$

where $(Au - \lambda u)_i$ denotes the i-th component of $Au - \lambda u$.

We can see that this system of equations is not exactly the problem of the form (3.31). However, by some simple reformulation, it can be transformed into a cone-constrained eigenvalue problem corresponding to the convex cone $K_{N,2N} = \mathbb{R}^N_+ \times \mathbb{R}^{2N}$. Indeed, there always exists a permutation σ, which is a bijective from $\{1, 2, .., 3N\}$ into itself, such that

$$\sigma(\{1, 2, .., N\}) = \{3i : i = 1, 2, .., N\}.$$

Denote $I = \{\sigma(1), \sigma(2), .., \sigma(N)\}$ and $J = \{\sigma(N+1), \sigma(2), .., \sigma(3N)\}$. Set $\tilde{u}_i = u_{\sigma(i)}$ and

$$\tilde{A} = \begin{bmatrix} A_{II} & A_{IJ} \\ A_{JI} & A_{JJ} \end{bmatrix}.$$

It can be seen that solving system (3.35) is equivalent to solving the following system of equations

$$\begin{aligned} 0 \leq \tilde{u}_1 \perp (\lambda\tilde{u} - \tilde{A}\tilde{u})_1 \geq 0, \\ \vdots \\ 0 \leq \tilde{u}_N \perp (\lambda\tilde{u} - \tilde{A}\tilde{u})_N \geq 0, \\ (\lambda\tilde{u} - \tilde{A}\tilde{u})_{N+1} = 0, \\ (\lambda\tilde{u} - \tilde{A}\tilde{u})_{N+2} = 0, \\ \vdots \\ (\lambda\tilde{u} - \tilde{A}\tilde{u})_{3N} = 0, \\ \tilde{u} \in \mathbb{R}^N \backslash \{0\}, \end{aligned} \tag{3.36}$$

which is a cone-constrained eigenvalue problem corresponding to the convex cone $K_{N,2N}$.

Now we give some numerical results for problem (3.35) by using the three methods including MPCM, NPIPM and LPM. Data for this numerical result are from IFP Energies Nouvelles (IFPEN). We consider 6 problems with N ranging from $2, 3, 15, 61, 500$ to 1500, which means that the maximum size of the matrix A we have to deal with is 4500.
First of all, in order to use NPIPM, we reformulate (3.35) using slack and relaxation variables like it was done previously: Find $(u, w, \lambda, \mu) \in \mathbb{R}^{3N} \times \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^+$ such that

$$\begin{aligned} Au - \lambda u + g(w) = 0, \\ u_3 w_1 - \mu = 0, \\ \vdots \\ u_{3N} w_N - \mu = 0, \\ \sum_{i=1}^{3N} u_i^2 - 1 = 0, \\ \frac{1}{2}\sum_{i=1}^{N}(\min\{u_{3i}, 0\}^2 + \min\{w_i, 0\}^2) + \mu^2 + \epsilon\mu = 0, \end{aligned} \tag{3.37}$$

where $g(w) = [0, 0, w_1, 0, 0, w_2, ..., 0, 0, w_N]^T$, and $\epsilon > 0$ is fixed. The normalization made for u in the above system is to prevent it from being identical to zero. We apply NPIPM to the system (3.37). On the other hand, we can also use MPCM to solve this kind of problem by doing exactly what has been described for MPCM in Section 3.2 with a difference that

in this case the function F in (3.22) would be substituted by

$$F(u,w) = \begin{bmatrix} Au - b + g(w) \\ u_3 w_1 \\ \vdots \\ u_{3N} w_N \end{bmatrix}.$$

Table 3.2 compares the 3 methods, namely MPCM, NPIPM and LPM on Problem

N	MPCM				NPIPM				LPM			
	N_{max}	Iter	T(s)	F (%)	N_{max}	Iter	T(s)	F (%)	N_{max}	Iter	T(s)	F (%)
2	5	12	0.0005	5	5	22	0.0007	0.2	5	9	0.0003	42
3	6	12	0.0005	12	6	20	0.0008	5.7	4	10	0.0003	62
15	5	11	0.001	4	4	19	0.002	1.6	3	9	0.0008	32
61	21	16	0.03	19	17	21	0.03	4.2	16	24	0.03	54
500	4	12	3.3	2	3	21	5.3	0.5	4	15	2.5	67
1500	30	20	48	39	16	29	34	3	14	32	34	94

Table 3.2: Comparison of MPCM, NPIPM and LPM on the 6 problems

(3.35) with 6 matrices from IFPEN, where "N_{max}" denotes the number of solutions found, "Iter"and "T" denote the average number of iterations and computing time respectively while "F" denotes the percentage of failure. We observe that all the methods converge for the 6 matrices in the sense that they can all manage to find at least a solution, especially in the cases $N = 500$ and $N = 1500$. Another point that can be observed from Table 3.2 is that NPIPM is the most robust solver among the three with respect to initial points. In terms of computing time, LPM outperforms the others, while MPCM and NPIPM have roughly the same computing time except for the last case where NPIPM is quite faster. With respect to the average number of iterations, we can say in general that MPCM performs best, followed respectively by LPM and NPIPM. We point out that LPM experienced many failures and found the fewest number of solutions. Finally, it is clear that MPCM found the most solutions compared to the others.

Remark 3.10 The number of failures of LPM is rather high and tends to increase with the problem's size. This event could be explained by the fact that bad conditioning during the computation of the projection could occur in concrete situations. On the other hand, interior point methods, particularly NPIPM do not suffer from this drawback because we are in the interior.

3.6 Extension of MPCM and NPIPM for solving quadratic pencils under conic constraints

MPCM and NPIPM can be adjusted so as to deal with more general cone-constrained eigenvalue problems. For simplicity, we consider the quadratic pencil eigenvalue complementarity problem presented as follows.
Given a triplet (A_0, A_1, A_2) of three matrices of size $n \times n$, we define the corresponding quadratic pencil as follows

$$M(\lambda) = A_0 + \lambda A_1 + \lambda^2 A_2.$$

Then, the quadratic pencil eigenvalue complementarity problem corresponding to the quadratic pencil $M(\lambda)$ is the problem of finding $\lambda \in \mathbb{R}$ and $x \in \mathbb{R}^n \setminus \{0\}$ such that

$$0 \leq x \perp M(\lambda)x \geq 0. \tag{3.38}$$

Problem (3.38) can be reformulated into

$$\begin{aligned} M(\lambda)x - y &= 0, \\ \langle 1_n^T, x\rangle - 1 &= 0, \\ x \geq 0, \quad y &\geq 0. \end{aligned} \tag{3.39}$$

Now we can see that MPCM and NPIPM can be applied in an attempt to solve the system (3.39).
Remark 3.11 Let us observe that MPCM and NPIPM not only can be applied to solve the quadratic pencil eigenvalue complementarity problem but might be applicable for solving similar problems in which the quadratic pencil in (3.38) is substituted by a matrix pencil of order m, with $m > 2$.
Consider the following quadratic pencil

$$M(\lambda) = \lambda^2 \begin{bmatrix} 2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 10 \end{bmatrix} + \lambda \begin{bmatrix} 7 & 0 & 0 \\ 0 & 30 & 0 \\ 0 & 0 & 20 \end{bmatrix} + \begin{bmatrix} -2 & 6 & 0 \\ 2 & 16 & 3 \\ 0 & 5 & 0 \end{bmatrix}. \tag{3.40}$$

Solving the corresponding problem associated with this quadratic pencil by MPCM and NPIPM with a sample of 104 random initial points gives 12 solutions as shown in Table 3.3.
Table 4 summarizes the results of MPCM and NPIPM on the problem corresponding to the pencil.(3.40).

Pareto eigenvalue	Pareto eigenvector			Dual vector		
	x_1	x_2	x_3	y_1	y_2	y_3
$\lambda_1 = -4.3930$	0	1	0	6	0.0000	5
$\lambda_2 = -3.7656$	1	0	0	0	2	0
$\lambda_3 = -3.6524$	0.8712	0.1288	0	0	0	0.6438
$\lambda_4 = -2.0000$	0	0	1	0	3	0
$\lambda_5 = -1.9613$	0	0.1318	0.8682	0.7909	0	0
$\lambda_6 = -1.9580$	0.0954	0.1277	0.7769	0	0	0
$\lambda_7 = -0.7689$	0.3877	0.4006	0.2116	0	0	0
$\lambda_8 = -0.6986$	0.5036	0.4964	0	0	0	2.4820
$\lambda_9 = -0.6820$	0	0.6426	0.3574	3.8554	0	0
$\lambda_{10} = -0.6070$	0	1	0	6	0	5
$\lambda_{11} = 0.0000$	0	0	1	0	3	0
$\lambda_{12} = 0.2656$	1	0	0	0	2	0

Table 3.3: Solutions of the problem corresponding to the pencil (3.40) solved by MPCM and NPIPM

Methods	N_{max}	Iter	Failure (%)
MPCM	12	8	9.7
NPIPM	12	9	0.2

Table 3.4: Comparison between MPCM and NPIPM on the pencil (3.40)

It can be seen that while the number of solutions found and the average number of iterations of MPCM and NPIPM are roughly the same, the percentage of failure of NPIPM is less than MPCM, which means that NPIPM is a more robust method with respect to initial points. This has been seen when we carried out the numerical experiments in Section 3.5.

For a given quadratic pencil (3.38) under conic constraints, a standard approach consists in using a reduction technique. More precisely, problem (3.38) can be reduced into an affine pencil as follows: Find $\lambda \in \mathbb{R}$ and $\begin{bmatrix} x \\ u \end{bmatrix} \in \mathbb{R}^{2n} \setminus \{0\}$ such that

$$\mathbb{R}^n_+ \times \mathbb{R}^n \ni \begin{bmatrix} x \\ u \end{bmatrix} \perp \begin{bmatrix} A_0 & A_1 \\ 0_{n\times n} & I_n \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} + \lambda \begin{bmatrix} 0_{n\times n} & A_2 \\ -I_n & 0_{n\times n} \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} \in \mathbb{R}^n_+ \times \{0\}^n . \tag{3.41}$$

This equivalence can be seen by first rewriting (3.38) as follows

$$\begin{aligned} u &= \lambda x, \\ 0 \leq x \perp A_0 x + A_1 u + \lambda A_2 u &\geq 0. \end{aligned} \tag{3.42}$$

It is clear that expressing (3.42) in the matrix form gives (3.41).
Using MPCM and NPIPM for solving the equivalent problem (3.41) with the data in (3.40), we get the following table

Methods	N_{max}	Iter	Failure (%)
MPCM	6	15	63
NPIPM	6	13	37

Table 3.5: Comparison between MPCM and NPIPM with the data (3.40)

Looking at Table 3.5, we see that with the reformulated problem (3.41), MPCM and NPIPM take more iterations to reach a solution and experience much more failures than they do with the initial problem (3.38). Furthermore, in this case both methods can find only 6 solutions given the same number of initial points. Regarding the number of failures, NPIPM is seen to be more robust.

3.7 Conclusions

In this chapter, we have considered two interior-point methods for solving eigenvalue complementarity problems. We have also presented an application of the methods to a geomechanical problem. Numerical experiments have been made to compare the performances of MPCM and NPIPM relative to two other common methods, namely LPM and SM. NPIPM has proved to be an efficient and robust method for solving eigenvalue complementarity problems, especially through its display on the geomechanical fracture problem. More precisely, we can observe that its percentage of failure remains very low in all six problems. Particularly in the case where the problem's size is 4500, its percentage of failure is around 3%, which is a very appealing property. The performance of MPCM is generally equivalent to that of NPIPM except in terms of robustness. There are some open questions regarding the use of MPCM and NPIPM or interior point methods in general for solving eigenvalue complementarity problems. The first issue could be the way to select efficiently initial points. It would be also interesting to consider other interior point methods and compare their performances to MPCM and NPIPM in the context of eigenvalue complementarity problems. As well, inexact Newton methods could be of great interest to use in our context. This would allow us to solve large scale problems. The question of local and global convergence of our algorithms for the Pareto eigenvalue

complementarity problem is open and needs further investigation. These points are out of the scope of the current dissertation and will be the subject of a future research project.

4

Solving inverse Pareto eigenvalue problems

Contents

This chapter covers the material discussed in the published paper [14], which was produced in collaboration with S. Adly

4.1 Introduction

Recall that the Pareto Eigenvalue Problem (PEP) corresponding to an $n \times n$ matrix A consists in finding a scalar $\lambda \in \mathbb{R}$ such that the linear complementarity system

$$x \geq 0_n, \quad \lambda x - Ax \geq 0_n, \quad \langle x, \lambda x - Ax \rangle = 0$$

admits a nonzero solution $x \in \mathbb{R}^n$. Such an x is called a Pareto eigenvector of A corresponding to the Pareto eigenvalue λ. We have used some standard notations here: the symbol 0_n refers to the n-dimensional zero vector, $x \geq 0_n$ indicates that all components of x are nonnegative, and $\langle \cdot, \cdot \rangle$ denotes the usual inner product of $\mathbb{R}^n$. The concept of Pareto eigenvalue arises naturally in a variety of applications such as the dynamic analysis of structural mechanical systems, vibro-acoustic systems, electrical circuit simulation, signal processing, fluid dynamic, contact problems in mechanics (see for instance [105–108, 123]). It should be noted that finding all Pareto eigenvalues of a medium or high order matrix can be challenging since the number of Pareto eigenvalues grows exponentially with the order of the matrix A, see [136]. For instance, a matrix of order 20 could have more than 1.5 million Pareto eigenvalues. A rich variety of algorithms for computing Pareto eigenvalues have been proposed in [13,15–17,88,90–92,96,102,116,129], just to mention a few references.

On the other hand, in this chapter we are interested in the Inverse Pareto Eigenvalue Problem (IPEP) which consists in constructing a matrix $A \in \mathcal{M}_n(\mathbb{R})$ whose Pareto spectrum contains a prescribed set $\Theta = \{\lambda_1, ..., \lambda_p\}$ of distinct reals. Here, the Pareto spectrum of A refers to the set of all Pareto eigenvalues of A and $\mathcal{M}_n(\mathbb{R})$ denotes the set

of matrices of order n with real entries. If denoting by $X = [x_1, ..., x_p]$ the rectangular matrix containing the (unknown) Pareto eigenvectors of A and by $\vec{\lambda} = (\lambda_1, ..., \lambda_p)^T$ the vector containing the given Pareto eigenvalues, we can present IPEP abstractly as

$$\begin{cases} X \geq \mathbf{O}, \quad X\text{diag}(\vec{\lambda}) - AX \geq \mathbf{O}, \quad \langle X, X\text{diag}(\vec{\lambda}) - AX \rangle = 0, \\ \text{Each column of } X \text{ is not identical to the zero vector,} \end{cases}$$

where $\mathbf{O}$ is a zero matrix of appropriate size, $\langle X, Y \rangle = \text{tr}(X^T Y)$ is the trace inner product, and $\text{diag}(\vec{\lambda})$ refers to the diagonal matrix whose diagonal is $\vec{\lambda}$. It should be noted that if one solution to the above problem is found, one can produce many others by left-right permutation operations, see [76] for details. There are other inverse Pareto eigenvalue problems different from what we consider in this note. For instance, in [137], the authors have considered the problem of finding a matrix A of order n whose Pareto spectrum is exactly a prescribed set. However, they only studied the solvability of such the problem but put aside the numerical resolution.

In this note, we will formulate the inverse Pareto eigenvalue problem under consideration as different systems of nonlinear equations. The resulting systems can be smooth or nonsmooth depending on how it is formulated. Newton-type methods can then be employed to solve such nonlinear systems. The IPEP can also be formulated as nonlinear optimization problems, see for instance [61]. In principle, there is no relation between the order of the matrix A and the cardinality of the target set $\Theta = \{\lambda_1, ..., \lambda_p\}$. However, to avoid the overdetermination for the resulting system (which will be made precise later, see Remark 4.1), the authors have intentionally assumed that p should not exceed n^2. Apparently, the case in which $p \leq n$ is not of interest because the IPEP can then be solved explicitly by taking as matrix A any diagonal matrix that contains the λ_k's on its diagonal. Even more, in our numerical tests, we have avoided the case in which p is greater than the triangular number $\tau_n := n(n+1)/2$. This is due to the fact that it has been shown in [76] that if $p \leq \tau_n$, then the IPEP is solvable with a matrix of order n for an arbitrary sample of cardinality p. By contrast, the case in which $p > \tau_n$ has not been well understood yet and may need further investigations.

The rest of this chapter is organized as follows. In Section 4.2, we respectively formulate the problem as two smooth nonlinear systems of equations; the first one contains complementarity conditions and we adapt the Mehrotra predictor corrector method [110] to address it; the second system is yielded with the help of the "squaring trick" and therefore does not impose any constraint on the variables; as a result, it is then solved by a Newton type method. In Section 4.3, we first present two methods based on complementarity function techniques, namely SNM_{min} and SNM_{FB}, and then the lattice projection method proposed in [15]. We conduct some numerical experiments in Section 4.4 to compare the performances of all the methods with respect to the average number of iterations and the

percentage of failures. Section 4.5 is devoted to the extension of MPCM, ST, $\text{SNM}_{\min}$ and SNM_{FB} to inverse quadratic pencil eigenvalue complementarity problems. We end the chapter with some concluding remarks and perspectives in Section 4.6.

4.2 Smooth approach

4.2.1 The Mehrotra Predictor Corrector Method (MPCM)

Mathematically speaking, the problem at hand is that of solving a system of the form

$$\begin{cases} x_k \geq 0_n, \quad \lambda_k x_k - A x_k \geq 0_n, \quad \langle x_k, \lambda_k x_k - A x_k \rangle = 0 \\ \\ \langle 1_n, x_k \rangle = 1 \end{cases} \qquad \text{for all } k \in \{1, .., p\}, \tag{4.1}$$

where the unknown variables are the columns of the matrix $A \in \mathcal{M}_n(\mathbb{R})$ and the vectors $x_1, ..., x_p \in \mathbb{R}^n$, and 1_n denotes the n dimensional vector of all ones. The last equation is added to ensure that x_k is not identical to the zero vector.

By introducing the slack variables $y_k = \lambda_k x_k - A x_k$, we reformulate (4.1) as

$$\begin{cases} x_k \geq 0_n \\ y_k \geq 0_n \\ \langle x_k, y_k \rangle = 0 \\ (A - \lambda_k I_n) x_k + y_k = 0_n \\ \langle 1_n, x_k \rangle - 1 = 0 \end{cases} \qquad \text{for all } k \in \{1, .., p\}, \tag{4.2}$$

The Mehrotra predictor corrector method was first proposed in 1989 by Sanjay Mehrotra [110], as a variant of the primal-dual interior point method for optimization problems. Because of their efficiency, most of today's interior-point general-purpose software for linear and nonlinear programming is based on predictor-corrector algorithms like the one of Mehrotra. Here we adapt this method to deal with the IPEP of the form (4.2). Let

us define $F : \mathbb{R}^{np} \times \mathbb{R}^{np} \times \mathbb{R}^{n^2} \longrightarrow \mathbb{R}^{2np+p}$ as

$$F(Z) = \begin{bmatrix} Ax_1 - \lambda_1 x_1 + y_1 \\ \vdots \\ Ax_p - \lambda_p x_p + y_p \\ \langle 1_n, x_1 \rangle - 1 \\ \vdots \\ \langle 1_n, x_p \rangle - 1 \\ \hat{X} \bullet \hat{Y} \end{bmatrix}, \quad Z = (\hat{X}, \hat{Y}, \hat{A}), \tag{4.3}$$

where

$$\begin{cases} \hat{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \in \mathbb{R}^{np}, \quad \hat{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} \in \mathbb{R}^{np}, \quad \hat{A} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^{n^2}, \\ x_i \in \mathbb{R}^n, \quad y_i \in \mathbb{R}^n, \quad a_i \in \mathbb{R}^n, \quad A = [a_1, ..., a_n] \in \mathcal{M}_n(\mathbb{R}). \end{cases}$$

Here, the notation $p \bullet q$ denotes the Hadamard product of $p \in \mathbb{R}^m$ and $q \in \mathbb{R}^m$. That is, $(p \bullet q)_i = p_i q_i$ for all $i \in \{1, ..., m\}$.

We can easily show that the Jacobian matrix of F can be expressed in terms of block matrices as

$$J_F(Z) = \left[\begin{array}{c|c|c} M & I_{np} & X^T \otimes I_n \\ \hline I_p \otimes 1_n^T & 0_{p \times np} & 0_{p \times n^2} \\ \hline \mathrm{diag}(\hat{Y}) & \mathrm{diag}(\hat{X}) & 0_{np \times n^2} \end{array}\right],$$

where M is the block diagonal matrix whose i-th diagonal block is $A - \lambda_i I_n$, $X = [x_1, ..., x_p]$, $\otimes$ stands for the tensor product, and $\vec{\lambda} = (\lambda_1, ..., \lambda_p)^T$

The description of this method reads as follows. First we choose initial points such that $\hat{X}^0 > 0_{np}, \hat{Y}^0 > 0_{np}, \hat{A}^0 \in \mathbb{R}^{n^2}$. Set $Z^k = (\hat{X}^k, \hat{Y}^k, \hat{A}^k)$ and let $k = 0$. At the predictor step, MPCM first computes the affine scaling (predictor) direction d_a^k, which is given by the least norm solution of the underdetermined linear system $J_F(Z^k) d_a^k = -F(Z^k)$, and then compute a step size $\alpha_a^k \in (0, 1]$ that ensures

$$\hat{X}^k + \alpha_a^k d\hat{X}_a^k > 0_{np}, \tag{4.4}$$

$$\hat{Y}^k + \alpha_a^k d\hat{Y}_a^k > 0_{np}, \tag{4.5}$$

where

$$d_a^k = \begin{bmatrix} d\hat{X}_a^k \\ d\hat{Y}_a^k \\ d\hat{A}_a^k \end{bmatrix} \in \mathbb{R}^{np} \times \mathbb{R}^{np} \times \mathbb{R}^{n^2}.$$

Now, the algorithm uses the information from the predictor step to compute the corrector direction d_c^k by finding the least norm solution of the following linear system

$$J_F(Z^k)d_c^k = -F(Z^k) + B^k, \text{ with } B^k = \begin{pmatrix} 0_{(np+p)} \\ \mu^k 1_{np} - d\hat{X}_a^k \bullet d\hat{Y}_a^k \end{pmatrix}, \tag{4.6}$$

where $\mu^k = \gamma^k \sigma^k$ with $\gamma^k = \frac{1}{n}\langle \hat{X}^k, \hat{Y}^k \rangle$ and $\sigma^k = \left(\frac{r_a^k}{r^k}\right)^3$ is the adaptively chosen centering parameter with

$$r^k = \tfrac{1}{n}\langle \hat{X}^k, \hat{Y}^k \rangle, \tag{4.7}$$

$$r_a^k = \tfrac{1}{n}\langle \hat{X}^k + \alpha_a^k d\hat{X}_a^k, \hat{Y}^k + \alpha_a^k d\hat{Y}_a^k \rangle. \tag{4.8}$$

Finally, we find a step size $\alpha_c^k \in (0, 1]$ such that

$$\hat{X}^k + \alpha_c^k d\hat{X}_c^k > 0_{np}, \tag{4.9}$$

$$\hat{Y}^k + \alpha_c^k d\hat{Y}_c^k > 0_{np}, \tag{4.10}$$

and then compute the next iterate $Z^{k+1} = Z^k + \alpha_c^k d_c^k$ and update $k = k + 1$.

Remark 4.1 At each iteration, we have to solve a linear system with $2pn + p$ equations and $2pn + n^2$ unknown variables which represent entries of eigenvectors, dual vectors and the matrix A. If $p > n^2$ which is the overdetermined case, the system is likely to have no solution. This clarifies what we have said in the introduction, which is that we should stick to the case where $p \leq n^2$.

4.2.2 The Squaring Trick (ST)

As its name suggests, the squaring technique helps to get rid of the nonnegativity constraints in (4.2) with the setting

$$x_k = u_k^{[2]} = u_k \bullet u_k,$$
$$y_k = v_k^{[2]} = v_k \bullet v_k.$$

One gets in this way a smooth system

$$H(Z) = \begin{bmatrix} Au_1^{[2]} - \lambda_1 u_1^{[2]} + v_1^{[2]} \\ \vdots \\ Au_p^{[2]} - \lambda_p u_p^{[2]} + v_p^{[2]} \\ \|u_1\|^2 - 1 \\ \vdots \\ \|u_p\|^2 - 1 \\ \hat{U} \bullet \hat{V} \end{bmatrix} = 0, \quad Z = (\hat{U}, \hat{V}, \hat{A}), \tag{4.11}$$

with $2pn + p$ equations and $2pn + n^2$ unknown variables, namely

$$\begin{cases} \hat{U} = \begin{bmatrix} u_1 \\ \vdots \\ u_p \end{bmatrix} \in \mathbb{R}^{np}, \quad \hat{V} = \begin{bmatrix} v_1 \\ \vdots \\ v_p \end{bmatrix} \in \mathbb{R}^{np}, \quad \hat{A} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^{n^2}, \\ u_i \in \mathbb{R}^n, \quad v_i \in \mathbb{R}^n, \quad a_i \in \mathbb{R}^n, \quad A = [a_1, ..., a_n] \in \mathcal{M}_n(\mathbb{R}). \end{cases}$$

The Jacobian matrix of H is of the form

$$J_H(Z) = \left[\begin{array}{c|c|c} 2M\mathrm{diag}(\hat{U}) & 2\mathrm{diag}(\hat{V}) & (U^{[2]})^T \otimes I_n \\ \hline N & 0_{p\times np} & 0_{p\times n^2} \\ \hline \mathrm{diag}(\hat{V}) & \mathrm{diag}(\hat{U}) & 0_{np\times n^2} \end{array}\right],$$

where M is the block diagonal matrix whose i-th diagonal block is $A - \lambda_i I_n$, N is the $p-$block diagonal matrix whose $i-$th diagonal block is $2u_i^T$, the square operator $[\cdot]^{[2]}$ is taken componentwisely, and $U = [u_1, u_2, ..., u_p]$.

In order to solve (4.11), we use a Newton method specially tailored for dealing with underdetermined systems of equations. It consists in applying the recursive formula

$$Z^{k+1} = Z^k - [J_H(Z^k)]^\dagger H(Z^k) \quad k = 0, 1, 2, ..., \tag{4.12}$$

where $M^\dagger$ denotes the Moore-Penrose inverse of the rectangular matrix M.

In the case of convergence of the sequence generated by (4.12), to guarantee that its limit is a zero of H, we assume that $J_H(Z^k)$ is of full row rank for $k \in \mathbb{N}$ (see [149]). In this case we can compute the term $\Delta Z^k = -[J_H(Z^k)]^\dagger H(Z^k)$ as the least norm solution of the linear system

$$J_H(Z^k)\Delta Z^k = -H(Z^k).$$

Interested readers in more details about theoretical and convergence analysis of the fixed point iteration (4.12) can consult references such as [85, 149].

Remark 4.2 As we look at $J_H(Z)$, clearly a necessary condition for it to be of full row rank is that $U_i \neq 0$ and $V_i \neq 0$ for all $i \in \{1, 2, ..., np\}$ which amounts to saying that we have strict complementarity conditions.

4.3 Nonsmooth approach

4.3.1 Nonlinear complementarity functions

Throughout this section, we denote by Φ^l an l-complementarity function (NCP-function) which is defined as follows. The function $\Phi^l : \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}^l$ is called an l-complementarity function if for all $X \in \mathbb{R}^l, Y \in \mathbb{R}^l$, and $i \in \{1, ..., l\}$, we have

$$(\Phi(X, Y))_i = 0 \Longleftrightarrow X_i \geq 0, \quad Y_i \geq 0, \quad X_i Y_i = 0.$$

Due to this property, the complementarity function technique consists in reformulating (4.2) as

$$L(Z) = \begin{bmatrix} Ax_1 - \lambda_1 x_1 + y_1 \\ \vdots \\ Ax_p - \lambda_p x_p + y_p \\ \langle 1_n, x_1 \rangle - 1 \\ \vdots \\ \langle 1_n, x_p \rangle - 1 \\ \Phi^{np}(\hat{X}, \hat{Y}) \end{bmatrix} = 0, \quad Z = (\hat{X}, \hat{Y}, \hat{A}), \tag{4.13}$$

where

$$\begin{cases} \hat{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \in \mathbb{R}^{np}, \quad \hat{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} \in \mathbb{R}^{np}, \quad \hat{A} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^{n^2}, \\ x_i \in \mathbb{R}^n, \quad y_i \in \mathbb{R}^n, \quad a_i \in \mathbb{R}^n, \quad A = [a_1, ..., a_n] \in \mathcal{M}_n(\mathbb{R}), \end{cases}$$

and $\Phi^{np} : \mathbb{R}^{np} \times \mathbb{R}^{np} \longrightarrow \mathbb{R}^{np}$ can be any np-complementarity function.

In this note, we consider Φ^{np} being one of the following NCP functions

$$\Phi^{np}_{\mathrm{FB}}(X,Y) = X + Y - [X^{[2]} + Y^{[2]}]^{[1/2]},$$
$$\Phi^{np}_{\min}(X,Y) = \min(X,Y),$$

where the square root operation $[\cdot]^{[1/2]}$ and the min function is carried out componentwisely. The two given NCP functions are not differentiable, but they are locally Lipschitz and semismooth [17]. The former was introduced by Fischer [74] and has been proven to be a useful tool for studying complementarity problems. Let us now recall some basic facts about the semismooth Newton method.

Given a locally Lipschitz mapping $\phi : \mathbb{R}^m \longrightarrow \mathbb{R}^m$, Rademacher's theorem ensures the existence of the Jacobian matrix $J_\phi(z)$ at almost every $z \in \mathbb{R}^m$. The B-subdifferential of ϕ at a point $z \in \mathbb{R}^m$ is defined by

$$\partial_B \phi(z) = \left\{ \lim_k J_\phi(z_k) : \exists (z_k) \subset D_\phi : z_k \to z \right\},$$

where D_ϕ is the set of differentiability points of ϕ. The Clarke generalized Jacobian [67] of ϕ is given by

$$\partial \phi(z) = \mathrm{co}\, \partial_B \phi(z),$$

where "co" stands for the convex hull of the set $\partial_B \phi(z)$. The function ϕ is said to be semismooth [111] at $z \in \mathbb{R}^m$ if it is locally Lipschitz around z, directionally differentiable at z and satisfies the following condition

$$\sup_{M \in \partial \phi(z+h)} \|\phi(z+h) - \phi(z) - Mz\| = o(\|h\|).$$

For solving the equation $\phi(x) = 0$ with ϕ being semismooth, the semismooth Newton method [128] refers to applying the following recursive formulation

$$z^{k+1} = z^k + h^k, \quad k = 0, 1, 2, ...,$$

where h^k is given by solving a linear system $M^k h^k = -\phi(z^k)$ for some $M^k \in \partial \phi(z^k)$. Under the semismoothness of the function ϕ and nonsingularity conditions of its generalized Jacobian, the sequence (z_k) generated by the semismooth Newton method converges at least superlinearly. Precisely, we have the following theorem, see [128].

Theorem 4.1 *Let $\bar{z}$ be a zero of the locally Lipschitz function ϕ. Suppose that ϕ is semismooth at $\bar{z}$, and all matrices in $\partial \phi(\bar{z})$ are nonsingular. Then, there exists a neighborhood*

V of $\bar{z}$ such that the semismooth Newton method initialized at any $z_0 \in V$ generates a sequence (z_k) that converges at least superlinealy to $\bar{z}$

We note that the two considered NCP functions are semismooth and the remaining components of L are continuously differentiable. This follows that L is semismooth. Now we turn to solving our system $L(Z) = 0$. We solve it by using the following trivial extension of the semismooth Newton method just described above to underdetermined systems.

- Choose an initial point Z^0.
- One has a current point Z^k. Pick $M^k \in \partial L(Z^k)$ and compute

$$Z^{k+1} = Z^k - (M^k)^\dagger L(Z^k).$$

The Clarke generalized Jacobian of L at Z is the set of all the matrices having the form

$$\left[\begin{array}{c|c|c} M & I_{np} & X^T \otimes I_n \\ \hline I_p \otimes 1_n^T & 0_{p\times np} & 0_{p\times n^2} \\ \hline U & V & 0_{p\times n^2} \end{array}\right],$$

where M is the block diagonal matrix whose i-th diagonal block is $A-\lambda_i I_n$, $X = [x_1, ..., x_p]$, and U and V are matrices of order np such that

$$[U, V] \in \partial\Phi^{np}(\hat{X}, \hat{Y}).$$

Here, Φ^{np} is one of the two np-complementarity functions Φ^{np}_{FB} or $\Phi^{np}_{\min}(X, Y)$. In these cases, we refer to [119] for the formulas of U and V.

4.3.2 The Lattice Projection Method (LPM)

In contrast with the complementarity function technique described above, the Lattice Projection method [15] does not employ any NCP function but rather utilizes the following observation: For any $\lambda > 0$, we have

$$x \geq 0, \lambda x - Ax \geq 0, \langle x, \lambda x - Ax\rangle \geq 0 \Longleftrightarrow \max(Ax, 0) = \lambda x,$$

where the max function is taken componentwisely.

Throughout this section, without loss of generality we assume $\lambda_i > 0$ for all $i \in \{1, 2, ..., p\}$. This is due to the fact that given $A \in M_n(\mathbb{R})$, we have $\sigma(A + \mu I_n) = \sigma(A) + \mu$ for all $\mu \in \mathbb{R}$. Here $\sigma(A)$ denotes the Pareto spectrum of A. As a result, we can always transform the initial problem into an equivalent one where all the prescribed Pareto eigenvalues are positive.

Using the slack variables $y_k = Ax_k$ and taking into account the above equivalence, we have the corresponding reformulation of the IPEP

$$K(Z) = \begin{bmatrix} Ax_1 - y_1 \\ \vdots \\ Ax_p - y_p \\ \max(y_1, 0) - \lambda_1 x_1 \\ \vdots \\ \max(y_p, 0) - \lambda_p x_p \\ \langle 1_n, x_1 \rangle - 1 \\ \vdots \\ \langle 1_n, x_p \rangle - 1 \end{bmatrix} = 0, \quad Z = (\hat{X}, \hat{Y}, \hat{A}), \tag{4.14}$$

where

$$\begin{cases} \hat{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \in \mathbb{R}^{np}, \quad \hat{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} \in \mathbb{R}^{np}, \quad \hat{A} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^{n^2}, \\ x_i \in \mathbb{R}^n, \quad y_i \in \mathbb{R}^n, \quad a_i \in \mathbb{R}^n, \quad A = [a_1, ..., a_n] \in \mathcal{M}_n(\mathbb{R}). \end{cases}$$

Clearly this system is locally Lipchitz and semismooth. Analogously to the previous section, we adopt the Semismooth Newton method to solve this system. Given a matrix Q having m columns, we define $Q(:) := \begin{bmatrix} q_1 \\ \vdots \\ q_m \end{bmatrix}$ where q_i is the i-th column of Q. The Clarke generalized Jacobian of K is the set of all the matrices having the form

$$\left[\begin{array}{c|c|c} I_p \otimes A & -I_{np} & X^T \otimes I_n \\ \hline -\text{diag}\left((1_{n\times p}\text{diag}(\vec{\lambda}))(:)\right) & C & 0_{np\times n^2} \\ \hline I_p \otimes 1_n^T & 0_{p\times np} & 0_{p\times n^2} \end{array}\right],$$

where $X = [x_1, x_2, ..., x_p]$ and C belongs to the Clarke generalized Jacobian of the function $\mathbb{R}^{np} \ni Y \mapsto \max(Y, 0)$ at $\hat{Y}$. For the formula of the Clarke generalized Jacobian of such the function, one can consult in, for instance, [67].

4.4 Numerical tests

First of all, we present how to choose initial points for each method.

MPCM: we generate a random matrix A uniformly distributed on $[-1,1]^{n\times n}$ and random vectors $\omega_1,...,\omega_p$ uniformly distributed on $]0,1]^n$. Then we set

$$\begin{aligned} X &= \left[\frac{\omega_1}{\langle 1_n,\omega_1\rangle},\ldots,\frac{\omega_1}{\langle 1_n,\omega_p\rangle}\right],\\ Y &= X\mathrm{diag}(\vec{\lambda})-AX. \end{aligned}$$

Finally, the initial points are set as follows

$$\hat{X}^0 = X(:), \hat{Y}^0 = \max(Y(:), 0.01), \text{ and } \hat{A}^0 = A(:),$$

where the max operator is carried out componentwisely.

ST: we generate a random matrix A with uniform distribution on $[-1,1]^{n\times n}$, random vectors $\omega_1,\ldots,\omega_p$ uniformly distributed on $[-1,1]^n$ and a random matrix D uniformly distributed on $\{-1,1\}^{n\times p}$. Then we set

$$\begin{aligned} U &= \left[\frac{\omega_1}{\|\omega_1\|},\ldots,\frac{\omega_1}{\|\omega_p\|}\right]\\ X &= U^{[2]},\\ Y &= X\mathrm{diag}(\vec{\lambda})-AX,\\ V &= D\bullet|Y|^{[1/2]}, \end{aligned}$$

where the absolute value operator $|\cdot|$ is carried out componentwisely. Finally, the initial points are set as follows

$$\hat{U}^0 = U(:), \hat{V}^0 = V(:), \text{ and } \hat{A}^0 = A(:).$$

$\mathbf{SNM_{min}}$, $\mathbf{SNM_{FB}}$: we generate a random matrix A uniformly distributed on $[-1,1]^{n\times n}$ and random vectors $\omega_1,...,\omega_p$ uniformly distributed on $[-1,1]^n$. Then we set

$$\begin{aligned} X &= \left[\frac{\omega_1}{\langle 1_n,\omega_1\rangle},\ldots,\frac{\omega_1}{\langle 1_n,\omega_p\rangle}\right],\\ Y &= X\mathrm{diag}(\vec{\lambda})-AX. \end{aligned}$$

Finally, the initial points are set as follows

$$\hat{X}^0 = X(:), \hat{Y}^0 = Y(:), \text{ and } \hat{A}^0 = A(:).$$

LPM: we generate a random matrix A uniformly distributed on $[-1,1]^{n\times n}$ and random vectors $\omega_1, ..., \omega_p$ uniformly distributed on $[-1,1]^n$. Then we set

$$\begin{aligned} X &= \left[\frac{\omega_1}{\langle 1_n, \omega_1\rangle}, \ldots, \frac{\omega_1}{\langle 1_n, \omega_p\rangle}\right], \\ Y &= AX. \end{aligned}$$

Finally, the initial points are set as follows

$$\hat{X}^0 = X(:), \hat{Y}^0 = Y(:), \text{ and } \hat{A}^0 = A(:).$$

Secondly, we mention that a solution is called to be found by a method if the norm of the objective function corresponding to that method at an iterate is less than or equal to 10^{-8}. A failure is declared when the number of iterations exceeds 100, or the (generalized) Jacobian matrix is ill-conditioned according to Matlab's criterion.

This section is divided into two parts. The first one is to consider a particular example and then apply all our methods to solve it. The solution matrix A resulting from each method is provided; moreover, we also provide the Pareto spectrum of those solutions. After that, we will compare the five methods using performance profiles in which we take into consideration the average number of iterations and the percentage of failures as performance measures. To this end, we now first consider an example in which $\Theta = \{1, 2, 4, 6, 8, 12\}$ is the set of prescribed eigenvalues and $n = 3$ is the order of matrices to be found. Applying all five methods yields different solutions as summarized in Table 4.1. Now, we compare the 5 discussed solvers. In order to complete this experiment, we choose the *performance profiles* developed by E. D. Dolan and J. J. Moré [72] as a tool for comparing the solvers. The performance profiles give for each $t \in \mathbb{R}$, the proportion $\rho_s(t)$ of test problems on which each solver under comparison has a performance within the factor t of the best possible ratio.
We have chosen a set P of 30 problems corresponding to 30 random vectors with positive components. For each problem, we run the five methods with 1000 initial points, and we look for matrices A of order n such that n is the smallest integer satisfying $p \leq \tau_n := n(n+1)/2$, where p is the number of Pareto eigenvalues. The average number of iterations and the percentage of failures are used as performance measures. Let S be

Method	A	$\sigma(A)$
MPCM	$\begin{bmatrix} 20 & -534.1521 & -278.6133 \\ 0.1643 & -3.6195 & 976.5402 \\ 0.0008 & -0.0438 & 9.6195 \end{bmatrix}$	$\Theta \cup \{-3.6195\}$
ST	$\begin{bmatrix} 1 & 49.9629 & 190.639 \\ -0.0038 & 12 & 0.067 \\ -0.0787 & 122.2038 & 9 \end{bmatrix}$	$\Theta \cup \{1.0174\}$
SNM_{FB}	$\begin{bmatrix} 20.3198 & -9.6546 & -13.8467 \\ 6.2351 & 6 & -13.5499 \\ 2.3572 & -0.369 & 2 \end{bmatrix}$	$\Theta \cup \{8.3198\}$
SNM_{min}	$\begin{bmatrix} 21 & -4.2692 & 0.8295 \\ 29.399 & -4.6058 & 4.4803 \\ 22.3138 & -7.4444 & 9.6058 \end{bmatrix}$	$\Theta \cup \{-4.6058\}$
LPM	$\begin{bmatrix} 8 & 0.242 & -27.2918 \\ -5.4576 & 12.3301 & -80.2682 \\ 0.2931 & -0.1412 & 2 \end{bmatrix}$	$\Theta \cup \{1.809\}$

Table 4.1: Five different solutions resulted from the five methods

the set of the five solvers that will be compared. The *performance ratio* is defined by

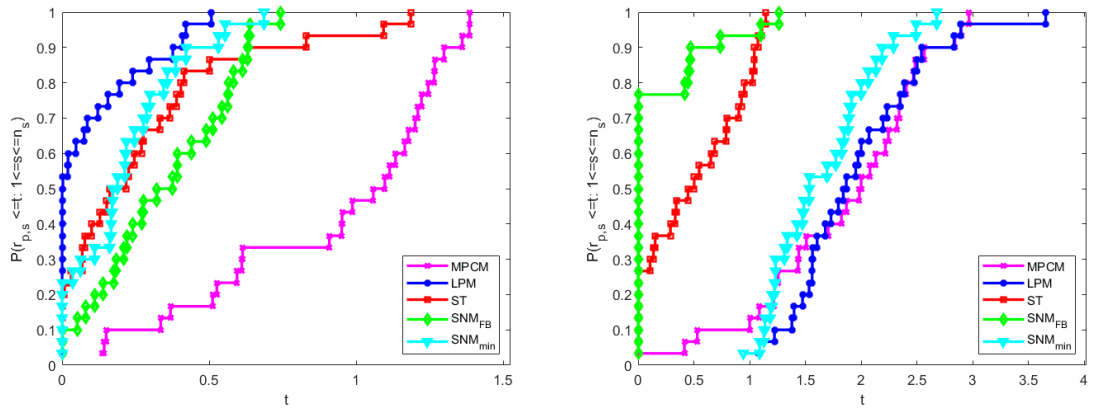
$$r_{p,s} = \frac{t_{p,s}}{\min\{t_{p,s} : s \in S\}},$$

where $p \in P$, $s \in S$, and $t_{p,s}$ is a performance measure. The performance of the solver $s \in S$ is defined by

$$\rho_s(t) = \frac{1}{n_p}\text{size}\{p \in P \ : \ \log_2(r_{p,s}) \leq t\},$$

where, n_p is the number of problems, and t is a real factor, and size(O) denotes the cardinality of O. For more details, we refer to [72].

Figure 1 represents the performance profiles of the 5 methods, namely MPCP, LPM, ST (Squaring Trick), SNM_{FB} (Semismooth Newton Method with Φ_{FB}), and SNM_{min} (Semismooth Newton Method with Φ_{min}) on a set of 30 random problems in which the average percentage of failures and the average number of iterations are taken as performance measures. It can be seen that with respect to the average percentage of failures, SNM_{FB} performs the best when there are nearly 80 % of problems it wins over all other methods. ST ranks second in this regard; the performances of SNM_{min}, LPM and MPCM regarding this criteria are generally the same. By contrast, in terms of the average number of iterations, LPM has the most number of wins; ST and SNM_{min} do not differ much in their performances and are followed respectively by SNM_{FB} and MPCM. Regardless of



(a) $t_{p,s}$ = the average number of iterations **(b)** $t_{p,s}$ = the percentage of failures

Figure 4.1: Performance profiles of MPCM, ST, SNM_{FB}, SNM_{min} and LPM on inverse Pareto eigenvalue problems

the criterion considered, we can see that MPCM performs the worst. In addition, we have tested with various problems and found that MPCM can only solve small size problems.

4.5 Extension to inverse quadratic eigenvalue complementarity problems

In this section, we show how these methods (except LPM) can naturally be extended to cope with more general inverse eigenvalue complementarity problems. To fix the idea, the case corresponding to the quadratic pencil (defined later) will be considered. The corresponding problem is called the inverse quadratic pencil eigenvalue complementarity problem.

First, let us recall that $M_r(\lambda)$ is called a pencil with respect to a finite collection $\{A_0, A_2, ..., A_r\}$ $(r \geq 1)$ of real matrices of order n if it admits the following form

$$M_r(\lambda) = \sum_{k=0}^{r} \lambda^k A_k.$$

The Pareto eigenvalue problem is just a particular case of the following model

$$0 \leq x \perp M_r(\lambda)x \geq 0, \tag{4.15}$$

where $M_r(\lambda)$ is a pencil of order r.
We generalize the inverse Pareto eigenvalue problem by considering the inverse problem

to (4.15). More precisely, given a prescribed set of distinct real numbers $\Theta = \{\lambda_1, ..., \lambda_p\}$, the inverse problem to (4.15) concerns finding a collection $\{A_0, ..., A_r\}$ of $n \times n$ matrices such that $A_r \neq 0_{n\times n}$ and for each $\lambda \in \Theta$, (4.15) admits a non trivial solution $x \in \mathbb{R}^n$. We refer to this type of problem as the inverse pencil eigenvalue complementarity problem; moreover, in the particular case in which $r = 2$, they are referred to as the inverse quadratic pencil eigenvalue complementarity problem.

Formally, the inverse quadratic pencil eigenvalue complementarity problem consists of solving a system of the form

$$\begin{cases} x_k \geq 0_n,\ A_0x_k + \lambda_k A_1 x_k + \lambda_k^2 A_2 x_k \geq 0_n,\ \langle x_k, A_0x_k + \lambda_k A_1 x_k + \lambda_k^2 A_2 x_k\rangle = 0,\ k = 1, \ldots, p, \\ \langle 1_n, x_k\rangle = 1,\ k = 1, \ldots, p, \\ A_2 \neq 0_{n\times n}, \end{cases} \tag{4.16}$$

where the unknown variables are the columns of the $n \times n$ matrices A_0, A_1 and A_2, and the vectors $x_1, ..., x_p \in \mathbb{R}^n$.
By introducing the slack variables $y_k = (A_0 + \lambda_k A_1 + \lambda_k^2 A_2)x_k$, we reformulate (4.16) as

$$\begin{cases} x_k \geq 0_n \\ y_k \geq 0_n \\ \langle x_k, y_k\rangle = 0 \\ (A_0 + \lambda_k A_1 + \lambda_k^2 A_2)x_k - y_k = 0_n \\ \langle 1_n, x_k\rangle - 1 = 0 \\ \|A_2\|_F^2 - 1 = 0 \end{cases} \quad \text{for all } k \in \{1, .., p\}, \tag{4.17}$$

where $\|A_2\|_F = \left(\sum_{i=1}^n \sum_{j=1}^n (A_2)_{ij}^2\right)^{1/2}$ is the Frobenius norm of A_2. The last equation is imposed without loss of generality in order to ensure that $A_2 \neq 0_{n\times n}$.

Apparently, all the methods that we have considered (except LPM), including MPCM, ST, SNM_{FB} and $\text{SNM}_{\min}$, can be employed to solve the system (4.17). In what follows, we will provide the functions as well as the Jacobians/generalized Jacobians associated with each of those methods.

4.5.1 Applying MPCM

The function $F : \mathbb{R}^{2np+3n^2} \to \mathbb{R}^{2np+p+1}$ corresponding to the MPCM is defined as follows

$$F(Z) = \begin{bmatrix} (A_0 + \lambda_1 A_1 + \lambda_1^2 A_2)x_1 - y_1 \\ \vdots \\ (A_0 + \lambda_p A_1 + \lambda_p^2 A_2)x_p - y_p \\ \langle 1_n, x_1 \rangle - 1 \\ \vdots \\ \langle 1_n, x_p \rangle - 1 \\ \|A_2\|_F^2 - 1 \\ \hat{X} \bullet \hat{Y} \end{bmatrix}, \quad Z = (\hat{X}, \hat{Y}, \hat{A}_0, \hat{A}_1, \hat{A}_2), \tag{4.18}$$

where

$$\begin{cases} \hat{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \in \mathbb{R}^{np}, \quad \hat{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} \in \mathbb{R}^{np}, \\ \hat{A}_j = \begin{bmatrix} a_1^j \\ \vdots \\ a_n^j \end{bmatrix} \in \mathbb{R}^{n^2}, \quad A_j = \left[a_1^j, ..., a_n^j\right] \in \mathcal{M}_n(\mathbb{R}) \quad \forall j \in \{0, 1, 2\}, \\ x_i \in \mathbb{R}^n, \quad y_i \in \mathbb{R}^n, \quad a_i^j \in \mathbb{R}^n \quad \forall i \in \{1, ..., n\}, \forall j \in \{0, 1, 2\}. \end{cases}$$

The Jacobian matrix of F has the form

$$J_F(Z) = \left[\begin{array}{c|c|c} M & -I_{np} & [X^T \quad \mathrm{diag}(\vec{\lambda})X^T \quad \mathrm{diag}(\vec{\lambda}^{[2]})X^T] \otimes I_n \\ \hline I_p \otimes 1_n^T & 0_{p\times np} & 0_{p\times 3n^2} \\ \hline 0_{1\times np} & 0_{1\times np} & [0_{1\times n^2} \quad 0_{1\times n^2} \quad 2\hat{A}_2^T] \\ \hline \mathrm{diag}(\hat{Y}) & \mathrm{diag}(\hat{X}) & 0_{np\times 3n^2} \end{array}\right],$$

where M is the block diagonal matrix whose i-th diagonal block is $A_0 + \lambda_i A_1 + \lambda_i^2 A_2$, $\vec{\lambda} = (\lambda_1, ..., \lambda_p)^T$ and $X = [x_1, ..., x_p]$.

4.5.2 Applying ST

The function $H : \mathbb{R}^{2np+3n^2} \to \mathbb{R}^{2np+p+1}$ corresponding to the ST is defined as follows

$$H(Z) = \begin{bmatrix} (A_0 + \lambda_1 A_1 + \lambda_1^2 A_2)u_1^{[2]} - v_1^{[2]} \\ \vdots \\ (A_0 + \lambda_p A_1 + \lambda_p^2 A_2)u_p^{[2]} - v_p^{[2]} \\ \|u_1\|^2 - 1 \\ \vdots \\ \|u_p\|^2 - 1 \\ \|A_2\|_F^2 - 1 \\ \hat{U} \bullet \hat{V} \end{bmatrix}, \quad Z = (\hat{U}, \hat{V}, \hat{A}_0, \hat{A}_1, \hat{A}_2), \tag{4.19}$$

where

$$\begin{cases} \hat{U} = \begin{bmatrix} u_1 \\ \vdots \\ u_p \end{bmatrix} \in \mathbb{R}^{np}, \quad \hat{V} = \begin{bmatrix} v_1 \\ \vdots \\ v_p \end{bmatrix} \in \mathbb{R}^{np}, \\ \hat{A}_j = \begin{bmatrix} a_1^j \\ \vdots \\ a_n^j \end{bmatrix} \in \mathbb{R}^{n^2}, \quad A_j = \left[a_1^j, ..., a_n^j\right] \in \mathcal{M}_n(\mathbb{R}) \quad \forall j \in \{0,1,2\}, \\ u_i \in \mathbb{R}^n, \quad v_i \in \mathbb{R}^n, \quad a_i^j \in \mathbb{R}^n \quad \forall i \in \{1, ..., n\}, \forall j \in \{0,1,2\}. \end{cases}$$

The Jacobian matrix of H is of the form

$$J_H(Z) = \left[\begin{array}{c|c|c} 2M\mathrm{diag}(\hat{U}) & -2\mathrm{diag}(\hat{V}) & [(U^{[2]})^T \quad \mathrm{diag}(\vec{\lambda})(U^{[2]})^T \quad \mathrm{diag}(\vec{\lambda}^{[2]})(U^{[2]})^T] \otimes I_n \\ \hline N & 0_{p\times np} & 0_{p\times 3n^2} \\ \hline 0_{1\times np} & 0_{1\times np} & [0_{1\times n^2} \quad 0_{1\times n^2} \quad 2\hat{A}_2^T] \\ \hline \mathrm{diag}(\hat{V}) & \mathrm{diag}(\hat{U}) & 0_{np\times 3n^2} \end{array}\right].$$

where M is the block diagonal matrix whose i-th diagonal block is $A_0 + \lambda_i A_1 + \lambda_i^2 A_2$, N is a $p-$block diagonal matrix whose $i-$th diagonal block is $2u_i^T$ and $U = [u_1, ..., u_p]$

4.5.3 Applying SNM_{FB} and SNM_{min}

The function $L : \mathbb{R}^{2np+3n^2} \to \mathbb{R}^{2np+p+1}$ corresponding to SNM_{FB} or SNM_{min} is defined as follows

$$L(Z) = \begin{bmatrix} (A_0 + \lambda_1 A_1 + \lambda_1^2 A_2)x_1 - y_1 \\ \vdots \\ (A_0 + \lambda_p A_1 + \lambda_p^2 A_2)x_p - y_p \\ \langle 1_n, x_1 \rangle - 1 \\ \vdots \\ \langle 1_n, x_p \rangle - 1 \\ \|A_2\|_F^2 - 1 \\ \Phi^{np}(\hat{X}, \hat{Y}) \end{bmatrix} = 0, \quad Z = (\hat{X}, \hat{Y}, \hat{A}_0, \hat{A}_1, \hat{A}_2), \tag{4.20}$$

where

$$\begin{cases} \hat{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \in \mathbb{R}^{np}, \quad \hat{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} \in \mathbb{R}^{np}, \\ \hat{A}_j = \begin{bmatrix} a_1^j \\ \vdots \\ a_n^j \end{bmatrix} \in \mathbb{R}^{n^2}, \quad A_j = \left[a_1^j, ..., a_n^j\right] \in \mathcal{M}_n(\mathbb{R}) \quad \forall j \in \{0, 1, 2\}, \\ x_i \in \mathbb{R}^n, \quad y_i \in \mathbb{R}^n, \quad a_i^j \in \mathbb{R}^n \quad \forall i \in \{1, ..., n\}, \forall j \in \{0, 1, 2\}, \end{cases}$$

and ϕ^{np} is one of the two complementarity functions considered in Section 4.3.

The Clarke generalized Jacobian of L at Z contains all the matrices having the form

$$\left[\begin{array}{c|c|c} M & -I_{np} & [X^T \quad \text{diag}(\vec{\lambda})X^T \quad \text{diag}(\vec{\lambda}^{[2]})X^T] \otimes I_n \\ \hline I_p \otimes 1_n^T & 0_{p \times np} & 0_{p \times 3n^2} \\ \hline 0_{1 \times np} & 0_{1 \times np} & [0_{1 \times n^2} \quad 0_{1 \times n^2} \quad 2\hat{A}_2^T] \\ \hline U & V & 0_{np \times 3n^2} \end{array}\right],$$

where M is the block diagonal matrix whose i-th diagonal block is $A_0 + \lambda_i A_1 + \lambda_i^2 A_2$, $X = [x_1, ..., x_p]$, and U and V are matrices of order np such that

$$[U, V] \in \partial \Phi^{np}(\hat{X}, \hat{Y}).$$

Here, Φ^{np} is one of the two np-complementarity functions Φ_{FB}^{np} or $\Phi_{\text{min}}^{np}(X, Y)$.

By using a similar technique as the proof of Proposition 3.4 in [76], we can have a similar result concerning a condition on p relative to n that ensures given p arbitrary distinct real numbers λ_i's we can always find a triplet (A_0, A_1, A_2) of $n \times n$ matrices whose corresponding quadratic pencil eigenvalue complementarity problem admits all λ_i's as solutions. Before diving into the result, we need the following lemma

Lemma 4.1 *Let A_0, A_1 and A_2 be matrices of order n. Then. for all $\alpha \in \mathbb{R}$ we have*

$$\gamma \in \sigma(A_0, A_1, A_2) \Longleftrightarrow \gamma - \alpha \in \sigma(A_0 + \alpha A_1 + \alpha^2 A_2, A_1 + 2\alpha A_2, A_2).$$

Proof. Set $\lambda = \gamma - \alpha$, we have

$$\begin{aligned}
\gamma \in \sigma(A_0, A_1, A_2) &\Longleftrightarrow \lambda + \alpha \in \sigma(A_0, A_1, A_2) \\
&\Longleftrightarrow 0 \leq x \perp \left[A_0 + (\lambda + \alpha)A_1 + (\lambda + \alpha)^2 A_2\right] x \geq 0, \text{ for some } x \neq 0 \\
&\Longleftrightarrow 0 \leq x \perp \left[A_0 + \alpha A_1 + \alpha^2 A_2 + \lambda(A_1 + 2\alpha A_2) + \lambda^2 A_2\right] x \geq 0 \\
&\Longleftrightarrow \lambda \in \sigma(A_0 + \alpha A_1 + \alpha^2 A_2, A_1 + 2\alpha A_2, A_2),
\end{aligned}$$

which completes the proof of Lemma 4.1. ∎

Proposition 4.1 *For an arbitrary collection $\Theta = \{\lambda_1, ..., \lambda_p\}$ of distinct real numbers with p not exceeding $\tau_n = n(n+1)/2$ with $n \geq 2$, there exists at least a triplet (A_0, A_1, A_2) of $n \times n$ matrices such that all the λ_i's, $i = 1, \ldots, p$, are solutions of the corresponding problem (4.15) with $r = 2$.*

Proof. It is sufficient to prove this statement in the case $p = \tau_n$. Let us observe first, according to Lemma 4.1 that we can assume without loss of generality that $\lambda_i > 0$ for all $i \in \{1, ..., p\}$. We set $A_1 = \text{diag}(\lambda_1, .., \lambda_n)$, where $\lambda_1, ..., \lambda_n$ are assumed to be the first n smallest numbers in Θ and $A_2 = -I_n$. The remaining $\tau_n - n = C_n^2$ elements $\lambda_{n+1}, \lambda_{n+2}, ..., \lambda_p$ of Θ can be indexed over the set $\Delta = \{(i, j) : 1 \leq i < j \leq n\}$, which means we can have a one to one correspondence, denoted by μ, that associates each pair $(i, j) \in \Delta$ with μ_{ij} belonging to the set $\{\lambda_{n+1}, ..., \lambda_p\}$. Now we define A_0 as follows

$$A_0(i, j) = \begin{cases} 0, & i = j, \\ \mu_{ij}^2 - \mu_{ij}\lambda_i, & i < j, \\ \mu_{ji}^2 - \mu_{ji}\lambda_i, & j < i. \end{cases}$$

We can verify that for all $1 \leq i \leq n$

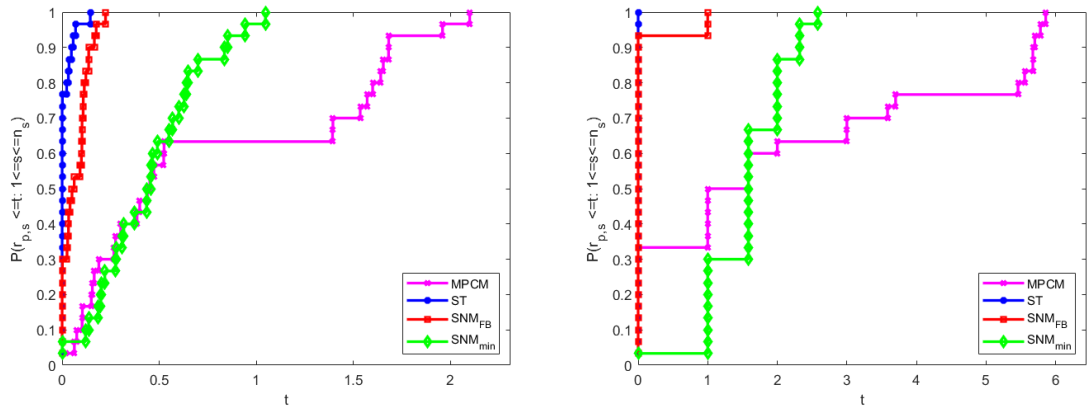
$$0 \leq e_i \perp (A_0 + \lambda_i A_1 + \lambda_i^2 A_2)e_i \geq 0,$$

where e_i denotes the n-th basis vector of $\mathbb{R}^n$.
We can also show that for $(i, j) \in \Delta$

$$0 \leq (e_i + e_j) \perp (A_0 + \mu_{ij} A_1 + \mu_{ij}^2 A_2)(e_i + e_j) \geq 0.$$

Therefore, the proof is completed. ■

Now we give some numerical results for the inverse quadratic pencil eigenvalue complementarity problem. As a first example, we consider $\Theta = \{1, 2, 4, 6, 8, 12\}$ and $n = 3$. Table 4.2 shows how the 4 methods presented in this section give different solutions to the given problem. Analogously to Section 4, we also compare MPCM, ST, SNM_{FB} and SNM_{min} on the quadratic pencil eigenvalue complementarity problem by using the performance profiles. As can be seen from Figure 4.2, the two most attractive methods



(a) $t_{p,s}$ = the average number of iterations **(b)** $t_{p,s}$ = the percentage of failures

Figure 4.2: Performance profiles of MPCM, ST, SNM_{FB} and SNM_{min} on inverse quadratic pencil eigenvalue complementarity problems

among the four to solve this type of problem are ST and SNM_{FB}. SNM_{min} ranks third while MPCM performs the worst as it does in inverse Pareto eigenvalue problems. We can see that these figures of performance profiles are consistent with those in the case of the inverse Pareto eigenvalue problem.

Method	(A_0, A_1, A_2)	$\sigma(A_0, A_1, A_2)$
MPCM	$A_0 = \begin{bmatrix} 0.4485 & 2.6022 & -1.3082 \\ 5.7842 & 10.7159 & 4.8616 \\ 0.3623 & 5.4577 & 2.9938 \end{bmatrix}$	$\Theta \cup \{-3.4113, -3.4021, -1.6773, 12.4942\}$
	$A_1 = \begin{bmatrix} -0.6728 & -1.194 & -4.6145 \\ -0.7954 & 5.4956 & 1.0699 \\ 0.5499 & 2.3657 & 0.5036 \end{bmatrix}$	
	$A_2 = \begin{bmatrix} 0.2243 & 0.1453 & 0.7278 \\ 0.188 & -0.5324 & -0.1022 \\ -0.1938 & -0.1417 & -0.1097 \end{bmatrix}$	
ST	$A_0 = \begin{bmatrix} -0.4428 & -0.6096 & -0.2637 \\ -0.364 & -0.0004 & 0.3199 \\ -1.4729 & -0.9470 & 0.0078 \end{bmatrix}$	$\Theta \cup \{-10.7246, 1.6019, 1.9634, 5.9484\}$
	$A_1 = \begin{bmatrix} 2.7137 & 2.3261 & -0.1233 \\ 0.5981 & 0.0003 & -0.1339 \\ -2.5919 & 0.559 & 1.494 \end{bmatrix}$	
	$A_2 = \begin{bmatrix} 0.2569 & -0.1802 & -0.4104 \\ 0.2516 & 0 & 0.0135 \\ 0.6905 & -0.039 & -0.4375 \end{bmatrix}$	
SNM_{FB}	$A_0 = \begin{bmatrix} -2.6976 & -0.2255 & -0.607 \\ -0.7477 & 0.7451 & -0.8772 \\ 0.2708 & -1.7924 & -1.4494 \end{bmatrix}$	$\Theta \cup \{-7.7249, 0.7392\}$
	$A_1 = \begin{bmatrix} -0.9479 & 1.8798 & -0.185 \\ 0.3456 & 1.469 & 0.89 \\ -2.3656 & 0.6169 & 2.3231 \end{bmatrix}$	
	$A_2 = \begin{bmatrix} 0.1721 & -0.2877 & 0.3694 \\ -0.0576 & -0.067 & 0.5595 \\ 0.4113 & 0.1444 & -0.4902 \end{bmatrix}$	
SNM_{min}	$A_0 = \begin{bmatrix} -1.2595 & -1.113 & -4.6623 \\ 0.5841 & 4.6952 & 5.8726 \\ 0.4464 & -7.2773 & 0.1566 \end{bmatrix}$	$\Theta \cup \{-323.7934\}$
	$A_1 = \begin{bmatrix} 1.2556 & 0.764 & 11.4069 \\ -0.2102 & -1.9563 & -3.9018 \\ 1.8169 & 1.6159 & -0.0326 \end{bmatrix}$	
	$A_2 = \begin{bmatrix} 0.0039 & 0.1104 & -0.786 \\ 0.005 & 0.1956 & 0.4686 \\ 0.0106 & 0.3348 & 0.0016 \end{bmatrix}$	

Table 4.2: Four different solutions resulted from the four methods

4.6 Conclusions

In this chapter, we have presented five methods for solving the inverse Pareto eigenvalue problem. Both smooth and nonsmooth approaches are considered. To compare the 5

given methods, we have used the performance profiles by Dolan & Moré [72]. Numerical experiment showed that the interior point method, namely MPCM, is not a good method for solving this type of problem. Furthermore, depending on the criterion considered, we can decide whether LPM or SNM_{FB} is the best. In particular, while SNM_{FB} shows that it experiences the fewest failures among the 5 methods, LPM takes the fewest number of iterations to find a solution. Finally, we show that 4 (out of 5) considered methods including MPCM, ST, SNM_{FB} and $\text{SNM}_{\min}$ can be extended to deal with inverse quadratic pencil eigenvalue complementarity problems. We consider only the case of the nonnegative orthant $K = \mathbb{R}^n_+$, which corresponds to Pareto eigenvalue problems. It would be interesting to extend, the methods considered in this chapter, to the more general case of cone-constrained eigenvalue problems involving a general closed convex cone K of $\mathbb{R}^n$. The convergence analysis of the Mehrotra predictor corrector method (MPCM) for IPEP as well as the improvement of the upper bound τ_n in Proposition 4.1 need further investigations and are open questions. This is out of the scope of the current version and will be the subject of a future research project.

5

First order inertial optimization algorithms with threshold effects associated with dry friction

Contents

This chapter covers the material discussed in the published paper [8], which was produced in collaboration with S. Adly and H. Attouch

5.1 Introduction

Throughout this chapter, $\mathcal{H}$ is a real Hilbert space equipped with the scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $\| \cdot \|$. The objective function $f : \mathcal{H} \to \mathbb{R}$ is assumed to be differentiable with Lipschitz continuous gradient. Unless otherwise specified, f is not assumed to be convex. When we consider the continuous dynamic on which the algorithms are based, and where the Hessian is involved, more regularity is needed for f which is then assumed to be C^2. Weakening these assumptions by removing the smoothness of f will be examined at the end of the chapter. We will analyze the convergence properties of several algorithms that can be obtained by temporal discretization of the differential inclusion

$$\boxed{\ddot{x}(t) + \gamma \dot{x}(t) + \partial \varphi \Big(\dot{x}(t) + \beta \nabla f(x(t)) \Big) + \beta \nabla^2 f(x(t)) \dot{x}(t) + \nabla f(x(t)) \ni 0,} \tag{5.1}$$

where $\gamma > 0$ and $\beta > 0$ are respectively the viscous damping and Hessian damping coefficients, and φ is a dry friction potential function with a sharp minimum at the origin. This type of autonomous system, with a damping which acts as a closed loop control of the sum of the velocity and gradient terms, was recently introduced by Attouch, Bot, and Csetnek in [27]. It falls within the general framework of the use of inertial dynamics in optimization to accelerate algorithms, as mechanical intuition naturally suggests. An abundant literature has been devoted to the link between damped inertial dynamics and corresponding optimization algorithms obtained by temporal discretization, see e.g. [31, 35, 58, 59, 120, 125, 143] for recent developments on the subject. The term $\gamma \dot{x}(t)$ in (5.1) models the viscous damping with a *fixed* positive coefficient $\gamma > 0$. Thus our algorithms are linked to the heavy ball with friction method of Polyak [125] (as opposed to

the Nesterov acceleration method [143] which corresponds to a viscous damping coefficient which vanishes asymptotically ($\gamma(t) \to 0$ as $t \to +\infty$) . This framework is well suited to dry friction, and will allow us to provide first order algorithms which are robust and converge for nonconvex and nonsmooth optimization problems.

Dry friction Following [3–5, 7], we say that the potential function φ satisfies the dry friction property $(\mathrm{DF})_r$, $r > 0$, if the following properties are satisfied:

$$(\mathrm{DF})_r \quad \begin{cases} \varphi : \mathcal{H} \to \mathbb{R}_+ \text{ is convex continuous,} \\ \min_{\xi \in \mathcal{H}} \varphi(\xi) = \varphi(0) = 0, \\ \varphi(\xi) \geq r\|\xi\| \ \ \forall \xi \in \mathcal{H}. \end{cases}$$

The function $\varphi(x) = r\|x\|$, $r > 0$ is a model example of potential which satisfies the dry friction property. In what follows, the friction potential function φ is assumed to satisfy the dry friction property. An important property associated with dry friction is stated in the lemma below (see [3–5] for further details).

Lemma 5.1 *Suppose that $\varphi : \mathcal{H} \to \mathbb{R}_+$ satisfies $(\mathrm{DF})_r$. Then one has $\overline{\mathbb{B}}(0, r) \subset \partial\varphi(0)$, and therefore*

$$\|x\| \leq \lambda r \Longrightarrow \mathrm{prox}_{\lambda\varphi}(x) = 0.$$

In the above formula, $\mathrm{prox}_{\varphi} : \mathcal{H} \to \mathcal{H}$ denotes the proximal mapping associated with the convex function φ. Recall that, for any $x \in \mathcal{H}$, for any $\lambda > 0$

$$\mathrm{prox}_{\lambda\varphi}(x) = \mathrm{argmin}_{\xi \in \mathcal{H}} \left\{ \lambda\varphi(\xi) + \tfrac{1}{2}\|x - \xi\|^2 \right\}.$$

Lemma 5.1 establishes a thresholding property for the proximal operator associated with a dry friction potential. It will play a key role in showing that after a finite number of steps our algorithm will arrive at the regime of the steepest descent method.

The algorithm We will focus on various temporal discretizations of (5.1) and their links with numerical optimization. Our main results concern the convergence properties of the proximal-gradient algorithm

$$\text{(IPAHDD-C1)} \qquad \begin{cases} y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1}), \\ x_{k+1} = x_k - \beta h \nabla f(x_k) + h\, \mathrm{prox}_{\frac{h}{1+\gamma h}\varphi}\left(\frac{1}{1+\gamma h} y_k + \frac{(\gamma\beta - 1)h}{1+\gamma h}\nabla f(x_k)\right), \end{cases}$$

which comes from the temporal discretization with step size $h > 0$ of (5.1). (IPAHDD) is the terminology introduced by Adly and Attouch [3] for this type of algorithm, which is a shorthand for Inertial Proximal Algorithm with Hessian Damping and Dry friction. The suffix C refers to the Composite form in which the dry friction acts in (5.1). (IPAHDD-C1) is a first-order autonomous algorithm whose behavior has some similarities with the heavy

ball with friction method, and with the steepest descent method (when φ is a prox-friendly function). Indeed, as we will see, it can be advantageously compared to these two methods in the presence of errors/perturbations. Specifically, the algorithm handles errors more efficiently than the other two methods when dealing with external errors, while it performs at least as well in situations involving additive errors.

Motivation As a specific property of (5.1), the dry friction term $\partial\varphi\Big(\dot{x}(t)+\beta\nabla f(x(t))\Big)$ involves both the velocity vector and the gradient of f. This makes this dynamic different from that studied previously, where the term of dry friction concerns only the velocity vector. This is a simple yet nontrivial modification which undoubtedly makes the dynamics totally different from that studied in [3–5, 7]. A major advantage of considering the dry friction term in this new form is that the iterates generated by our algorithm will converge towards a critical point of f (a minimizer in the case where f is convex). In fact, any stationary point x_∞ of the dynamic (5.1) satisfies: $\partial\varphi(\beta\nabla f(x_\infty)) + \nabla f(x_\infty) \ni 0$. This is equivalent to $\beta\nabla f(x_\infty) = \text{prox}_{\beta\varphi}(0)$, which, combining with the fact that the potential φ satisfies the dry friction property $(\text{DF})_r$, implies that $\nabla f(x_\infty) = 0$ (see Lemma 5.1), i.e., x_∞ is a critical point of f. By contrast, for each sequence (x_k) generated by the algorithms in [3–5, 7], there is only convergence of (x_k) towards an "approximate" critical point x_∞ of f, that is, $-\nabla f(x_\infty) \in \partial\varphi(0)$. Dry friction is an important subject in mechanics. It produces stabilization of mechanical systems in finite-time. This contrasts with the viscous damping that can asymptotically produce many small oscillations. This makes it an attractive tool for optimization. The use of dry friction in optimization is a relatively new topic. First results concerning the property of finite convergence under the action of dry friction were obtained by Adly, Attouch, and Cabot [7]. Corresponding results for Partial Differential Equations have been obtained by Amann and Diaz in [22]. Despite certain formal analogies, our study clearly stands out from the study of optimization algorithms using the notion of sharp minimum, because in our situation the sharpness property relates to the velocity and not the function to be minimized.

Hessian driven damping The combination of viscous friction with dry friction and Hessian driven damping has been considered by Adly and Attouch in [3–5]. Even if the dynamic (5.1) requires that the potential f is twice differentiable, the associated algorithm is a first-order one. In fact, since the term $\nabla^2 f(x(t))\dot{x}(t)$ is the time derivative of $\nabla f(x(t))$, we obtain that its temporal discretization contains only the gradients of f at two consecutive steps, and is therefore relevant to first-order algorithms. The Hessian driven damping has a natural connection with the strong damping property in mechanics and physics, see [82]. It helps to control and attenuate the oscillation effects that occur

naturally with inertial systems. The first result involving the Hessian-driven damping concerned the dynamic with fixed viscous damping

$$(\mathrm{DIN})_{\gamma,\beta} \qquad \ddot{x}(t) + \gamma\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0$$

see [21]. The terminology (DIN) refers to the interpretation of this system as a (regularized) Dynamic Inertial Newton method. Several recent papers have been devoted to the combination of this dynamic with the Nesterov accelerated gradient method, see [34, 42, 60, 64, 95, 101, 138, 143].

Results Under suitable conditions on the damping parameters γ, β and the step size h, we show that any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C1) converges weakly to a minimizer of f when f is convex, and to a critical point of f when f is a nonconvex function which satisfies the Kurdyka-Lojasiewicz property. Moreover, the sequence $(x_k)_k$ follows the steepest descent method after a finite number of steps, and the summability property is satisfied $\sum \|\nabla f(x_k)\|^2 < +\infty$. The convergence results tolerate the presence of errors, under weak assumptions. When f is strongly convex, (IPAHDD-C1) achieves exponential convergence. We show that various discretizations of the dynamic (5.1) lead to different algorithms which share similar convergence properties, including the combination of dry friction and Hessian-driven damping with the extrapolation method of Nesterov. We finally consider corresponding splitting algorithms for composite minimization, including the case of nonsmooth nonconvex d.c. programming, and Lasso problems.

Contents In section 5.2, we proceed with the Lyapunov analysis of the inertial proximal-gradient algorithm (IPAHDD-C1). In section 5.3, we analyze the convergence properties of (IPAHDD-C1) and successively examine the case of a general convex function f, then the strongly convex case, and finally the case of a nonconvex function f which satisfies the Kurdyka-Lojasiewicz property. In section 5.4, we show the robustness of the algorithm (IPAHDD-C1) with respect to perturbations, and errors. In section 5.5, we examine two variants of the algorithm which have a structure similar to that of the accelerated gradient method of Nesterov. In section 5.6, based on the variational properties of Moreau's envelope, we extend our results to the case where $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a convex lower semicontinuous and proper function, and then we examine the case of nonsmooth d.c. problems. In section 5.7 we extend our analysis to the case of additive composite optimization problems of Lasso type, and obtain a corresponding splitting algorithm. Section 5.8 is devoted to numerical experiments. We complete the chapter with some concluding remarks and perspectives.

5.2 Lyapunov analysis of the (IPAHDD-C1) algorithm

Given a constant step size $h > 0$, we consider the following temporal discretization of (5.1), which is implicit with respect to the nonsmooth operator $\partial\varphi$, and explicit with respect to the smooth operator ∇f:

$$\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\gamma}{h}(x_{k+1} - x_k) + \partial\varphi\left(\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k)\right) + \frac{\beta}{h}(\nabla f(x_k) - \nabla f(x_{k-1})) + \nabla f(x_k) \ni 0. \quad (5.2)$$

Set $y_k := \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})$, $k \geq 1$. Let us reformulate (5.2) with the help of y_k. We obtain

$$y_{k+1} + \tfrac{h}{1+\gamma h}\partial\varphi(y_{k+1}) \ni \tfrac{1}{1+\gamma h}y_k + \tfrac{(\gamma\beta-1)h}{1+\gamma h}\nabla f(x_k).$$

Equivalently,

$$y_{k+1} = \operatorname{prox}_{\frac{h}{1+\gamma h}\varphi}\left(\frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta-1)h}{1+\gamma h}\nabla f(x_k)\right), \quad (5.3)$$

which gives $x_{k+1} = x_k - \beta h\nabla f(x_k) + h\operatorname{prox}_{\frac{h}{1+\gamma h}\varphi}\left(\frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta-1)h}{1+\gamma h}\nabla f(x_k)\right)$.
Therefore, we obtain the following algorithm

(IPAHDD-C1)

Initialize : $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$.

$y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})$.

$x_{k+1} = x_k - \beta h\nabla f(x_k) + h\operatorname{prox}_{\frac{h}{1+\gamma h}\varphi}\left(\frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta-1)h}{1+\gamma h}\nabla f(x_k)\right)$.

Note that the discretization of the first order equivalent system in time and space introduced in [27] leads to a similar algorithm.

5.2.1 Energy estimates

We can now state our main result concerning the algorithm (IPAHDD-C1).

Theorem 5.1 *Let $f : \mathcal{H} \to \mathbb{R}$ be a differentiable function such that $\inf_{\mathcal{H}} f > -\infty$, and whose gradient is L-Lipschitz continuous. Assume that the friction potential $\varphi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive*

parameters h, γ, β satisfy the relation

$$hL \leq \frac{2\gamma}{\gamma\beta + 1}. \tag{5.4}$$

Let $(x_k)_k$ be a sequence generated by (IPAHDD-C1). *Then, the energy-like sequence $(E_k)_k$*

$$E_k := \frac{1}{2}\|\frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})\|^2 + (\gamma\beta + 1)\left(f(x_k) - \inf_{x\in H} f(x)\right)$$

is non-increasing, and the following energy properties are satisfied:

$$\sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < +\infty \text{ and } \sum_{k=1}^{+\infty} \|x_{k+1} - x_k\|^2 < +\infty.$$

Proof. Multiplying (5.2) by h and rewriting it using y_k, we obtain for $k \geq 1$

$$y_{k+1} - y_k + \gamma(x_{k+1} - x_k) + h\partial\varphi(y_{k+1}) + h\nabla f(x_k) \ni 0. \tag{5.5}$$

Taking the scalar product of (5.5) with y_{k+1}, we obtain

$$\begin{aligned}\|y_{k+1}\|^2 - \langle y_k, y_{k+1}\rangle + \gamma\langle x_{k+1} - x_k, \frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k)\rangle + h\langle\partial\varphi(y_{k+1}), y_{k+1}\rangle \\ + h\langle\nabla f(x_k), \frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k)\rangle = 0.\end{aligned}$$

Equivalently

$$\begin{aligned}\underbrace{\|y_{k+1}\|^2 - \langle y_k, y_{k+1}\rangle}_{A} + \underbrace{\frac{\gamma}{h}\|x_{k+1} - x_k\|^2 + (\gamma\beta + 1)\langle x_{k+1} - x_k, \nabla f(x_k)\rangle}_{B} \\ + h\langle\partial\varphi(y_{k+1}), y_{k+1}\rangle + \beta h\|\nabla f(x_k)\|^2 = 0.\end{aligned} \tag{5.6}$$

1) **Estimate** $h\langle\partial\varphi(y_{k+1}), y_{k+1}\rangle$. Using the convexity of φ and $\varphi(0) = 0 = \min_{\mathcal{H}}\varphi$, we have

$$h\langle\partial\varphi(y_{k+1}), y_{k+1}\rangle \geq h\varphi(y_{k+1}). \tag{5.7}$$

2) **Estimate** A. We have

$$A \geq \|y_{k+1}\|^2 - \|y_k\|\|y_{k+1}\| \geq \|y_{k+1}\|^2 - \frac{1}{2}\left(\|y_{k+1}\|^2 + \|y_k\|^2\right) = \frac{1}{2}\|y_{k+1}\|^2 - \frac{1}{2}\|y_k\|^2. \tag{5.8}$$

3) **Estimate** B. According to the classical gradient descent lemma, we obtain

$$\begin{aligned} B &\geq \frac{\gamma}{h}\|x_{k+1}-x_k\|^2+(\gamma\beta+1)(f(x_{k+1})-f(x_k)-\frac{L}{2}\|x_{k+1}-x_k\|^2) \\ &\geq (\gamma\beta+1)(f(x_{k+1})-f(x_k))+\left(\frac{\gamma}{h}-\frac{L}{2}(\gamma\beta+1)\right)\|x_{k+1}-x_k\|^2 \\ &\geq (\gamma\beta+1)(f(x_{k+1})-f(x_k)). \end{aligned} \tag{5.9}$$

where the last inequality follows from the assumption (5.4) on the parameters, which gives equivalently $\frac{\gamma}{h}-\frac{L}{2}(\gamma\beta+1)\geq 0$. By combining (5.6), (5.7), (5.8) and (5.9), we obtain

$$\frac{1}{2}\|y_{k+1}\|^2-\frac{1}{2}\|y_k\|^2+(\gamma\beta+1)(f(x_{k+1})-f(x_k))+h\varphi(y_{k+1})+\beta h\|\nabla f(x_k)\|^2\leq 0. \tag{5.10}$$

Equivalently

$$E_{k+1}-E_k+h\varphi(y_{k+1})+\beta h\|\nabla f(x_k)\|^2\leq 0, \tag{5.11}$$

where

$$E_k:=\frac{1}{2}\|y_k\|^2+(\gamma\beta+1)\left(f(x_k)-\inf_{x\in H}f(x)\right).$$

By summing the inequalities (5.11) from $k=1$ to N, and using that $E_k\geq 0$, we obtain

$$h\sum_{k=1}^{N}\varphi\Big(\frac{1}{h}(x_{k+1}-x_k)+\beta\nabla f(x_k)\Big)+\beta h\sum_{k=1}^{N}\|\nabla f(x_k)\|^2\leq E_1-E_{N+1}\leq E_1.$$

Letting $N\to+\infty$, and since h,β are supposed to be positive, we obtain

$$\sum_{k=1}^{+\infty}\|\nabla f(x_k)\|^2<+\infty \text{ and } \sum_{k=1}^{+\infty}\varphi\Big(\frac{1}{h}(x_{k+1}-x_k)+\beta\nabla f(x_k)\Big)<+\infty. \tag{5.12}$$

Since φ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r>0$, we also deduce that

$$\sum_{k=1}^{+\infty}\|\frac{1}{h}(x_{k+1}-x_k)+\beta\nabla f(x_k)\|<+\infty, \text{ that is } \sum_{k=1}^{+\infty}\|y_k\|<+\infty. \tag{5.13}$$

Therefore, $\lim_k y_k=0$, which implies $\|y_k\|^2\leq\|y_k\|$ for k large enough, and hence $\sum_{k=1}^{+\infty}\|y_k\|^2<+\infty$. This property, combined with $\sum_{k=1}^{+\infty}\|\nabla f(x_k)\|^2<+\infty$ immediately gives

$$\sum_{k=1}^{+\infty}\|x_{k+1}-x_k\|^2<+\infty.$$

The proof is thereby completed. ■

5.2.2 Finite time transition to the steepest descent method

Let us now prove that after a finite number of steps, the sequence $(x_k)_k$ follows the steepest descent method.

Theorem 5.2 *Let $f : \mathcal{H} \to \mathbb{R}$ be a differentiable function such that $\inf_{\mathcal{H}} f > -\infty$, and whose gradient is L-Lipschitz continuous. Assume that the friction potential $\varphi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive parameters h, γ, β satisfy the relation*

$$hL \leq \frac{2\gamma}{\gamma\beta + 1}. \tag{5.14}$$

Let $(x_k)_k$ be a sequence generated by (IPAHDD-C1)*. Then, after a finite number of steps*

$$\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k) = 0,$$

i.e. the sequence $(x_k)_k$ follows the steepest descent method.

Proof. The proof relies on Lemma 5.1. Recall that, according to (5.3), we have the following equivalent formulation of the algorithm (IPAHDD-C1): $y_{k+1} = \mathrm{prox}_{\frac{h}{1+\gamma h}\varphi}(z_k)$, where

$$z_k = \frac{1}{1+\gamma h} y_k + \frac{(\gamma\beta - 1)h}{1+\gamma h}\nabla f(x_k).$$

According to (5.12), (5.13), and since the general term of a convergent series necessarily goes to zero,

$$\lim_k \nabla f(x_k) = \lim_k y_k = 0.$$

According to the definition of z_k, we get $\lim_k z_k = 0$. Therefore, there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$,

$$\|z_k\| \leq \frac{hr}{1+\gamma h}.$$

According to Lemma 5.1, this implies that $y_{k+1} = \mathrm{prox}_{\frac{h}{1+\gamma h}\varphi}(z_k) = 0$ for all $k \geq k_0$. Equivalently, $\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k) = 0$ for all $k \geq k_0$, which means that after a finite number of steps, the sequence $(x_k)_k$ follows the steepest descent algorithm. This completes the proof. ■

Remark 5.1 When $\mathcal{H}$ is of finite dimension, let us give another proof of the fact that $y_k = 0$ after a finite number of steps, *i.e.* , $(x_k)_k$ follows the steepest descent. We argue by contradiction, which leads to suppose that there exists a subsequence $(y_{k_l})_l$ such that $\|y_{k_l+1}\| > 0$ for all $l \in \mathbb{N}$. From (5.5), we have

$$-\frac{1}{h}(y_{k_l+1} - y_{k_l}) - \frac{\gamma}{h}(x_{k_l+1} - x_{k_l}) - \nabla f(x_{k_l}) \in \partial\varphi(y_{k_l+1}).$$

Due to the monotoncity of the subdifferential $\partial\varphi$, we have

$$\langle -\frac{1}{h}(y_{k_l+1} - y_{k_l}) - \frac{\gamma}{h}(x_{k_l+1} - x_{k_l}) - \nabla f(x_{k_l}) - \partial\varphi(0), \frac{y_{k_l+1}}{\|y_{k_l+1}\|}\rangle \geq 0 \quad \forall l \in \mathbb{N}.$$

Since the sequence $w_l = \left(\frac{y_{k_l+1}}{\|y_{k_l+1}\|}\right)_l$ is bounded in a finite dimensional space, it has a convergent subsequence. For notational convenience, we use the same notation and therefore assume $w_l \to w$. It is clear that $\|w\| = 1$. Letting $l \to \infty$ in the above inequality, it follows that

$$\langle \partial\varphi(0), w\rangle \leq 0.$$

Since $\overline{\mathbb{B}}(0, r) \subset \partial\varphi(0)$, the above inequality implies that

$$\langle ru, w\rangle \leq 0 \quad \forall u \in \overline{\mathbb{B}}(0, 1).$$

Choose $u = w$, it follows that $r\|w\|^2 \leq 0$, and hence $w = 0$. This is a contradiction with $\|w\| = 1$.

5.2.3 Estimating the transition process

Let us give some information about the number of steps after which the iterates $(x_k)_k$ follow the steepest descent algorithm. According to the proof of Theorem 5.1, this is satisfied as soon as $\|z_k\| \leq \frac{hr}{1+\gamma h}$, where $z_k = \frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta-1)h}{1+\gamma h}\nabla f(x_k)$. Let us take advantage of the summation estimates that we have obtained in the proof of Theorem 5.1, namely

$$\sum_{k=1}^{+\infty} \|y_k\| \leq \frac{E_1}{hr}, \quad \sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < \frac{E_1}{h\beta}. \tag{5.15}$$

According to the definition of z_k, elementary algebra gives

$$\|z_k\|^2 \leq \frac{2}{(1+\gamma h)^2}\|y_k\|^2 + \frac{2(\gamma\beta - 1)^2h^2}{(1+\gamma h)^2}\|\nabla f(x_k)\|^2.$$

According to (5.15) and the inequality $\sum_{k=1}^{+\infty}\|y_k\|^2 \leq (\sum_{k=1}^{+\infty}\|y_k\|)^2$, we infer

$$\sum_{k=1}^{+\infty}\|z_k\|^2 \;\leq\; \frac{2}{(1+\gamma h)^2}\left(\frac{E_1}{hr}\right)^2 + \frac{2(\gamma\beta-1)^2h^2}{(1+\gamma h)^2}\frac{E_1}{h\beta}.$$

Set $M := \frac{2}{(1+\gamma h)^2}\left(\frac{E_1}{hr}\right)^2 + \frac{2(\gamma\beta-1)^2h^2}{(1+\gamma h)^2}\frac{E_1}{h\beta}$. We have

$$\sum_{k=1}^{+\infty}\|z_k\|^2 \geq \sum_{i=k}^{2k}\|z_i\|^2 \geq k\inf_{k\leq i\leq 2k}\|z_i\|^2.$$

Therefore $\inf_{k\leq i\leq 2k}\|z_i\| \leq \sqrt{\frac{M}{k}}$. Combining the above results, we obtain that

$$k \geq \frac{M(1+h\gamma)^2}{h^2r^2} \Longrightarrow \exists i,\ k\leq i\leq 2k \text{ such that } \frac{1}{h}(x_{i+1}-x_i)+\beta\nabla f(x_i)=0.$$

5.2.4 Exponential convergence rate of (y_k) to zero

Recall that $y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1}),\ k \geq 1$.

Proposition 5.1 *Set $q = \frac{1}{\sqrt{1+2\gamma h}} \in (0,1)$. Then, there exists $k_0 \in \mathbb{N}$ such that*

$$\|y_k\| \leq q^{k-k_0}\|y_{k_0}\| \quad \forall k > k_0.$$

Proof. The convergence rate of $(y_k)_k$ can be established as follows. First, we have

$$y_{k+1} - y_k + \gamma(x_{k+1}-x_k) + h\partial\varphi(y_{k+1}) + h\nabla f(x_k) \ni 0.$$

Taking the scalar product of the above inclusion with y_{k+1}, and using the convexity of φ, we obtain

$$\|y_{k+1}\|^2 - \langle y_k, y_{k+1}\rangle + \gamma\langle x_{k+1}-x_k, y_{k+1}\rangle + h\varphi(y_{k+1}) + h\langle\nabla f(x_k), y_{k+1}\rangle \leq 0. \tag{5.16}$$

Since $\nabla f(x_k) \to 0$, we have $(\gamma\beta-1)\nabla f(x_k) \in \partial\varphi(0)$ for k sufficiently large due to the dry friction condition. By definition of the subdifferential, we deduce that

$$\varphi(y_{k+1}) \geq (\gamma\beta-1)\langle\nabla f(x_k), y_{k+1}\rangle.$$

Equivalently

$$\varphi(y_{k+1}) + \langle\nabla f(x_k), y_{k+1}\rangle \geq \gamma\beta\langle\nabla f(x_k), y_{k+1}\rangle.$$

According to the above inequality and the Cauchy-Schwarz inequality, from (5.16) we deduce that

$$\frac{1}{2}\|y_{k+1}\|^2 - \frac{1}{2}\|y_k\|^2 + \gamma\langle x_{k+1} - x_k, y_{k+1}\rangle + \gamma\beta h\langle \nabla f(x_k), y_{k+1}\rangle \leq 0.$$

Equivalently

$$\frac{1}{2}\|y_{k+1}\|^2 - \frac{1}{2}\|y_k\|^2 + \gamma h\langle \frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k), y_{k+1}\rangle \leq 0.$$

According to the definition of y_{k+1}, this gives

$$(1 + 2\gamma h)\|y_{k+1}\|^2 \leq \|y_k\|^2.$$

Set $q = \frac{1}{\sqrt{1+2\gamma h}} \in (0,1)$, we finally deduce that $\|y_{k+1}\| \leq q\|y_k\|$ for k sufficiently large, say $k \geq k_0$. Therefore,

$$\|y_k\| \leq q^{k-k_0}\|y_{k_0}\| \quad \forall k > k_0.$$

The proof is thereby completed. ∎

We now analyze the convergence of the sequences (x_k) generated by the algorithm (IPAHDD − C1).

5.3 Convergence results

Let us state our main result concerning the convergence properties of the sequences generated by the algorithm (IPAHDD-C1). They rely on the fact that after a finite number of steps the sequences follow the steepest descent method, and the well-known results relating to this algorithm. We proceed with a unified statement, then examine successively the different cases, f convex, strongly convex, and f non-convex satisfying the (KL) property.

Theorem 5.3 *Let $f : \mathcal{H} \to \mathbb{R}$ be a differentiable function whose gradient is L-Lipschitz continuous, and such that $\inf_{\mathcal{H}} f > -\infty$. Assume that the friction potential $\varphi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive parameters h, γ, β satisfy the relation*

$$hL \leq \frac{2\gamma}{\gamma\beta + 1}. \tag{5.17}$$

Then for any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C1)*, we have the following convergence properties, described below based on the geometric properties of f:*

(i) *Case f convex with $\operatorname{argmin}_{\mathcal{H}} f \neq \emptyset$. Then $(x_k)_k$ converges weakly, and its limit is a minimizer of f.*

(ii) *Case f μ-strongly convex with parameter $\mu > 0$ such that either $h\beta = \frac{1}{L}$ or $h\beta \leq \frac{2}{\mu+L}$.*
Let x_∞ be the unique minimizer of f. Then, we have linear strong convergence of $(x_k)_k$ to x_∞.

(iii) *Take $\mathcal{H} = \mathbb{R}^N$ and suppose that $f : \mathcal{H} \to \mathbb{R}$ satisfies the* (KL) *property. Then $(x_k)_k$ converges, and its limit is a critical point of f.*

Proof. (i) According to Theorem 5.1, after a finite number of steps, say $k \geq k_0$

$$x_{k+1} = x_k - h\beta\nabla f(x_k)$$

i.e. , the sequence $(x_k)_{k\geq k_0}$ follows the classical gradient scheme with the fixed step size $s = h\beta > 0$. It is then a classical result (see for example [51, Corollary 28.9]) that the sequence converges weakly, and its limit is a minimizer of f, whenever the step size s satisfies $s = h\beta < \frac{2}{L}$. Clearly this is satisfied, because, under the assumption (5.17) on the parameters, we have

$$hL \leq \frac{2\gamma}{\gamma\beta + 1} < \frac{2\gamma}{\gamma\beta} = \frac{2}{\beta}.$$

Let us recall that the Opial's lemma is the key ingredient to prove the weak convergence of the iterates.

(ii) We have shown that, after a finite number of steps, the sequence (x_k) follows the steepest descent method. Therefore, the conclusion follows from the classical result concerning the convergence rate of the steepest descent method for strongly convex objective functions, see for example [48].

(iii) $\mathcal{H} = \mathbb{R}^N$ and f satisfies (KL). Basic facts concerning the (KL) properties are recalled in the appendix. Since the sequence (x_k) follows the steepest descent method, the conclusion follows from the convergence result of Attouch, Bolte and Svaiter [26, Theorem 3.2] concerning the convergence of the gradient method for functions satisfying the (KL) property. ■

The following theorem concerns the convergence property of the algorithm (IPAHDD − C1) in the case where f is strongly convex and where the friction potential function φ is under a different setting.

Theorem 5.4 *Let $f : \mathcal{H} \to \mathbb{R}$ be a μ-strongly convex and smooth function whose gradient is L-Lipschitz continuous. Let x_∞ be the unique minimizer of f. Assume that the function $\varphi : \mathcal{H} \to \mathbb{R}$ is a convex function which is differentiable and satisfies $\min_{\xi\in\mathcal{H}} \varphi(\xi) = \varphi(0) = 0$, and whose gradient is Lipschitz continuous on any bounded subset of $\mathcal{H}$ such that*

(i) there exists a positive constant α such that for all u in some ball centered at zero in $\mathcal{H}$

$$\langle \nabla\varphi(u), u\rangle \geq \alpha\|u\|^2.$$

(ii) there exist $p \geq 1, r > 0$ such that for all $u \in \mathcal{H}$, $\varphi(u) \geq r\|u\|^p$.
Suppose that the positive parameters h, γ, β satisfy the relation

$$\frac{\gamma}{h} - \frac{L}{2}(\gamma\beta + 1) \geq 0.$$

Then, for any sequence $(x_k)_k$ generated by the algorithm (IPAHDD − C1)*, we have exponential convergence rate to zero as $k \to \infty$ for $f(x_k) - f(x_\infty), \|x_k - x_\infty\|$ and $\|y_k\|$, where $y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})$. Moreover, if $\frac{1}{h} - L\beta \geq 0$, we have exponential convergence rates to zero as $k \to \infty$, in the following ergodic sense: there exists $C > 1$ such that*

$$(\frac{1}{h^2} - \frac{L\beta}{h})\frac{\sum_{k=1}^n (1+\frac{h}{2\beta})^k\|x_k - x_{k-1}\|^2}{\sum_{k=1}^n (1+\frac{h}{2\beta})^k} + \beta^2\frac{\sum_{k=1}^n (1+\frac{h}{2\beta})^k\|\nabla f(x_{k-1})\|^2}{\sum_{k=1}^n (1+\frac{h}{2\beta})^k} = O\Big(\frac{1}{C^n}\Big).$$

Proof. Repeating the proof of Theorem 5.1 with the awareness of the assumption (ii), we infer that

$$\sum_{k=1}^{\infty}\|y_k\|^p < \infty.$$

It follows that the sequence $(y_k)_k$ is convergent to zero or in particular bounded. We call K the Lipschitz constant of $\nabla\varphi$ on the bounded set containing $\{y_k : k \in \mathbb{N}\}$. As a result, we have for all $k \geq 1$

$$\|\nabla\varphi(y_k)\| \leq K\|y_k\|.$$

From assumption (i) and repeating the proof of Theorem 5.1, we have

$$\frac{1}{2}\|y_{k+1}\|^2 - \frac{1}{2}\|y_k\|^2 + (\gamma\beta + 1)(f(x_{k+1}) - f(x_k)) + h\alpha\|y_{k+1}\|^2 + \beta h\|\nabla f(x_k)\|^2 \leq 0.$$

Set $E_k = \frac{1}{2}\|y_k\|^2 + (1 - \epsilon\beta + \gamma\beta)(f(x_k) - f(x_\infty)) + \epsilon\langle x_k - x_\infty, y_k\rangle$ where $\epsilon > 0$ will be chosen later. We have

$$\begin{aligned} E_{k+1} - E_k = \frac{1}{2}\|y_{k+1}\|^2 - \frac{1}{2}\|y_k\|^2 + (1 - \epsilon\beta + \gamma\beta)(f(x_{k+1}) - f(x_k)) \\ + \epsilon\langle x_{k+1} - x_\infty, y_{k+1}\rangle - \epsilon\langle x_k - x_\infty, y_k\rangle. \end{aligned}$$

Now we estimate the last two terms in the above equality. We have

$$\epsilon\langle x_{k+1}-x_\infty, y_{k+1}\rangle - \epsilon\langle x_k - x_\infty, y_k\rangle = \epsilon\langle x_{k+1}-x_k, y_{k+1}\rangle + \epsilon\langle x_k - x_\infty, y_{k+1}-y_k\rangle.$$

Substituting $y_{k+1}-y_k$ by $-\gamma(x_{k+1}-x_k) - h\nabla\varphi(y_{k+1}) - h\nabla f(x_k)$ gives

$$\begin{aligned}
&\epsilon\langle x_{k+1}-x_\infty, y_{k+1}\rangle - \epsilon\langle x_k - x_\infty, y_k\rangle\\
&= \epsilon\langle x_{k+1}-x_k, y_{k+1}\rangle + \epsilon\langle x_k - x_\infty, -\gamma(x_{k+1}-x_k) - h\nabla\varphi(y_{k+1}) - h\nabla f(x_k)\rangle\\
&= \frac{\epsilon}{h}\|x_{k+1}-x_k\|^2 + \epsilon\beta\langle x_{k+1}-x_k, \nabla f(x_k)\rangle + \epsilon h\langle x_\infty - x_k, \nabla f(x_k)\rangle\\
&+ \epsilon\langle x_k - x_\infty, \gamma(x_k - x_{k+1}) - h\nabla\varphi(y_{k+1})\rangle\\
&\le \frac{\epsilon}{h}\|x_{k+1}-x_k\|^2 + \epsilon\beta(f(x_{k+1}) - f(x_k)) + \epsilon h\langle x_\infty - x_k, \nabla f(x_k)\rangle\\
&+ \epsilon\|x_k - x_\infty\|(\gamma\|x_k - x_{k+1}\| + Kh\|y_{k+1}\|)\\
&\le \frac{\epsilon}{h}\|x_{k+1}-x_k\|^2 + \epsilon\beta(f(x_{k+1}) - f(x_k)) + \epsilon h\langle x_\infty - x_k, \nabla f(x_k)\rangle\\
&+ \epsilon h\frac{\mu}{2}\|x_k - x_\infty\|^2 + \frac{\epsilon h}{2}\Big(\frac{\gamma}{h\sqrt{\mu}}\|x_k - x_{k+1}\| + \frac{K}{\sqrt{\mu}}\|y_{k+1}\|\Big)^2\\
&\le \frac{\epsilon}{h}\|x_{k+1}-x_k\|^2 + \epsilon\beta(f(x_{k+1}) - f(x_k)) + \epsilon h\langle x_\infty - x_k, \nabla f(x_k)\rangle\\
&+ \frac{\epsilon h\mu}{2}\|x_k - x_\infty\|^2 + \frac{\epsilon\gamma^2}{h\mu}\|x_k - x_{k+1}\|^2 + \frac{\epsilon h K^2}{\mu}\|y_{k+1}\|^2.
\end{aligned}$$

To summarize, we have

$$\begin{aligned}
&\epsilon\langle x_{k+1}-x_\infty, y_{k+1}\rangle - \epsilon\langle x_k - x_\infty, y_k\rangle\\
&\le (\frac{\epsilon}{h} + \frac{\epsilon\gamma^2}{h\mu})\|x_{k+1}-x_k\|^2 + \epsilon\beta(f(x_{k+1}) - f(x_k)) + \epsilon h(\langle x_\infty - x_k, \nabla f(x_k)\rangle\\
&+ \frac{\mu}{2}\|x_k - x_\infty\|^2) + \frac{\epsilon h K^2}{\mu}\|y_{k+1}\|^2\\
&\le (\frac{\epsilon}{h} + \frac{\epsilon\gamma^2}{h\mu})\|x_{k+1}-x_k\|^2 + \epsilon\beta(f(x_{k+1}) - f(x_k)) + \epsilon h(f(x_\infty) - f(x_k)) + \frac{\epsilon h K^2}{\mu}\|y_{k+1}\|^2.
\end{aligned}$$

Accordingly, we have

$$
\begin{aligned}
&E_{k+1} - E_k \\
&\leq \frac{1}{2}\|y_{k+1}\|^2 - \frac{1}{2}\|y_k\|^2 + (1+\gamma\beta)(f(x_{k+1}) - f(x_k)) + (\frac{\epsilon}{h} + \frac{\epsilon\gamma^2}{h\mu})\|x_{k+1} - x_k\|^2 \\
&+ \epsilon h(f(x_\infty) - f(x_k)) + \frac{\epsilon h K^2}{\mu}\|y_{k+1}\|^2 \\
&\leq -h\alpha\|y_{k+1}\|^2 - \beta h\|\nabla f(x_k)\|^2 + (\frac{\epsilon}{h} + \frac{\epsilon\gamma^2}{h\mu})\|x_{k+1} - x_k\|^2 + \epsilon h(f(x_\infty) - f(x_k)) \\
&+ \frac{\epsilon h K^2}{\mu}\|y_{k+1}\|^2 \\
&= (\frac{\epsilon h K^2}{\mu} - h\alpha)\|y_{k+1}\|^2 - \beta h\|\nabla f(x_k)\|^2 + (\frac{\epsilon}{h} + \frac{\epsilon\gamma^2}{h\mu})\|x_{k+1} - x_k\|^2 + \epsilon h(f(x_\infty) - f(x_k)) \\
&\leq (\frac{\epsilon h K^2}{\mu} - h\alpha + 2h\epsilon + \frac{2h\epsilon\gamma^2}{\mu})\|y_{k+1}\|^2 + (2h\epsilon\beta^2 + \frac{2h\epsilon\gamma^2\beta^2}{\mu} - \beta h)\|\nabla f(x_k)\|^2 \\
&+ \epsilon h(f(x_\infty) - f(x_k)).
\end{aligned}
$$

We choose $\epsilon > 0$ to be sufficiently small such that there exists $C_1 > 0$ such that

$$
E_{k+1} - E_k \leq -C_1(\|y_{k+1}\|^2 + \|\nabla f(x_k)\|^2 + f(x_k) - f(x_\infty)). \tag{5.18}
$$

Moreover, for $\epsilon > 0$ so small that $1 - \epsilon\beta + \gamma\beta > 0$, we have

$$
\begin{aligned}
&E_{k+1} \\
&= \frac{1}{2}\|y_{k+1}\|^2 + (1 - \epsilon\beta + \gamma\beta)(f(x_{k+1}) - f(x_\infty)) + \epsilon\langle x_{k+1} - x_\infty, y_{k+1}\rangle \\
&= \frac{1}{2}\|y_{k+1}\|^2 + (1 - \epsilon\beta + \gamma\beta)(f(x_{k+1}) - f(x_k)) + \epsilon\langle x_{k+1} - x_k, y_{k+1}\rangle + \frac{\epsilon}{h}\langle x_k - x_\infty, x_{k+1} - x_k\rangle \\
&+ \epsilon\beta(\langle x_k - x_\infty, \nabla f(x_k)\rangle + f(x_\infty) - f(x_k)) + (1+\gamma\beta)(f(x_k) - f(x_\infty)) \\
&\leq \frac{1}{2}\|y_{k+1}\|^2 + (1 - \epsilon\beta + \gamma\beta)(\langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2) + \epsilon\langle x_{k+1} - x_k, y_{k+1}\rangle \\
&+ \frac{\epsilon}{2h}\|x_k - x_\infty\|^2 + \frac{\epsilon}{2h}\|x_{k+1} - x_k\|^2 + \frac{\epsilon\beta}{2\mu}\|\nabla f(x_k)\|^2 + (1+\gamma\beta)(f(x_k) - f(x_\infty)) \\
&\leq \frac{1}{2}\|y_{k+1}\|^2 + (1 - \epsilon\beta + \gamma\beta)(\|\nabla f(x_k)\|\|x_{k+1} - x_k\| + \frac{L}{2}\|x_{k+1} - x_k\|^2) + \epsilon\|x_{k+1} - x_k\|\|y_{k+1}\| \\
&+ \frac{\epsilon}{2h}\frac{2}{\mu}(f(x_k) - f(x_\infty)) + \frac{\epsilon}{2h}\|x_{k+1} - x_k\|^2 + \frac{\epsilon\beta}{2\mu}\|\nabla f(x_k)\|^2 + (1+\gamma\beta)(f(x_k) - f(x_\infty)).
\end{aligned}
$$

Taking into account that $\|x_{k+1} - x_k\| \leq h^2\|y_{k+1}\| + h^2\beta\|\nabla f(x_k)\|$, we choose $\epsilon > 0$ small

enough so that there exists $C_2 > 0$ satisfying

$$E_{k+1} \leq C_2(\|y_{k+1}\|^2 + \|\nabla f(x_k)\|^2 + f(x_k) - f(x_\infty)). \tag{5.19}$$

From (5.18) and (5.19), we have

$$E_{k+1} \leq C_3 E_k,$$

where $C_3 = \frac{C_2}{C_1+C_2} \in (0,1)$. This apparently implies that $E_k \leq C_3^{k-1} E_1$.

On the other hand,

$$\begin{aligned} E_k &= \frac{1}{2}\|y_k\|^2 + (1 - \epsilon\beta + \gamma\beta)(f(x_k) - f(x_\infty)) + \epsilon\langle x_k - x_\infty, y_k\rangle \\ &\geq \frac{1}{2}\|y_k\|^2 + (1 - \epsilon\beta + \gamma\beta)(f(x_k) - f(x_\infty)) - \frac{\epsilon}{2}\|x_k - x_\infty\|^2 - \frac{\epsilon}{2}\|y_k\|^2 \\ &\geq (\frac{1}{2} - \frac{\epsilon}{2})\|y_k\|^2 + (1 - \epsilon\beta + \gamma\beta - \frac{\epsilon}{\mu})(f(x_k) - f(x_\infty)). \end{aligned}$$

Choose $\epsilon > 0$ to be small enough so that $\frac{1}{2} - \frac{\epsilon}{2} > 0$ and $1 - \epsilon\beta + \gamma\beta - \frac{\epsilon}{\mu} > 0$. As a result, there exists $C_4 > 0$ such that

$$E_k \geq C_4(\|y_k\|^2 + f(x_k) - f(x_\infty)).$$

From the above inequality, the strong convexity of f and the fact that $E_k \leq C_3^{k-1} E_1$, we derive the first part of the theorem.

Now we turn to the second part. In this part, recall that we assume $\frac{1}{h} - L\beta \geq 0$.
We have shown that

$$C_3^{k-1} E_1 \geq C_4(\|y_k\|^2 + f(x_k) - f(x_\infty)).$$

Equivalently,

$$\begin{aligned}
&C_5C_3^{k-1} \geq \|y_k\|^2 + f(x_k) - f(x_\infty), \quad \text{where} \quad C_5 = \frac{E_1}{C_4} > 0\\
&\geq \frac{1}{h^2}\|x_k - x_{k-1}\|^2 + \beta^2\|\nabla f(x_{k-1})\|^2 + \frac{2\beta}{h}\langle x_k - x_{k-1}, \nabla f(x_{k-1})\rangle + f(x_k) - f(x_\infty)\\
&\geq (\frac{1}{h^2} - \frac{L\beta}{h})\|x_k - x_{k-1}\|^2 + \beta^2\|\nabla f(x_{k-1})\|^2 + \frac{2\beta}{h}(f(x_k) - f(x_{k-1})) + f(x_k) - f(x_\infty)\\
&\geq (\frac{1}{h^2} - \frac{L\beta}{h})\|x_k - x_{k-1}\|^2 + \beta^2\|\nabla f(x_{k-1})\|^2 + (\frac{2\beta}{h} + 1)(f(x_k) - f(x_\infty))\\
&- \frac{2\beta}{h}(f(x_{k-1}) - f(x_\infty)).
\end{aligned}$$

Set $u_k = \frac{(h+2\beta)^{k+1}}{h(2\beta)^k}(f(x_k) - f(x_\infty)), k \geq 0$ and $m = 1 + \frac{h}{2\beta} > 1$. Multiplying the above inequality with m^k gives

$$\frac{C_5}{C_3}(C_3m)^k \geq (\frac{1}{h^2} - \frac{L\beta}{h})m^k\|x_k - x_{k-1}\|^2 + \beta^2 m^k\|\nabla f(x_{k-1})\|^2 + u_k - u_{k-1}.$$

Taking the sum on both sides of the above inequality over $k = 1, 2, .., N$ gives

$$\frac{C_5}{C_3}\sum_{k=1}^{N}(C_3m)^k \geq (\frac{1}{h^2} - \frac{L\beta}{h})\sum_{k=1}^{N} m^k\|x_k - x_{k-1}\|^2 + \beta^2\sum_{k=1}^{N} m^k\|\nabla f(x_{k-1})\|^2 + u_N - u_0.$$

This implies that

$$\begin{aligned}
\frac{C_5}{C_3}\sum_{k=1}^{N}(C_3m)^k + \frac{h+2\beta}{h}(f(x_0) - f(x_\infty)) \geq &(\frac{1}{h^2} - \frac{L\beta}{h})\sum_{k=1}^{N} m^k\|x_k - x_{k-1}\|^2\\
&+ \beta^2\sum_{k=1}^{N} m^k\|\nabla f(x_{k-1})\|^2.
\end{aligned}$$

The multiplication of the left hand side of the above inequality with $\frac{1}{\sum_{k=1}^{N} m^k}$ is equal to

$$h(N) = \begin{cases}
\frac{C_5}{C_3}\frac{N(m-1)}{m^{N+1}-1} + \frac{h+2\beta}{h}(f(x_0) - f(x_\infty))\frac{m-1}{m^{N+1}-1}, & C_3m = 1,\\
\\
\frac{C_5}{C_3}\frac{(C_3m)^{N+1}-1}{C_3m-1}\frac{m-1}{m^{N+1}-1} + \frac{h+2\beta}{h}(f(x_0) - f(x_\infty))\frac{m-1}{m^{N+1}-1}, & C_3m \neq 1.
\end{cases}$$

It is straightforward to see that $h(N) = O\Big(\frac{1}{C^N}\Big)$ for some $C > 1$. This completes our proof. ∎

With another different setting for the dry friction potential φ, we have the following theorem which deals with nonconvex objective function f using the Kurdyka-Lojasiewicz theory. We will consider in this case $\mathcal{H} = \mathbb{R}^N$. We will show that the convergence of a sequence generated by (IPAHDD-C1) is still achieved provided that some assumptions on f and φ are imposed. First of all, let us recall the definition of the Kurdyka-Lojasiewicz property.

Definition 5.1 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a proper and continuously differentiable function. f is called to have the Kurdyka-Lojasiewicz (KL) property at u^* which is a critical point of f if there exists $\eta \in (0, \infty]$, a neighborhood U of u^*, and a concave, continuous function $\theta : [0, \eta) \to \mathbb{R}_+$ which vanishes at 0, and which is smooth on $(0, \eta)$, and such that $\theta' > 0$ such that*

$$\theta'(f(u) - f(u^*))\|\nabla f(u)\| \geq 1, \quad \forall u \in U \cap [f(u^*) < f < f(u^*) + \eta].$$

The function θ in the above definition is called the desingularizing function associated with the KL function f. Now we state the convergence result.

Theorem 5.5 *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a C^2 function whose gradient is L-Lipschitz continuous, and such that $\inf_{\mathbb{R}^N} f > -\infty$. Let $\varphi : \mathbb{R}^N \to \mathbb{R}_+$ be a convex function which is differentiable. Let $(x_n)_n$ be a bounded sequence generated by (IPAHDD-C1). We make the following assumptions on the data f, φ and the positive parameters γ, β, h:*

- *(assumptions on φ): Suppose that φ satisfies the following growth conditions: there exist positive c, ϵ and δ such that $\varphi(u) \geq c\|u\|^2$ for all $u \in \mathbb{R}^N$, and $\|\nabla\varphi(u)\| \leq \delta\|u\|$ for all u with $\|u\| \leq \epsilon$.*
- *(assumptions on γ, β and h): Suppose that these parameters satisfy the following relation*

$$\frac{\gamma}{h} - \frac{L}{2}(\gamma\beta + 1) > 0,$$
$$1 - \beta\delta > 0.$$

- *(assumption on f): Suppose that the function H satisfies the (KL) property, where $H : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is defined by*

$$H(x, y) = (\gamma\beta + 1)f(x) + \frac{1}{2}\left\|\frac{1}{h}(x - y) + \beta\nabla f(y)\right\|^2.$$

Moreover, f is supposed to satisfy

$$\left\| -\frac{1}{h}I_N + \beta\nabla^2 f(x)\right\| \leq D, \quad \forall x \in \mathbb{R}^N \quad \text{for some } D \geq 0.$$

Then, the following properties are fulfilled

(i) $\sum_{n=1}^{\infty} \|x_{n+1} - x_n\| < \infty$.
(ii) $x_n \to x_\infty$ *as* $n \to \infty$, *where* x_∞ *is a critical point of* f.

Proof. Repeating the proof of Theorem 5.1, we have

$$E_{k+1} - E_k + h\varphi(X_{k+1}) + \beta h\|\nabla f(x_k)\|^2 + (\frac{\gamma}{h} - \frac{L}{2}(\gamma\beta + 1))\|x_{k+1} - x_k\|^2 \leq 0,$$

where X_k and E_k are defined as in the proof of Theorem 5.1. From this inequality, we can conclude that

$$\sum_{n=1}^{\infty} \|x_{k+1} - x_k\|^2 < \infty, \quad \text{and hence} \quad \|x_{k+1} - x_k\| \to 0 \text{ as } k \to \infty,$$
$$\sum_{n=1}^{\infty} \|\nabla f(x_k)\|^2 < \infty, \quad \text{and hence} \quad \|\nabla f(x_k)\| \to 0 \text{ as } k \to \infty,$$

and that

$$E_{k+1} - E_k + (\frac{\gamma}{h} - \frac{L}{2}(\gamma\beta + 1))\|x_{k+1} - x_k\|^2 \leq 0.$$

Equivalently

$$H(x_{k+1}, x_k) + l\|x_{k+1} - x_k\|^2 \leq H(x_k, x_{k-1}), \tag{5.20}$$

where $l = \frac{\gamma}{h} - \frac{L}{2}(\gamma\beta + 1) > 0$.
Since f is bounded from below, this implies that

$$\lim_{k\to\infty} H(x_k, x_{k-1}) \in \mathbb{R}.$$

Denote by $\omega((x_n)_n)$ the set of limit points of $(x_n)_n$ and by $crit(f)$ the set of critical points of f, that is $x \in crit(f)$ if and only if $\nabla f(x) = 0$.
We notice that $\omega((x_n)_n) \subset crit(f)$. This together with the fact that $x_{k+1} - x_k \to 0$ implies that $\omega((x_{n+1}, x_n)_n) \subset crit(H)$. From (5.20), we can infer that H is constant on $\omega((x_{n+1}, x_n)_n)$. Indeed, for $x^* \in \omega((x_n)_n)$, we have

$$H(x^*, x^*) = (\gamma\beta + 1)f(x^*) = \lim_{k\to\infty} H(x_k, x_{k-1}).$$

Since H satisfies the (KL) property, we denote by θ its desingularizing function. Now, we consider 2 cases.

Case 1: There exists $\bar{k}$ such that $H(x_{\bar{k}+1}, x_{\bar{k}}) = H(x^*, x^*)$. From the decreasing property

(5.20), this follows that $(x_n)_n$ is a constant sequence from which the conclusion is immediate.

Case 2: For all $k \geq 0$, $H(x_{k+1}, x_k) > H(x^*, x^*)$. Since θ is concave and $\theta' > 0$, we derive from (5.20) that

$$\begin{aligned}\Delta_k &:= \theta(H(x_k, x_{k-1}) - H(x^*, x^*)) - \theta(H(x_{k+1}, x_k) - H(x^*, x^*)) \\ &\geq \theta'(H(x_k, x_{k-1}) - H(x^*, x^*))(H(x_k, x_{k-1}) - H(x_{k+1}, x_k)) \\ &\geq \theta'(H(x_k, x_{k-1}) - H(x^*, x^*))l\|x_{k+1} - x_k\|^2 \\ &\geq \frac{l\|x_{k+1} - x_k\|^2}{\|\nabla H(x_k, x_{k-1})\|},\end{aligned}$$

where the last inequality is true for sufficiently large k and obtained by applying Lemma 6 in [56] to the non empty compact set $\Omega = \omega((x_{k+1}, x_k)_k)$.
Moreover, a direct calculation yields

$$\nabla H(x_k, x_{k-1}) = \begin{bmatrix} (\gamma\beta + 1)\nabla f(x_k) + \frac{1}{h}\Big(\frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})\Big) \\ \Big(-\frac{1}{h}I_N + \beta\nabla^2 f(x_{k-1})\Big)\Big(\frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})\Big) \end{bmatrix}.$$

Recall that we have the following equality

$$\begin{aligned}\frac{1}{h^2}(x_{k+1} - x_k) - \frac{1}{h^2}(x_k - x_{k-1}) + \frac{\gamma}{h}(x_{k+1} - x_k) + \nabla\varphi(\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k)) \\ + \frac{\beta}{h}(\nabla f(x_k) - \nabla f(x_{k-1})) + \nabla f(x_k) = 0.\end{aligned}$$

According to the L-Lipschitz continuity of f, the growth conditions of φ and the fact that $\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k) \to 0$, we have for k large enough that

$$\|\nabla f(x_k)\| \leq (\frac{1}{h^2} + \frac{\gamma}{h} + \frac{\delta}{h})\|x_{k+1} - x_k\| + (\frac{1}{h^2} + \frac{L\beta}{h})\|x_k - x_{k-1}\| + \delta\beta\|\nabla f(x_k)\|,$$

or

$$(1 - \delta\beta)\|\nabla f(x_k)\| \leq (\frac{1}{h^2} + \frac{\gamma}{h} + \frac{\delta}{h})\|x_{k+1} - x_k\| + (\frac{1}{h^2} + \frac{L\beta}{h})\|x_k - x_{k-1}\|.$$

Since $1 - \delta\beta > 0$, this implies that there exists $C_1 > 0$ such that

$$\|\nabla f(x_k)\| \leq C_1\Big(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|\Big).$$

With this inequality, we can also prove that there exists $C_2 > 0$ such that

$$\|\nabla f(x_{k-1})\| \leq C_2\Big(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|\Big).$$

In view of the two above inequalities and the assumption $\Big\| -\frac{1}{h}I_N + \beta\nabla^2 f(x_{k-1})\Big\| \leq D$, we implies that there exists $C_3 > 0$ such that

$$\|\nabla H(x_k, x_{k-1})\| \leq C_3\Big(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|\Big).$$

Taking into account the above inequality, we continue the estimation of Δ_k as follows

$$\Delta_k \geq \frac{l}{C_3}\frac{\|x_{k+1} - x_k\|^2}{\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|}.$$

Equivalently,

$$\|x_{k+1} - x_k\|^2 \leq \frac{C_3}{l}\Delta_k(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|),$$

which implies

$$\begin{aligned}\|x_{k+1} - x_k\| &\leq \sqrt{\frac{C_3}{l}\Delta_k(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|)} \\ &\leq \frac{\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|}{4} + \frac{C_3}{l}\Delta_k.\end{aligned}$$

Finally we get for k large enough, say $k \geq k_0$

$$\|x_{k+1} - x_k\| \leq \frac{1}{3}\|x_k - x_{k-1}\| + C_4\Delta_k,$$

where $C_4 = \frac{4C_3}{3l}$.
Summing up the above inequality yields

$$\sum_{k=k_0}^{\infty}\|x_{k+1} - x_k\| \leq \frac{1}{2}\|x_{k_0} - x_{k_0-1}\| + \frac{3}{2}C_4\theta(H(x_{k_0}, x_{k_0-1}) - H(x^*, x^*)) < \infty.$$

Since $\mathbb{R}^N$ is complete, the above implies that $(x_n)_n$ is a convergent sequence. The proof is thereby completed. ∎

Remark 5.2 If in addition to the assumptions of the above theorem, we assume that f is coercive, then the boundedness of the sequence $(x_n)_n$ is guaranteed. On the other hand, if f is supposed further to be convex, then the last assumption on f is automatically satisfied.

In the above theorem, if the desingularizing function of H is of the form $s \mapsto cs^{1-\xi}, \xi \in (0, 1/2]$, we can have a result concerning convergence rate of the sequence $(x_n)_n$. More specifically,

Proposition 5.2 *Under the assumptions of Theorem 5.5, we assume furthermore that the desingularizing function of H is of the form $s \mapsto cs^{1-\xi}$, where $\xi \in (0, 1/2]$. Then there exist $c > 0$ and $Q \in [0, 1)$ such that*

$$\|x_k - x_\infty\| \leq cQ^k.$$

Proof. Recall from the previous proof that we have for k large enough

$$\|x_{k+1} - x_k\| \leq \frac{1}{3}\|x_k - x_{k-1}\| + C_4\Delta_k.$$

Summing up this inequality, we get

$$\sum_{p=k}^{\infty} \|x_{p+1} - x_p\| \leq \frac{1}{2}\|x_k - x_{k-1}\| + \frac{3}{2}C_4\theta(H(x_k, x_{k-1}) - H(x^*, x^*)).$$

Set $m_k = \sum_{p=k-1}^{\infty} \|x_{p+1} - x_p\|$. We can express the above in terms of m_k as follows

$$m_{k+1} \leq \frac{1}{2}(m_k - m_{k+1}) + \frac{3}{2}C_4\theta(H(x_k, x_{k-1}) - H(x^*, x^*)).$$

According to Lemma 6 of [56], we have for k large enough

$$\theta'(H(x_k, x_{k-1}) - H(x^*, x^*)) \geq \frac{1}{\|\nabla H(x_k, x_{k-1})\|}.$$

In view of the form of θ, this implies that there exists $C_5 > 0$ such that

$$(H(x_k, x_{k-1}) - H(x^*, x^*))^{\xi} \leq C_5\|\nabla H(x_k, x_{k-1})\| \leq C_5C_3\Big(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|\Big).$$

Further we have

$$\theta(H(x_k, x_{k-1}) - H(x^*, x^*)) = c.(H(x_k, x_{k-1}) - H(x^*, x^*))^{1-\xi}.$$

Therefore we have there exists $C_6 > 0$ such that

$$\begin{aligned}\theta(H(x_k, x_{k-1}) - H(x^*, x^*)) &\leq C_6\Big(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|\Big)^{\frac{1-\theta}{\theta}} \\ &\leq C_6(m_{k+1} - m_{k+2} + m_k - m_{k+1})^{\frac{1-\theta}{\theta}}.\end{aligned}$$

Coming back to the estimation of m_k we have

$$m_{k+1} \leq \frac{1}{2}(m_k - m_{k+1}) + \frac{3}{2}C_4C_6(m_k - m_{k+2})^{\frac{1-\theta}{\theta}}.$$

Since $\theta \in (0, 1/2]$, this follows $\frac{1-\theta}{\theta} \geq 1$. Hence, we get from the above inequality there exists $a > 0, b > 0$ such that

$$m_{k+1} \leq a(m_k - m_{k+1}) + b(m_k - m_{k+2}).$$

Equivalently

$$(1 + a)m_{k+1} - (a + b)m_k + bm_{k+2} \leq 0.$$

Consider the equation $bt^2 - (1 + a)t - (a + b) = 0$. Clearly this equation has two distinct solutions with opposite signs t_1 and t_2. WLOG, we assume that $t_1 < 0$. Therefore

$$m_{k+2} - (t_1 + t_2)m_{k+1} + t_1t_2m_k \leq 0.$$

This follows that

$$m_{k+2} - t_1m_{k+1} \leq t_2(m_{k+1} - t_1m_k).$$

Based the induction principle, this follows that there exist $c > 0$ and $Q \in [0, 1)$ such that for k large enough

$$m_{k+1} - t_1m_k \leq cQ^k,$$

which, due to $t_1 < 0$, implies $m_{k+1} \leq cQ^k$. By noticing the following inequality

$$m_{k+1} = \sum_{p=k}^{\infty} \|x_{p+1} - x_p\| \geq \|x_k - x_\infty\|,$$

the proof is completed. ∎

5.4 Errors, perturbations

Let us examine the effect of introducing errors, and then external perturbations.

5.4.1 Errors

When $\nabla f(x_k)$ is evaluated with an exogenous additive error e_k, the algorithm becomes

(IPAHDD-C1-errors)

$$y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta(\nabla f(x_{k-1}) + e_{k-1}).$$
$$x_{k+1} = x_k - \beta h(\nabla f(x_k) + e_k) + h\,\mathrm{prox}_{\frac{h}{1+\gamma h}\varphi}\left(\frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta-1)h}{1+\gamma h}(\nabla f(x_k) + e_k)\right).$$

The perturbed algorithm (IPAHDD-C1-errors) incorporates exogenous additive errors denoted by e_k. In each iteration, k, we calculate y_k and x_{k+1} using the current, and past states and gradients, along with exogenous errors. This refinement helps the algorithm to handle situations where gradient calculations include added errors, which often occur in real-world scenarios where exact gradient computations face inaccuracies. As a result, the algorithm proves its adaptability and usefulness outside ideal situations.

Theorem 5.6 *Let's make the assumptions of Theorem 5.1 and suppose furthermore that the inequality condition on γ, β and h in Theorem 5.1 is strict. Suppose that the sequence $(e_k)_k$ of perturbations, errors satisfies:*

$$\sum_k \|e_k\|^2 < +\infty.$$

Then any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C1-errors) *satisfies*

(i) $\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k) = -\beta e_k$ *after a finite number of steps.*

(ii) $\sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < +\infty$ *and* $\sum_{k=1}^{+\infty} \|x_{k+1} - x_k\|^2 < +\infty$. *So* $\nabla f(x_k) \to 0$, $x_{k+1} - x_k \to 0$.

Proof. Let us reconstruct the dynamic with respect to y_k from which this perturbed algorithm is derived. By definition of y_k, we have

$$y_{k+1} = \frac{1}{h}(x_{k+1} - x_k) + \beta(\nabla f(x_k) + e_k). \tag{5.21}$$

Dividing the second equation of (IPAHDD-C1-errors) by h, and reformulating it in terms of y_{k+1}, we obtain

$$y_{k+1} = \mathrm{prox}_{\frac{h}{1+\gamma h}\varphi}\left(\frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta-1)h}{1+\gamma h}(\nabla f(x_k) + e_k)\right). \tag{5.22}$$

By definition of the proximal operator, this gives

$$y_{k+1} + \frac{h}{1+\gamma h}\partial\varphi(y_{k+1}) \ni \frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta-1)h}{1+\gamma h}(\nabla f(x_k) + e_k).$$

Equivalently

$$(1+\gamma h)(y_{k+1}-y_k)+h\partial\varphi(y_{k+1})+\gamma h y_k \ni (\gamma\beta-1)h(\nabla f(x_k)+e_k).$$

This gives

$$(y_{k+1}-y_k)+h\partial\varphi(y_{k+1})+\gamma h y_{k+1}-(\gamma\beta-1)h(\nabla f(x_k)+e_k)\ni 0.$$

According to (5.21) we obtain

$$(y_{k+1}-y_k)+h\partial\varphi(y_{k+1})+\gamma h\Big(\frac{1}{h}(x_{k+1}-x_k)+\beta(\nabla f(x_k)+e_k)\Big)-(\gamma\beta-1)h(\nabla f(x_k)+e_k)\ni 0.$$

After simplification we get

$$(y_{k+1}-y_k)+h\partial\varphi(y_{k+1})+\gamma(x_{k+1}-x_k)+h\nabla f(x_k)\ni -he_k. \tag{5.23}$$

The proof is now parallel to that of Theorem 5.1. Taking the scalar product of (5.23) with y_{k+1}, we obtain

$$\begin{aligned}\|y_{k+1}\|^2-\langle y_k,y_{k+1}\rangle+\gamma\langle x_{k+1}-x_k,\frac{1}{h}(x_{k+1}-x_k)+\beta(\nabla f(x_k)+e_k)\rangle+h\langle\partial\varphi(y_{k+1}),y_{k+1}\rangle\\ +h\langle\nabla f(x_k),\frac{1}{h}(x_{k+1}-x_k)+\beta(\nabla f(x_k)+e_k)\rangle=-h\langle e_k,y_{k+1}\rangle.\end{aligned}$$

According to the assumption (5.14) on the parameters, similar calculation as in Theorem 5.1 gives

$$\begin{aligned}&\frac{1}{2}\|y_{k+1}\|^2-\frac{1}{2}\|y_k\|^2+(\gamma\beta+1)(f(x_{k+1})-f(x_k))+\left(\frac{\gamma}{h}-\frac{L}{2}(\gamma\beta+1)\right)\|x_{k+1}-x_k\|^2\\ &+hr\|y_{k+1}\|+\beta h\|\nabla f(x_k)\|^2\le h\|e_k\|\|y_{k+1}\|+\gamma\beta\|e_k\|\|x_{k+1}-x_k\|+h\beta\|e_k\|\|\nabla f(x_k)\|.\end{aligned}$$

Equivalently

$$\begin{aligned}&E_{k+1}-E_k+hr\|y_{k+1}\|+\beta h\|\nabla f(x_k)\|^2+\left(\frac{\gamma}{h}-\frac{L}{2}(\gamma\beta+1)\right)\|x_{k+1}-x_k\|^2\\ &\le h\|e_k\|\|y_{k+1}\|+\gamma\beta\|e_k\|\|x_{k+1}-x_k\|+h\beta\|e_k\|\|\nabla f(x_k)\|,\end{aligned}$$

where

$$E_k:=\frac{1}{2}\|y_k\|^2+(\gamma\beta+1)\left(f(x_k)-\inf_{x\in H}f(x)\right).$$

We deduce that

$$\begin{aligned}
&E_{k+1} - E_k + h(r - \|e_k\|)\|y_{k+1}\| + \beta h\|\nabla f(x_k)\|^2 + \left(\frac{\gamma}{h} - \frac{L}{2}(\gamma\beta + 1)\right)\|x_{k+1} - x_k\|^2 \\
&\leq \gamma\beta\|e_k\|\|x_{k+1} - x_k\| + h\beta\|e_k\|\|\nabla f(x_k)\| \\
&\leq \frac{1}{2}\left(\frac{\gamma}{h} - \frac{L}{2}(\gamma\beta + 1)\right)\|x_{k+1} - x_k\|^2 + \frac{\gamma^2\beta^2}{2\left(\frac{\gamma}{h} - \frac{L}{2}(\gamma\beta + 1)\right)}\|e_k\|^2 \\
&+\frac{1}{2}\beta h\|\nabla f(x_k)\|^2 + \frac{1}{2}\beta h\|e_k\|^2
\end{aligned}$$

Since $e_k \to 0$, we obtain the existence of a constant $C > 0$ such that for k sufficiently large

$$E_{k+1} - E_k + \frac{hr}{2}\|y_{k+1}\| + \frac{1}{2}\beta h\|\nabla f(x_k)\|^2 + \frac{1}{2}\left(\frac{\gamma}{h} - \frac{L}{2}(\gamma\beta + 1)\right)\|x_{k+1} - x_k\|^2 \leq C\|e_k\|^2.$$

By summing the above inequalities we deduce that

$$\sum_{k=1}^{+\infty}\|\nabla f(x_k)\|^2 < +\infty, \text{ and } \sum_{k=1}^{+\infty}\|y_k\| < +\infty. \tag{5.24}$$

Let us now prove that after a finite number of steps, the sequence $(x_k)_k$ follows the steepest descent method. The proof relies on Lemma 5.1. Recall that, according to (5.22), we have $y_{k+1} = \operatorname{prox}_{\frac{h}{1+\gamma h}\varphi}(z_k)$, where

$$z_k = \left(\frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta - 1)h}{1+\gamma h}(\nabla f(x_k) + e_k)\right).$$

According to (5.24), since the general term of a convergent series necessarily tends towards zero, we have that $\lim_k \nabla f(x_k) = \lim_k y_k = 0$. Therefore, according to the definition of z_k, and since e_k tends to zero, we have $\lim_k z_k = 0$. So, there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$,

$$\|z_k\| \leq \tfrac{hr}{1+\gamma h}.$$

By Lemma 5.1, this implies that $y_{k+1} = \operatorname{prox}_{\frac{h}{1+\gamma h}\varphi}(z_k) = 0$ for all $k \geq k_0$. Equivalently, $\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k) = -\beta e_k$ for all $k \geq k_0$, which means that after a finite number of steps, the sequence (x_k) follows a perturbed steepest descent algorithm. This ends the proof. ∎

As a consequence of Theorem 5.6, and of the properties of the perturbed steepest descent [120], we obtain the following convergence result.

Corollary 5.1 *Under the summability assumption* $\sum_k \|e_k\| < +\infty$*, there is convergence of the sequences* (x_k) *generated by* (IPAHDD-C1-errors) *in the convex case, and in the nonconvex case finite dimensional case under* (KL)*.*

5.4.2 External perturbation

The algorithm (IPAHDD-C1) enjoys remarkable strutural stability. Consider the perturbed version of (5.1)

$$\ddot{x}(t)+\gamma\dot{x}(t)+\partial\varphi\Big(\dot{x}(t)+\beta\nabla f(x(t))\Big)+\beta\nabla^2 f(x(t))\dot{x}(t)+\nabla f(x(t))\ni e(t),$$

where the right-hand side $e(\cdot)$ takes into account an external perturbation, of forcing term. A temporal discretization similar to that in Section 5.2 gives

$$\frac{1}{h^2}(x_{k+1}-2x_k+x_{k-1})+\frac{\gamma}{h}(x_{k+1}-x_k)+\partial\varphi(\frac{1}{h}(x_{k+1}-x_k)+\beta\nabla f(x_k))\\ +\frac{\beta}{h}(\nabla f(x_k)-\nabla f(x_{k-1}))+\nabla f(x_k)\ni e_k. \quad (5.25)$$

Solving the above inclusion with respect to x_{k+1} gives the following algorithm:

(IPAHDD-C1-pert)

$y_k=\frac{1}{h}(x_k-x_{k-1})+\beta\nabla f(x_{k-1})$.

$x_{k+1}=x_k-\beta h\nabla f(x_k)+h\,\mathrm{prox}_{\frac{h}{1+\gamma h}\varphi}\left(\frac{1}{1+\gamma h}y_k+\frac{(\gamma\beta-1)h}{1+\gamma h}\nabla f(x_k)+\frac{h}{1+\gamma h}e_k\right)$.

In the (IPAHDD-C1-pert) algorithm, we include e_k to account for unexpected changes or disturbances, often called external perturbations. These disturbances might come from measurement noise, unpredictable data, or random elements in the optimization process itself, common in real-world situations. By including e_k, our algorithm can better handle these unexpected changes. This makes the algorithm stronger and more flexible, allowing it to work well in real-world applications, even when conditions are not perfect.
We have the following convergence results for this perturbed version of (IPAHDD-C1), where we emphasize that the convergence results hold under a very weak assumption on the perturbation terms.

Theorem 5.7 *Let's make the assumptions of Theorem 5.1, and suppose that the sequence $(e_k)_k$ of perturbations satisfies:*

$$\lim_k\|e_k\|=0 \quad as \quad k\to+\infty.$$

Then any sequence $(x_k)_k$ generated by (IPAHDD-C1-pert) *satisfies the following properties:*

(i) $\frac{1}{h}(x_{k+1}-x_k)+\beta\nabla f(x_k)=0$ *after a finite number of steps.*

(ii) $\sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < +\infty$ *and* $\sum_{k=1}^{+\infty} \|x_{k+1} - x_k\|^2 < +\infty$. *So* $\nabla f(x_k) \to 0$, $x_{k+1} - x_k \to 0$.

(iii) *The sequence* (x_k) *converges in the following cases:*

(a) f *convex with* $\operatorname{argmin}_{\mathcal{H}} f \neq \emptyset$. *Then* $(x_k)_k$ *converges weakly, and its limit is a minimizer of* f.

(b) $f : \mathbb{R}^N \to \mathbb{R}$ *satisfies the* (KL) *property. Then* $(x_k)_k$ *converges to a critical point of* f.

Proof. The proof is a direct adaptation of Theorem 5.1. The basic energy estimate becomes

$$E_{k+1} - E_k + (hr - \|e_k\|)\|y_{k+1}\| + \beta h\|\nabla f(x_k)\|^2 \leq 0, \tag{5.26}$$

where

$$E_k := \frac{1}{2}\|y_k\|^2 + (\gamma\beta + 1)\left(f(x_k) - \inf_{x\in H} f(x)\right).$$

Since $\lim_k \|e_k\| = 0$, we deduce that for k large enough,

$$E_{k+1} - E_k + \frac{1}{2}hr\|y_{k+1}\| + \beta h\|\nabla f(x_k)\|^2 \leq 0. \tag{5.27}$$

We conclude by similar arguments as in Theorem 5.1. ∎

Remark 5.3 The above result suggests that, when combined with approximation or variation of the data (f, φ, γ, β), the algorithm (IPAHDD-C) converges under minimal assumptions, much weaker than the standard ones based on summability properties. In particular, in order for FISTA or the gradient descent to preserve the convergence results in the presence of perturbations, it is common practice to impose some stringent summability conditions on the perturbations which might fail to be satisfied in practice. For instance, FISTA when perturbed, has the same convergence results as in the error-free case as long as the perturbations satisfy $\sum k\|e_k\| < +\infty$. On the other hand, the corresponding condition for our algorithm is only that the perturbation sequence (e_k) converges to 0, which can cover various cases where $\sum k\|e_k\| < +\infty$ does not satisfy, for example when $\|e_k\| = 1/k$.

5.5 Variants using Nesterov extrapolation method

We construct algorithms, still obtained by temporal discretizations of the differential inclusion

$$\ddot{x}(t) + \gamma\dot{x}(t) + \partial\varphi\Big(\dot{x}(t) + \beta\nabla f(x(t))\Big) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) \ni 0,$$

and which have an analogous structure to the accelerated gradient method of Nesterov [114, 115]. Specifically, we consider the following discretization of the dynamic

$$\frac{1}{h^2}(x_{k+1}-2x_k+x_{k-1})+\frac{\gamma}{h}(x_{k+1}-x_k)+\partial\varphi(\frac{1}{h}(x_{k+1}-x_k)+\beta\nabla f(x_k)) \\ +\frac{\beta}{h}(\nabla f(x_k)-\nabla f(x_{k-1}))+\nabla f(z_k)\ni 0. \quad (5.28)$$

There is some flexibility in the choice of the point z_k where the gradient of f is computed. By taking $z_k = x_k$, we obtain the algorithm (IPAHDD-C1) studied in section 5.2. In this section, we consider two different choices for z_k, which are in accordance with the structure of the extrapolation step in the Nesterov accelerated gradient method (but here the extrapolation coefficient is fixed, taken less than one):

5.5.1 Case 1

Take $z_k = x_k + \frac{1}{1+\gamma h}(x_k - x_{k-1})$. With this choice of z_k in (5.28), elementary calculation gives the following algorithm:

(IPAHDD-C2)

Initialize : $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$.

$z_k = x_k + \frac{1}{1+\gamma h}(x_k - x_{k-1})$.

$w_k = \frac{1}{h}(z_k - x_k) + \frac{\beta}{1+\gamma h}\nabla f(x_{k-1}) + \frac{h\beta\gamma}{1+\gamma h}\nabla f(x_k) - \frac{h}{1+\gamma h}\nabla f(z_k)$

$x_{k+1} = x_k - \beta h\nabla f(x_k) + h\,\mathrm{prox}_{\frac{h}{1+\gamma h}\varphi}(w_k)$.

Theorem 5.8 *Let $f : \mathcal{H} \to \mathbb{R}$ be a differentiable function whose gradient is L-Lipschitz continuous, and such that $\inf_{\mathcal{H}} f > -\infty$. Assume that the friction potential function $\varphi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive parameters h, γ, β satisfy the relation*

$$\begin{cases} \gamma > \max\{2hL, L/2\}, \\ \beta < \min\left\{\frac{\gamma+\gamma^2 h-2Lh}{Lh}, \frac{2+(2\gamma-L)h}{\gamma^2 h+\gamma}\right\}. \end{cases}$$

Then any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C2) *satisfies the following properties:*

(i) $\frac{1}{h}(x_{k+1}-x_k)+\beta\nabla f(x_k)=0$ *after a finite number of steps.*
(ii) $\sum_{k=1}^{+\infty}\|\nabla f(x_k)\|^2 < +\infty$ *and* $\sum_{k=1}^{+\infty}\|x_{k+1}-x_k\|^2 < +\infty$.

Proof. Let us rewrite (5.28) with the help of $y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})$. Equivalently, we have

$$y_{k+1} - y_k + \gamma(x_{k+1} - x_k) + h\partial\varphi(y_{k+1}) + h\nabla f(z_k) \ni 0.$$

By taking the scalar product of the above inclusion with y_{k+1} we obtain

$$\|y_{k+1}\|^2 - \langle y_k, y_{k+1}\rangle + \gamma\langle x_{k+1} - x_k, y_{k+1}\rangle + h\langle\partial\varphi(y_{k+1}), y_{k+1}\rangle + h\langle\nabla f(z_k), y_{k+1}\rangle = 0. \tag{5.29}$$

We can easily check that

$$\gamma\langle x_{k+1} - x_k, y_{k+1}\rangle = \frac{\gamma h}{2}\|y_{k+1}\|^2 + \frac{\gamma}{2h}\|x_{k+1} - x_k\|^2 - \frac{\gamma h\beta^2}{2}\|\nabla f(x_k)\|^2. \tag{5.30}$$

According to the L-Lipschitz continuity of ∇f, we have

$$\begin{aligned}
&h\langle\nabla f(z_k), y_{k+1}\rangle = \\
&h\langle\nabla f(z_k) - \nabla f(x_k), y_{k+1}\rangle + h\langle\nabla f(x_k), y_{k+1}\rangle \\
&\geq \frac{-hL}{1+\gamma h}\|x_k - x_{k-1}\|\|y_{k+1}\| + h\langle\nabla f(x_k), y_{k+1}\rangle \\
&= \frac{-h^2L}{1+\gamma h}\|y_k - \beta\nabla f(x_{k-1})\|\|y_{k+1}\| + h\langle\nabla f(x_k), y_{k+1}\rangle \\
&\geq \frac{-h^2L}{1+\gamma h}\|y_k\|\|y_{k+1}\| - \frac{h^2L\beta}{1+\gamma h}\|\nabla f(x_{k-1})\|\|y_{k+1}\| + h\langle\nabla f(x_k), y_{k+1}\rangle \\
&\geq \frac{-h^2L}{1+\gamma h}\|y_k\|\|y_{k+1}\| - \frac{h^2L\beta}{2(1+\gamma h)}(\|\nabla f(x_{k-1})\|^2 + \|y_{k+1}\|^2) + h\langle\nabla f(x_k), y_{k+1}\rangle.
\end{aligned}$$

Moreover, according to the gradient descent lemma

$$\begin{aligned}
h\langle\nabla f(x_k), y_{k+1}\rangle &= \beta h\|\nabla f(x_k)\|^2 + \langle\nabla f(x_k), x_{k+1} - x_k\rangle \\
&\geq \beta h\|\nabla f(x_k)\|^2 + f(x_{k+1}) - f(x_k) - \frac{L}{2}\|x_{k+1} - x_k\|^2.
\end{aligned}$$

By combining the two estimates above, we obtain

$$\begin{aligned}
h\langle\nabla f(z_k), y_{k+1}\rangle \geq &\frac{-h^2L}{1+\gamma h}\|y_k\|\|y_{k+1}\| - \frac{h^2L\beta}{2(1+\gamma h)}(\|\nabla f(x_{k-1})\|^2 + \|y_{k+1}\|^2) \\
&+ \beta h\|\nabla f(x_k)\|^2 + f(x_{k+1}) - f(x_k) - \frac{L}{2}\|x_{k+1} - x_k\|^2.
\end{aligned} \tag{5.31}$$

By combining (5.29), (5.30) and (5.31), and using the dry friction property $\varphi(u) \geq r\|u\|$,

we obtain

$$\|y_{k+1}\|^2 - \langle y_k, y_{k+1}\rangle + \frac{\gamma h}{2}\|y_{k+1}\|^2 + \frac{\gamma}{2h}\|x_{k+1}-x_k\|^2 - \frac{\gamma h\beta^2}{2}\|\nabla f(x_k)\|^2 + hr\|y_{k+1}\|$$
$$- \frac{h^2L}{1+\gamma h}\|y_k\|\|y_{k+1}\| - \frac{h^2L\beta}{2(1+\gamma h)}(\|\nabla f(x_{k-1})\|^2 + \|y_{k+1}\|^2)$$
$$+ \beta h\|\nabla f(x_k)\|^2 + f(x_{k+1}) - f(x_k) - \frac{L}{2}\|x_{k+1}-x_k\|^2 \leq 0.$$

Therefore,

$$(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_{k+1}\|^2 - (1 + \frac{h^2L}{1+\gamma h})\|y_k\|\|y_{k+1}\| + (\frac{\gamma}{2h} - \frac{L}{2})\|x_{k+1}-x_k\|^2$$
$$+ (\beta h - \frac{\gamma h\beta^2}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|\nabla f(x_k)\|^2 + \frac{h^2L\beta}{2(1+\gamma h)}(\|\nabla f(x_k)\|^2 - \|\nabla f(x_{k-1})\|^2)$$
$$+ f(x_{k+1}) - f(x_k) + hr\|y_{k+1}\| \leq 0.$$

For each $k \geq 1$ set

$$E_k := \frac{1}{2}(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_k\|^2 + \frac{h^2L\beta}{2(1+\gamma h)}\|\nabla f(x_{k-1})\|^2 + f(x_k) - \inf_{\mathcal{H}} f. \quad (5.32)$$

We deduce that

$$E_{k+1} - E_k + \frac{1}{2}(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_{k+1}\|^2 - (1 + \frac{h^2L}{1+\gamma h})\|y_k\|\|y_{k+1}\|$$
$$+ \frac{1}{2}(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_k\|^2 + (\frac{\gamma}{2h} - \frac{L}{2})\|x_{k+1}-x_k\|^2$$
$$+ (\beta h - \frac{\gamma h\beta^2}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|\nabla f(x_k)\|^2 + hr\|y_{k+1}\| \leq 0.$$

According to the assumptions on γ, h and β, we have

$$\begin{cases} \frac{\gamma}{2h} - \frac{L}{2} \geq 0, \\ \beta h - \frac{\gamma h\beta^2}{2} - \frac{h^2L\beta}{2(1+\gamma h)} > 0. \end{cases}$$

Let us show that

$$\frac{1}{2}(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_{k+1}\|^2 - (1 + \frac{h^2L}{1+\gamma h})\|y_k\|\|y_{k+1}\|$$
$$+ \frac{1}{2}(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_k\|^2 \geq 0.$$

Indeed, a sufficient condition for this is

$$\begin{cases} 1+\frac{\gamma h}{2}-\frac{h^2L\beta}{2(1+\gamma h)}>0, \\ (1+\frac{h^2L}{1+\gamma h})^2-(1+\frac{\gamma h}{2}-\frac{h^2L\beta}{2(1+\gamma h)})^2\leq 0. \end{cases}$$

which is equivalent to (since $\gamma>0, h>0, \beta>0$)

$$1+\frac{h^2L}{1+\gamma h}\leq 1+\frac{\gamma h}{2}-\frac{h^2L\beta}{2(1+\gamma h)},$$

or

$$\beta\leq\frac{\gamma+\gamma^2h-2Lh}{Lh}, \tag{5.33}$$

which is fullfiled, according to our assumptions on γ, h and β. We have shown that

$$E_{k+1}-E_k+(\beta h-\frac{\gamma h\beta^2}{2}-\frac{h^2L\beta}{2(1+\gamma h)})\|\nabla f(x_k)\|^2+hr\|y_{k+1}\|\leq 0, \tag{5.34}$$

where E_k has been defined in (5.32). By summing the above inequalities, we obtain

$$\sum_{k=1}^{+\infty}\|\nabla f(x_k)\|^2<+\infty, \quad \sum_{k=1}^{+\infty}\|y_k\|<+\infty. \tag{5.35}$$

Let us now prove that after a finite number of steps, the sequence $(x_k)_k$ follows the steepest descent method. The proof relies on Lemma 5.1. Recall the following equivalent formulation of (IPAHDD-C2)

$$y_{k+1}=\operatorname{prox}_{\frac{h}{1+\gamma h}\varphi}(w_k),$$

where

$$w_k=\frac{1}{h}(z_k-x_k)+\frac{\beta}{1+\gamma h}\nabla f(x_{k-1})+\frac{h\beta\gamma}{1+\gamma h}\nabla f(x_k)-\frac{h}{1+\gamma h}\nabla f(z_k).$$

According to (5.35), and since the general term of a convergent series necessarily goes to zero, we have that $\lim_k\nabla f(x_k)=\lim_k y_k=0$. By definition of y_k this implies $\lim_k(x_k-x_{k-1})=0$, and hence $\lim_k(z_k-x_k)=0$. According to the Lipschitz continuity of ∇f, we have

$$\begin{aligned}\|\nabla f(z_k)\| &\leq \|\nabla f(z_k)-\nabla f(x_k)\|+\|\nabla f(x_k)\| \\ &\leq L\|z_k-x_k\|+\|\nabla f(x_k)\|\longrightarrow 0 \text{ as } k \text{ tends to } +\infty.\end{aligned}$$

Taking into account all these observations, we easily deduce that $\lim_k w_k = 0$. Therefore, there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$,

$$\|w_k\| \leq \tfrac{hr}{1+\gamma h}.$$

According to Lemma 5.1, this implies that $y_{k+1} = \operatorname{prox}_{\frac{h}{1+\gamma h}\varphi}(w_k) = 0$ for all $k \geq k_0$. Equivalently, $\frac{1}{h}(x_{k+1} - x_k) + \beta \nabla f(x_k) = 0$ which means that after a finite number of steps, the sequence (x_k) follows the steepest descent algorithm. This completes the proof. ■

5.5.2 Case 2

Take $z_k = x_k + \frac{1}{h(1+\gamma h)}(x_k - x_{k-1})$ in (5.28). With this choice of z_k, elementary calculation gives the following algorithm:

(IPAHDD-C3)

Initialize : $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$.

$$z_k = x_k + \tfrac{1}{h(1+\gamma h)}(x_k - x_{k-1}).$$
$$w_k = z_k - x_k + \tfrac{\beta}{1+\gamma h}\nabla f(x_{k-1}) + \tfrac{h\beta\gamma}{1+\gamma h}\nabla f(x_k) - \tfrac{h}{1+\gamma h}\nabla f(z_k)$$
$$x_{k+1} = x_k - \beta h \nabla f(x_k) + h \operatorname{prox}_{\frac{h}{1+\gamma h}\varphi}(w_k).$$

A similar proof to the one of Theorem 5.8 gives

Theorem 5.9 *Let $f : \mathcal{H} \to \mathbb{R}$ be a differentiable function whose gradient is L-Lipschitz continuous, and such that $\inf_{\mathcal{H}} f > -\infty$. Assume that the friction potential function $\varphi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive parameters h, γ, β satisfy the relation*

$$\begin{cases} \gamma > \max\left\{\frac{L}{2h}, 2L, Lh\right\}, \\ \beta < \min\left\{\frac{2+2\gamma h - L}{\gamma(1+\gamma h)}, \frac{\gamma + h\gamma^2 - 2L}{L}\right\}. \end{cases}$$

Then any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C3) *satisfies the following properties:*

(i) $\frac{1}{h}(x_{k+1} - x_k) + \beta \nabla f(x_k) = 0$ *after a finite number of steps.*

(ii) $\sum_{k=1}^{\infty} \|\nabla f(x_k)\|^2 < +\infty$ *and* $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\|^2 < +\infty$.

Remark 5.4 As an immediate consequence of Theorem 5.8 and 5.9, and of the classical properties of the steepest descent method, we obtain the convergence of the sequence (x_k) in the convex case, and in the nonconvex case under (KL). Similar results are still valid for the perturbed version of these algorithms.

Remark 5.5 In Theorems 5.8 and 5.9, a crucial assumption is $\gamma > \max\{2hL, L/2\}$, resp. $\gamma > \max\{L/2h, 2L, Lh\}$. Thus the viscous damping coefficient γ must remain sufficiently large. The above approach excludes the case where the viscous damping tends asymptotically to zero. It seems difficult to combine dry friction with the Nesterov accelerated gradient method because dry friction involves a finite time stabilization property whereas Nesterov method is based on the asymptotic vanishing of the damping coefficient.

5.6 Nonsmooth problems

We consider the extension of our study to two nonsmooth situations: the nonsmooth convex case, and the nonsmooth d.c. optimization.

5.6.1 Nonsmooth convex case

Suppose that $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a closed, convex and proper function such that $\operatorname{argmin}_{\mathcal{H}} f \neq \emptyset$. We will reduce to the smooth case by means of the Moreau-Yosida approximation of f. Recall that the Moreau envelope of f of index $\lambda > 0$ is the function $f_\lambda : \mathcal{H} \to \mathbb{R}$ defined by, for all $x \in \mathcal{H}$,

$$f_\lambda(x) = \min_{\xi \in \mathcal{H}} \left\{ f(\xi) + \tfrac{1}{2\lambda}\|x - \xi\|^2 \right\}.$$

As a classical result, f_λ is convex, differentiable and its gradient is $\frac{1}{\lambda}$-Lipschitz continuous. Moreover, we have $\operatorname{argmin}_{\mathcal{H}} f = \operatorname{argmin}_{\mathcal{H}} f_\lambda$ and $\min_{\mathcal{H}} f = \min_{\mathcal{H}} f_\lambda$. One can consult [29, 51, 62] for an in-depth study of the properties of the Moreau envelope in a Hilbert framework. Exploiting this property of the Moreau envelope, we can equivalently consider the problem in which f is substituted by its Moreau envelope, and hence we recover the smooth case. Since $\nabla f_\lambda(x) = \frac{1}{\lambda}(x - \operatorname{prox}_{\lambda f}(x))$, we obtain the following algorithm:

(IPAHDD-C-nonsmooth)

Initialize : $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$.

$$y_k = \tfrac{1}{h}(x_k - x_{k-1}) + \tfrac{\beta}{\lambda}(x_k - \operatorname{prox}_{\lambda f}(x_{k-1}))$$
$$w_k = \tfrac{1}{1+\gamma h} y_k + \tfrac{(\gamma\beta - 1)h}{(1+\gamma h)\lambda}(x_k - \operatorname{prox}_{\lambda f}(x_k))$$
$$x_{k+1} = x_k - \tfrac{\beta h}{\lambda}(x_k - \operatorname{prox}_{\lambda f}(x_k)) + h \operatorname{prox}_{\frac{h}{1+\gamma h}\varphi}(w_k)$$

The two nonsmooth functions f and φ enter the algorithm via their proximal mappings. In addition, these proximal steps are computed independently, which makes the algorithm (IPAHDD-C-nonsmooth) a splitting algorithm. Based on the properties of the Moreau

envelope, a direct adaptation of Theorem 5.1 gives the following convergence results for (IPAHDD-C-nonsmooth).

Theorem 5.10 *Let $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a closed, convex, proper function such that $\operatorname{argmin}_{\mathcal{H}} f \neq \emptyset$. Assume that the friction potential function $\varphi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive parameters $h, \gamma, \beta, \lambda$ satisfy the relation*

$$\frac{\gamma}{h} - \frac{1}{2\lambda}(\gamma\beta + 1) \geq 0.$$

Then any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C-nonsmooth) *converges weakly and its limit is a minimizer of f. Moreover,*

(i) $\frac{1}{h}(x_{k+1} - x_k) + \frac{\beta}{\lambda}(x_k - \operatorname{prox}_{\lambda f}(x_k)) = 0$ *after a finite number of steps;*

(ii) $\sum_{k=1}^{+\infty} \|x_k - \operatorname{prox}_{\lambda f}(x_k)\|^2 < +\infty$.

Proof. By replacing the Lipschitz constant L in Theorem 5.1 by $\frac{1}{\lambda}$, and using the equality $\nabla f_\lambda(x_k) = \frac{1}{\lambda}(x_k - \operatorname{prox}_{\lambda f}(x_k))$, the result follows immediately. ∎

Remark 5.6 It is worth mentioning that under the assumptions on the parameters in the above theorem, the first item is a relaxed proximal point algorithm

$$x_{k+1} = x_k + \frac{h\beta}{\lambda}(\operatorname{prox}_{\lambda f}(x_k) - x_k),$$

where the relaxation parameter $h\beta/\lambda < 2$. Since $\operatorname{prox}_{\lambda f}$ is firmly nonexpansive, the convergence can be also derived within the theory of Krasnosel'skiĭ-Mann iteration (see Corollary 5.15 in [51])

5.6.2 Nonsmooth nonconvex d.c. problems

Suppose that $f = g - h$ where $g, h : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are closed, convex and proper functions. Following Hiriart-Urruty [86], consider the problem in which f is substituted by the difference of the Moreau envelopes of g and h, so recovering the smooth case. Given $\lambda > 0$, according to the properties of the Moreau envelope, the regularized function $\psi_\lambda : \mathcal{H} \to \mathbb{R}$ defined by

$$\psi_\lambda = g_\lambda - h_\lambda,$$

is differentiable and its gradient is $\frac{2}{\lambda}$ Lipschitz continuous. Moreover, if x is a critical point of ψ_λ, we have

$$\begin{aligned} \nabla\psi_\lambda(x) &= \nabla g_\lambda(x) - \nabla h_\lambda(x) \\ &= -\frac{1}{\lambda}\left(\operatorname{prox}_{\lambda g}(x) - \operatorname{prox}_{\lambda h}(x)\right) = 0. \end{aligned}$$

Therefore, $u := \text{prox}_{\lambda g}(x) = \text{prox}_{\lambda h}(x)$, and the point u, which is so defined, verifies $\partial g(u) - \partial h(u) \ni 0$, which is a critical point of $f = g - h$ in the sense of Toland [144]. The algorithm now writes

(IPAHDD-CDC)

Initialize : $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$.

$$y_k = \tfrac{1}{h}(x_k - x_{k-1}) - \tfrac{\beta}{\lambda}(\text{prox}_{\lambda g}(x_{k-1}) - \text{prox}_{\lambda h}(x_{k-1}))$$

$$x_{k+1} = x_k + \tfrac{\beta h}{\lambda}(\text{prox}_{\lambda g}(x_k) - \text{prox}_{\lambda h}(x_k)) + h\,\text{prox}_{\frac{h}{1+\gamma h}\varphi}\left(\tfrac{1}{1+\gamma h}y_k - \tfrac{(\gamma\beta-1)h}{(1+\gamma h)\lambda}(\text{prox}_{\lambda g}(x_k) - \text{prox}_{\lambda h}(x_k))\right)$$

According to the above results, a direct adaptation of Theorem 5.1 gives the following result:

Theorem 5.11 *Let $f = g - h$ where $g, h : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are closed, convex and proper functions. Assume that the friction potential $\varphi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\text{DF})_r$ for some $r > 0$. Take $\lambda > 0$, and suppose that the positive parameters h, γ, β satisfy the relation*

$$\frac{h}{\lambda} \leq \frac{\gamma}{\gamma\beta + 1}. \tag{5.36}$$

Then, for any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-CDC)*, we have that $(x_k)_k$ satisfies*

(i) $\frac{1}{h}(x_{k+1} - x_k) + \beta(\nabla g_\lambda(x_k) - \nabla h_\lambda(x_k)) = 0$ *after a finite number of steps.*

(ii) $\sum_{k=1}^{\infty} \|\nabla g_\lambda(x_k) - \nabla h_\lambda(x_k)\|^2 < +\infty$ *and* $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\|^2 < +\infty$.

(iii) *If $\mathcal{H}$ is a finite dimensional space, and $g_\lambda - h_\lambda$ verifies the (KL) property, then the sequence (x_k) converges to some x_∞ such that $u := \text{prox}_{\lambda g}(x_\infty) = \text{prox}_{\lambda h}(x_\infty)$ is a critical point in the sense of Toland of $f = g - h$, i.e. ,*

$$\partial g(u) - \partial h(u) \ni 0.$$

Remark 5.7 As a particular case of practical importance, suppose that g and h are convex functions which are semialgebraic. Then their Moreau envelopes are still semialgebraic [26], and so is the difference of their Moreau envelopes. In this case, we have that $g_\lambda - h_\lambda$ verifies the (KL) property, and so the above convergence result is valid in this nonsmooth nonconvex situation.

5.7 Splitting algorithms for the Lasso-type problems

Take $\mathcal{H} = \mathbb{R}^n$. We consider Lasso-type splitting algorithms for additively structured minimization problems. The function f to be minimized is written as

$$f(x) = \tfrac{1}{2}\|Ax - b\|^2 + g(x),$$

where A is an $m \times n$ matrix, $b \in \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a closed, convex proper function.

A direct application of the nonsmooth algorithm (IPAHDD-C-nonsmooth) to this minimization problem would require calculating (at least approximately) the proximal operator of f. It's not easy in general. To overcome this difficulty, we use a change of metric, a technique already used in [3], [34]. For a symmetric and positive definite matrix $M \in \mathbb{R}^{n\times n}$, we denote by $\langle \cdot, \cdot \rangle_M = \langle M\cdot, \cdot \rangle$ the scalar product on $\mathbb{R}^n$ induced by M, and by $\|\cdot\|_M$ the associated norm. For a given closed, convex function f, the Moreau's envelope of index $\lambda > 0$ associated with the metric induced by M is the function $f_\lambda^M : \mathcal{H} \to \mathbb{R}$ defined by, for $x \in \mathbb{R}^n$,

$$f_\lambda^M(x) = \min_{\xi\in\mathcal{H}} \left\{ f(\xi) + \tfrac{1}{2\lambda}\|x - \xi\|_M^2 \right\}.$$

The Moreau envelope f_λ^M is a smooth function whose gradient for the Euclidean structure is given by

$$\nabla f_\lambda^M(x) = \frac{1}{\lambda} M(x - \operatorname{prox}_{\lambda f}^M(x)), \tag{5.37}$$

where $\operatorname{prox}_{\lambda f}^M(x) = \operatorname{argmin}_{\xi\in\mathcal{H}} \left\{ f(\xi) + \frac{1}{2\lambda}\|x - \xi\|_M^2 \right\}$. As a classical result, ∇f_λ^M is $\frac{1}{\lambda}$-Lipschitz continuous for the norm $\|\cdot\|_M$. From this, by using classical linear algebra, we easily deduce that

$$\|\nabla f_\lambda^M(x_1) - \nabla f_\lambda^M(x_2)\| \leq \frac{1}{\lambda}\sqrt{\frac{\mu_{\max}(M)}{\mu_{\min}(M)}}\|x_1 - x_2\| \quad \forall x_1 \in \mathcal{H}, x_2 \in \mathcal{H},$$

where $\mu_{\min}(M)$ and $\mu_{\max}(M)$ are respectively the smallest and the largest eigenvalue of M.

We set $M = I_n - \lambda A^T A$. If $\lambda \in [0, \frac{1}{\|A\|^2}[$, then M is symmetric positive definite. In this case, we have

$$\operatorname{prox}_{\lambda f}^M(x) = \operatorname{prox}_{\lambda g}(x - \lambda A^T(Ax - b)). \tag{5.38}$$

The formula (5.38) can be consulted in [66, section 4.6, p. 190]. Using (5.37) and (5.38), we get

$$\nabla f_\lambda^M(x) = \frac{1}{\lambda} M(x - \operatorname{prox}_{\lambda g}(x - \lambda A^T(Ax - b))).$$

Since $\operatorname{argmin}_{\mathcal{H}} f_\lambda^M = \operatorname{argmin}_{\mathcal{H}} f$, we can replace f with f_λ^M to recover the smooth case, and obtain

(IPAHDD-C-lasso)

Initialize : $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$.

$z_k = \frac{1}{\lambda} M(x_k - \operatorname{prox}_{\lambda g}(x_k - \lambda A^T(Ax_k - b)))$.

$y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta z_{k-1}$.

$x_{k+1} = x_k - \beta h z_k + h \operatorname{prox}_{\frac{h}{1+\gamma h}\varphi}\left(\frac{1}{1+\gamma h} y_k + \frac{(\gamma\beta - 1)h}{1+\gamma h} z_k\right)$.

Theorem 5.12 *Let A be an $m \times n$ matrix, $b \in \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a closed, convex proper function. Take $f = \frac{1}{2}\|A \cdot -b\|^2 + g$ and suppose that $\operatorname{argmin}_{\mathbb{R}^n} f \neq \emptyset$. Assume that $\varphi : \mathbb{R}^n \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Set $M = I_n - \lambda A^T A$ with $\lambda \in [0, \frac{1}{\|A\|^2}[$, and suppose that the positive parameters $h, \gamma, \beta, \lambda$ satisfy the relation*

$$\frac{\gamma}{h} - \frac{1}{2\lambda}\sqrt{\frac{\mu_{max}(M)}{\mu_{min}(M)}}(\gamma\beta + 1) \geq 0.$$

Then, for any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C-lasso)*, we have that $(x_k)_k$ converges, and its limit is a minimizer of f. Moreover*

(i) $\frac{1}{h}(x_{k+1} - x_k) + \beta z_k = 0$ *after a finite number of steps;*

(ii) $\sum_{k=1}^{\infty} \|z_k\|^2 < +\infty$, *where* $z_k = \frac{1}{\lambda} M(x - \operatorname{prox}_{\lambda g}(x - \lambda A^T(Ax - b)))$.

Proof. Replacing the Lipschitz constant L in Theorem 5.1 by $\frac{1}{\lambda}\sqrt{\mu_{\max}(M)/\mu_{\min}(M)}$, and recalling that $z_k = \nabla f_\lambda^M(x_k)$, then the result follows immediately. ■

5.8 Some numerical experiments

We use the performance profiles developed by Dolan and Moré as a tool for comparing different solvers. For each $t \in \mathbb{R}$, the performance profiles give the proportion $\rho_s(t)$ of test problems on which each solver s under comparison has a performance within the factor t of the best possible ratio. We choose the number of iterations found by each solver as a performance measure. We give a brief description of the performance profiles as follows (for more details, we refer to [72])

Let S be the set of the solvers that will be compared and n_s the number of solvers. The *performance ratio* is defined by

$$r_{p,s} = \log_2\left(\frac{t_{p,s}}{\min\{t_{p,s} : s \in S\}}\right),$$

where $p \in P$, $s \in S$, and $t_{p,s}$ is the performance measure (the number of iterations in our case).
The performance of the solver $s \in S$ is defined by

$$\rho_s(t) = \frac{1}{n_p}\text{size}\,\{p \in P \,:\, r_{p,s} \leq t\},$$

where, n_p is the number of problems, and t is a real factor.

5.8.1 Comparing the three algorithms (IPAHDD-C1), (IPAHDD-C2) and (IPAHDD-C3)

We perform numerical tests to compare the algorithms defined in the previous sections, and which deal with general differentiable function f with Lipschitz continuous gradient. We take $\varphi : \mathbb{R}^n \to \mathbb{R}$ given by $x \mapsto \varphi(x) = r\|x\|$, $r = 0.1$. First consider the simple situation where the function $f : \mathbb{R}^n \to \mathbb{R}$ is quadratic

$$f(x) = \tfrac{1}{2}\|Ax - b\|^2, \quad A \in \mathbb{R}^{m\times n},\ (m \leq n),\ b \in \mathbb{R}^m \text{ are chosen randomly.}$$

The matrices A are generated randomly. We have chosen a set P of 40 different problems with 40 matrices $A \in \mathbb{R}^{m\times n}$. The numerical experiments are carried out on an ordinary computer. All the codes are written and executed in MATLAB R2019a. We use the same initial points and the same stopping criterion, i.e., either the number of iterations exceeds 10^5 or $\|\nabla f(x_k)\| \leq 10^{-6}$. Regarding the choices of parameters h, β and γ, depending on the considered algorithm the parameters are chosen such that the assumptions in Theorem 5.1, or in Theorem 5.8, or in Theorem 5.9 are satisfied. In particular, the parameter selections are as follows.
For (IPAHDD-C1), to respect the condition in Theorem 5.1 we chose $\gamma = 0.3$, $\beta = 1$, and

$$h = \frac{2\gamma}{L(\gamma\beta + 1)}.$$

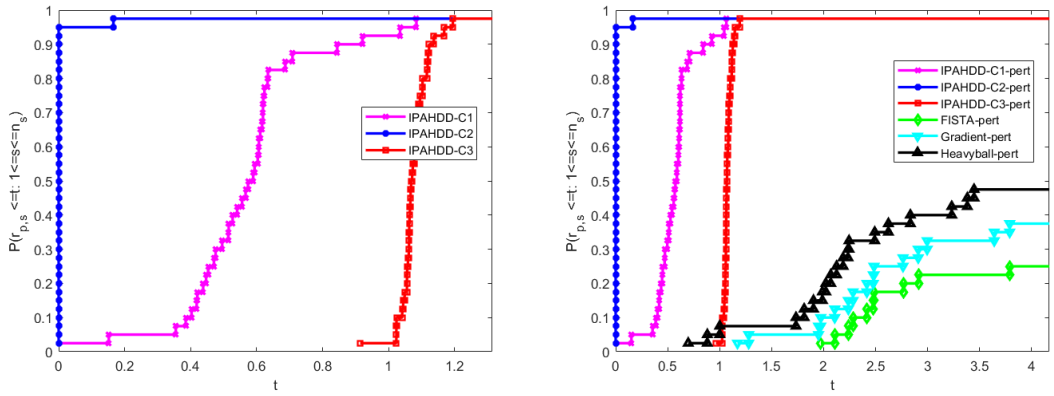
For (IPAHDD-C2), to respect the condition in Theorem 5.8 we chose $h = 0.5$, and

$$\gamma = 1.001\text{max}\,\{2Lh, L/2\},$$
$$\beta = 0.99\,\text{min}\left\{\frac{\gamma + \gamma^2 h - 2Lh}{Lh}, \frac{2 + (2\gamma - L)h}{\gamma^2 h + \gamma}\right\}.$$

For (IPAHDD-C3), to respect the condition in Theorem 5.9 we chose $h = 0.5$, and

$$\gamma = 1.001 \max\left\{\frac{L}{2h}, 2L, Lh\right\},$$
$$\beta = 0.99 \min\left\{\frac{2+2\gamma h - L}{\gamma(1+\gamma h)}, \frac{\gamma + h\gamma^2 - 2L}{L}\right\}.$$

For other variants of the algorithms, the parameters are chosen in the same spirit. Figure 5.1(a) reveals that (IPAHDD-C2) is the most efficient method out of the three in the sense that it requires the least number of iterations to reach a solution. Despite their good convergence properties, the algorithms which are based on the dry friction damping are not as fast as the FISTA method. This is easily understandable since our methods are proved to follow the steepest descent method regime after a finite number of steps. However, the situation is reversed if we introduce perturbations in the algorithms, as shown in the following experiments.



(a) **(b)**

Figure 5.1: Performance profiles of (IPAHDD-C1), (IPAHDD-C2) and (IPAHDD-C3) (left), (IPAHDD-C1-pert), (IPAHDD-C2-pert), (IPAHDD-C3-pert), (FISTA-pert), (Gradient-pert) and (Heavyball-pert) (right).

Following the suggestion of an anonymous reviewer, we conduct an experiment which compares a model algorithm, IPAHDD-C1, with the gradient descent (GD) and FISTA in the error-free case so that we can see the behavior of the algorithm at the early stage. Specifically, we test the three algorithms on the following toy least squares problem with the objective function $f(x_1, x_2) = \frac{1}{2}(x_1^2 + 1000x_2^2)$. As we can see from Figure 5.2, since our algorithm eventually becomes the gradient descent, it is reasonable that it performs slower than FISTA. At the early stage of (IPAHDD-C1) where the inertial effects are involved, we can see the oscillation effects which are commonly observed in inertial algorithms. After that, it loses its inertial effects and becomes the gradient descent. Indeed, we observe a

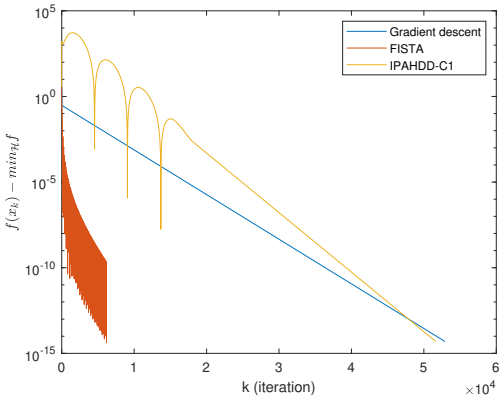


Figure 5.2: A concrete comparison between (IPAHDD-C1), FISTA and the gradient descent

linear convergence at the final stage of the algorithm, which is consistent with the strong convexity of the function considered. In addition, the convergence of (IPAHDD-C1) is comparable with the gradient descent.

5.8.2 Introducing errors

Recall that for the heavy ball method, introducing errors (e_k) does not affect the fast convergence property as long as $\sum \|e_k\| < +\infty$. For the FISTA algorithm, the condition is even more stringent, we need to assume that $\sum k\|e_k\| < +\infty$, see [35, Theorem 5.1] and [134]. A unified presentation of these results is given in [33, Theorem 2.1]. By contrast, in our situation, to preserve the convergence properties, we just need to assume that $\lim_k \|e_k\| = 0$. For the development of perturbations aspects of first order optimization methods, interested readers can consult [33, 35, 44, 46, 47, 50, 53, 134, 140, 147], and [38] in the case of the Hessian driven damping. We will now compare the perturbed versions of our algorithms, namely (IPAHDD-C1-pert), (IPAHDD-C2-pert) and (IPAHDD-C3-pert) (the two latter are respectively the perturbed version of (IPAHDD-C2) and (IPAHDD-C3) and defined in the same way as (IPAHDD-C1-pert)) with the perturbed gradient method, the perturbed Heavy Ball method and the perturbed FISTA method which are given below

(Gradient-pert)

Initialize : $x_0 \in \mathbb{R}^n$.

$$x_k = x_{k-1} - \gamma(\nabla f(x_{k-1}) + e_k).$$

(Heavyball-pert)

Initialize : $x_0 \in \mathbb{R}^n$.

$$x_{k+1} = x_k + \alpha(x_k - x_{k-1}) - \gamma(\nabla f(x_k) + e_k).$$

(FISTA-pert)
$$\boxed{\begin{array}{l}\text{Initialize : } y_0 = x_0 \in \mathbb{R}^n, (t_k)_{k\geq 1} : t_k = \frac{k+1}{2}.\\ x_k = y_{k-1} - \gamma(\nabla f(y_{k-1}) + e_k).\\ y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}).\end{array}}$$

The sequence $(t_k)_k$ in the above algorithm satisfies $t_1 = 1$ and $t_k^2 \geq t_{k+1}^2 - t_{k+1}$. Under this property, Beck and Teboulle [52] showed the $O(1/k^2)$ convergence rate for the above algorithm in the error-free case, i.e. when $e_k = 0$, $\forall k \geq 1$. Indeed, as explained above, under the summability property $\sum k\|e_k\| < +\infty$, the convergence rate is as in the error-free case (see [35] or [134]). For numerical purposes, we choose the sequence (e_k) such that $\|e_k\| = 1/k$; in fact, for each k we choose a random vector $\xi \in \mathbb{R}^n$ with the uniform distribution on $]0,1[^n$ and then set $e_k = (1/(k\|\xi\|))\xi$. In this way, the conditions $\sum k\|e_k\| < +\infty$ and $\sum \|e_k\| < +\infty$ are not satisfied, which allows us to check the advantage of our methods in presence of perturbations compared to (FISTA-pert), (Gradient-pert) and (Heavyball-pert). We use performance profiles on the quadratic problem, as we did before to carry out this comparison. As anticipated, we can see from Figure 5.1(b) that FISTA, the gradient method and the Heavy Ball method suffer substantially from the perturbations when the conditions $\sum k\|e_k\| < +\infty$ and $\sum \|e_k\| < +\infty$ are not satisfied, while the proposed algorithms prove their robustness and preserve their behavior as in the non-perturbed case. This naturally leads to considering stochastic versions of our algorithms.

Remark 5.8 Let us recall the dynamical system corresponding to the algorithm (IPAHDD-C1-pert) (also (IPAHDD-C2-pert) and (IPAHDD-C3-pert))

$$\ddot{x}(t) + \gamma\dot{x}(t) + \partial\varphi\Big(\dot{x}(t) + \beta\nabla f(x(t))\Big) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) \ni e(t). \tag{5.39}$$

In a mechanical context, the error $t \mapsto e(t)$ is interpreted as external forces or excitation applied to a mechanical system. We know that IPAHDD-C1-pert are obtained by temporal discretizations of this dynamical system. On the other hand, we would like to emphasize that (Gradient-pert), (Heavyball-pert) and (FISTA-pert) are actually the discretizations of the following dynamical system (see [35])

$$\ddot{x}(t) + \frac{\gamma}{t}\dot{x}(t) + \nabla f(x(t)) = e(t). \tag{5.40}$$

Here, it should be noticed that the equation (5.40) is a just special case of the inclusion (5.39) in which we set $\beta = 0, \varphi \equiv 0$, and the viscous damping coefficient is time dependent and of the form $\gamma(t) = \frac{\gamma}{t}$. This means that the inclusion (5.39) and the equation (5.40) have the same continuous structure. In the temporal discretizations of these dynamics, the term $e(t)$ becomes e_k, that could be interpreted as a perturbation coming from an external force. Therefore, comparing discretized versions of them is a totally fair comparison.

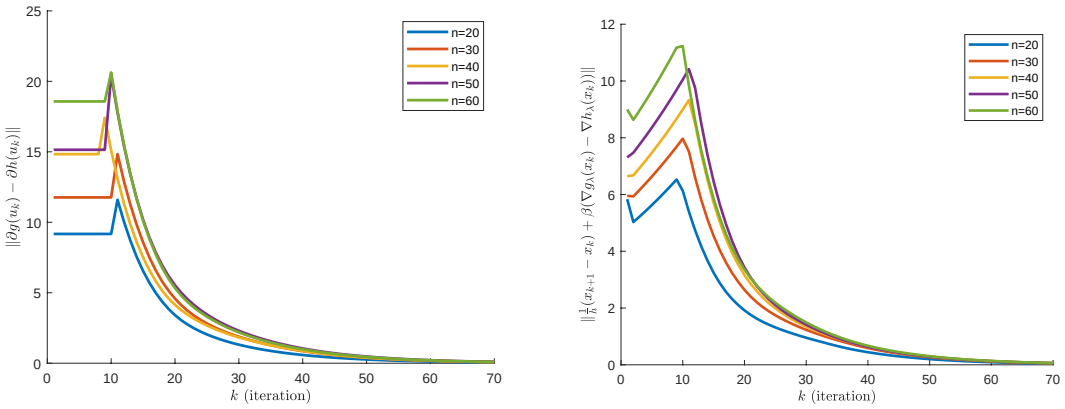
Another comment is that if we look at the equation 5.40, the perturbations in (Gradient-pert), (Heavyball-pert) and (FISTA-pert) should not be understood as born in the calculations of the gradients but should be understood as independent external errors. The reason why the formulas of these three algorithms seem like that they deal with gradient-associated errors is because they involve only one gradient term in their formulas. This is unlike our algorithms, where we have some additional "correcting terms" related to the gradients which stem from the dry friction and the Hessian driven damping.

5.8.3 Nonsmooth nonconvex d.c. problems

Let us illustrate the algorithm (IPAHDD-CDC) with nonsmooth nonconvex problems of DC type. Given $n \geq 2$, consider the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(x) = \|Ax - b\|_2^2 - \|A^T b\|_2 \|x\|_2, \tag{5.41}$$

where A is an orthogonal matrix of order n and $b \in \mathbb{R}^n$. We choose 5 random orthogonal matrices A of size ranging from 20 to 60 while b has all its coordinates equal to one. To apply the algorithm (IPAHDD-CDC), we rely on the "trivial" DC decomposition $f = g - h$ where $g : x \mapsto \|Ax - b\|_2^2$ and $h : x \mapsto \|A^T b\|_2 \|x\|_2$. Clearly, g and h are semialgebraic. The orthogonality of A is assumed only to facilitate the computations of prox_g. Therefore, according to Remark 5.7, we have that $g_\lambda - h_\lambda$ satisfies the (KL) property for $\lambda > 0$. As a result, under the assumptions of Theorem 5.11, the sequence (x_k) generated by the algorithm (IPAHDD-CDC) converges to some x_∞, and $\text{prox}_{\lambda h}(x_\infty)$ is a critical point of f in the sense of Toland. It is easy to show that u is a critical point of f in the sense of Toland if and only if $u \neq 0$ and $2A^T(Au - b) - \frac{\|A^T b\|_2 u}{\|u\|_2} = 0$. The stopping condition we use for (IPAHDD-CDC) is either the number of iterations exceeding 10^5 or $u_k \neq 0$ and $\left\| 2A^T(Au_k - b) - \frac{\|A^T b\|_2 u_k}{\|u_k\|_2} \right\|_2 \leq 10^{-6}$. Figure (5.3) depicts the behavior of the quantities $\|\partial g(u_k) - \partial h(u_k)\|$ and $\left\| \frac{1}{h}(x_{k+1} - x_k) + \beta(\nabla g_\lambda(x_k) - \nabla h_\lambda(x_k)) \right\|$ over iterations, where $u_k = \text{prox}_{\lambda h}(x_k)$, in five problems of different sizes. (IPAHDD-CDC) deals with the five problems successfully. In Figure 5.3(b), we observe that after a certain number of iterations, the norm of the sum of the discrete velocity vector and gradient terms is decreasing and approaching zero. This is in accordance with Theorem 5.11, which establishes that after a finite number of iterations, the algorithm follows the steepest descent regime. We now consider the algorithm (IPAHDD-CDC-pert) which



(a) Behavior of $\|\partial g(u_k) - \partial h(u_k)\|$

(b) Behavior of $\left\|\frac{1}{h}(x_{k+1} - x_k) + \beta(\nabla g_\lambda(x_k) - \nabla h_\lambda(x_k))\right\|$

Figure 5.3: Algorithm (IPAHDD-CDC) with $f : x \mapsto \|Ax - b\|_2^2 - \|A^T b\|_2 \|x\|_2$ and different initial data.

is a perturbed version of (IPAHDD-CDC).

(IPAHDD-CDC-pert)

Initialize : $x_0 \in \mathbb{R}^n$, $x_1 \in \mathbb{R}^n$.

$$y_k = \tfrac{1}{h}(x_k - x_{k-1}) - \tfrac{\beta}{\lambda}(\mathrm{prox}_{\lambda g}(x_{k-1}) - \mathrm{prox}_{\lambda h}(x_{k-1}))$$

$$x_{k+1} = x_k + \tfrac{\beta h}{\lambda}(\mathrm{prox}_{\lambda g}(x_k) - \mathrm{prox}_{\lambda h}(x_k)) + h\,\mathrm{prox}_{\frac{h}{1+\gamma h}\varphi}\left(\tfrac{1}{1+\gamma h}y_k - \tfrac{(\gamma\beta-1)h}{(1+\gamma h)\lambda}(\mathrm{prox}_{\lambda g}(x_k) - \mathrm{prox}_{\lambda h}(x_k)) + \tfrac{h}{1+\gamma h}e_k\right)$$

It is easy to check that under the assumptions of Theorem 5.11 together with $\lim_k \|e_k\| = 0$, the conclusions of Theorem 5.11 also hold true for the algorithm (IPAHDD-CDC-pert). It is well-known that the classical DC algorithm (DCA), introduced by Pham Dinh Tao et al [121] is one of the algorithms that solve effectively nonsmooth and nonconvex optimization problems of the form

$$\inf_{x\in\mathbb{R}^n} \{f(x) := g(x) - h(x)\},$$

where g and h are lower semicontinuous proper real extended valued convex functions. Briefly, the algorithm consists in constructing two sequences (x_k) and (y_k) such that the sequences of values of the primal and dual objective functions $\{g(x_k) - h(x_k)\}, \{g^*(x_k) - h^*(x_k)\}$ are decreasing, and their corresponding limits x_∞ and y_∞ satisfy local optimality con-

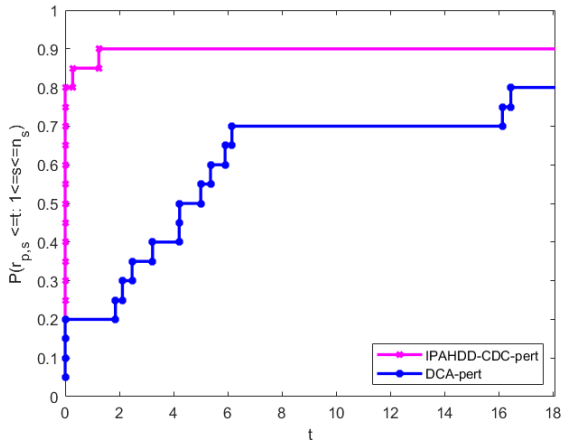


Figure 5.4: Performance profiles of (IPAHDD-CDC-pert) and (DCA-pert) on the problem (5.41)

ditions [97]. Precisely, the standard (DCA) reads as follows. Choose an initial point $x_0 \in \text{dom}(g) = \{x \in \mathbb{R}^n \ : \ g(x) < +\infty\}$, and for $k = 0, 1, \ldots$, set

$$y_k \in \partial h(x_k); \quad x_{k+1} \in \partial g^*(y_k) = \text{argmin}_{x \in \mathbb{R}^n} \left\{ g(x) - \langle y^k, x \rangle \right\}.$$

For the purpose of comparison with (IPAHDD-CDC-pert), we propose the following perturbed version of DCA

(DCA-pert)
$$\boxed{\begin{array}{l} y_k \in \partial h(x_k) \\ x_{k+1} \in \partial g^*(y_k) + e_k = \text{argmin}_{x \in \mathbb{R}^n} \left\{ g(x) - \langle y^k, x \rangle \right\} + e_k \end{array}}$$

Using performance profiles with the number of iterations as a performance measure, we make a comparison between (IPAHDD-CDC-pert) and (DCA-pert) on the d.c. problem (5.41). The perturbation sequence here is chosen in the same way as before, i.e., for each k we choose a random vector $\xi \in \mathbb{R}^n$ with the uniform distribution on $]0, 1[^n$ and then set $e_k = \frac{\xi}{k\|\xi\|}$. The performance profiles in Fig. 5.4 show that in the presence of perturbations, (IPAHDD-CDC-pert) outperforms (DCA-pert). Specifically, (IPAHDD-CDC-pert) wins over (DCA-pert) on 80% of the problems used for this experiment; moreover, the number of problems that can be solved by (IPAHDD-CDC-pert) is higher (compared to (DCA-pert)).

5.9 Concluding remarks

In this chapter, we presented a new way of handling dry friction in first order inertial algorithms. While in previous works, dry friction comes as a nonlinear action on the velocity, we now consider its action on a weighted sum of the velocity vector and the

gradient of the function f to be minimized. The sequences thus generated converge towards critical points of f (global minima when f is convex), whereas previously we only end up with approximate critical points of f. In addition, after a finite number of steps, the algorithm changes its nature and passes from an inertial algorithm to a steepest descent method. This combined with the Hessian-driven damping makes it possible to considerably reduce the oscillations: one benefits from the inertial effect at the beginning, then one passes to a method of gradient. In many ways, this closed loop control of the algorithm/dynamic has similarities to restart methods. Most importantly, the algorithm enjoys remarkable structural stability and robustness properties. It is a well known fact that there is a trade-off between fast convergence of optimization methods and their robustness to perturbations. Thus the algorithm is an interesting balance between fast convergence and robustness. This makes the algorithm a promising tool for dealing with stochastic/noisy situations in nonconvex, nonsmooth optimization. Its combination with approximation techniques is also promising. In addition, the technique that is developed is quite flexible. By relying on the threshold effect attached to dry damping, one can imagine controlling the dynamics, and thus switching to different regimes, forcing finite time synchronization of nonlinear oscillators, and many others. Several questions require additional investigations, concerning for example general composite optimization problems, as well as the study of the associated stochastic algorithms.

5.10 Appendix

5.10.1 Another proof of the iterate's weak convergence

We start by recalling the following well-known Opial's lemma that will be used in the proof.

Lemma 5.2 (Opial's lemma) *Let S be a nonempty set of a Hilbert space $\mathcal{H}$. Suppose that $(x_k)_k$ is a sequence in $\mathcal{H}$ which satisfies*

- $\lim_{k\to\infty} \|x_k - p\|$ *exists for all $p \in S$.*
- *For each subsequence $(x_{k_l})_l$ of $(x_k)_k$ that converges weakly to x, we have $x \in S$.*

Then, there exists $x \in S$ such that $(x_k)_k$ converges weakly to x.

In the item (i) of Theorem 5.3, let us give a direct proof that $(x_k)_k$ converges weakly to a minimizer of f, without using the fact that after a finite number of steps, the iterates follow the steepest descent method. The proof is based on Opial's lemma. According to the convexity of f, and hence the monotonicity of ∇f, we have for all $k \geq 1$ and for all $z \in \mathcal{H}$

$$\begin{aligned}\beta\langle \nabla f(x_{k-1}), x_{k-1} - z\rangle &= \beta\langle \nabla f(x_{k-1}) - \nabla f(z), x_{k-1} - z\rangle + \beta\langle \nabla f(z), x_{k-1} - z\rangle \\ &\geq \beta\langle \nabla f(z), x_{k-1} - z\rangle.\end{aligned}$$

Therefore,

$$\begin{aligned}\langle y_k, x_{k-1}-z\rangle &= \langle \frac{1}{h}(x_k-x_{k-1}), x_{k-1}-z\rangle + \beta\langle \nabla f(x_{k-1}), x_{k-1}-z\rangle \\ &\geq \frac{1}{2h}(\|x_k-z\|^2 - \|x_{k-1}-z\|^2 - \|x_k-x_{k-1}\|^2) + \beta\langle \nabla f(z), x_{k-1}-z\rangle,\end{aligned}$$

where $y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})$.
This, together with the Cauchy Schwarz inequality, implies

$$\frac{1}{2h}(\|x_k-z\|^2 - \|x_{k-1}-z\|^2) \leq (\|y_k\| + \beta\|\nabla f(z)\|)\|x_{k-1}-z\| + \frac{1}{2h}\|x_k-x_{k-1}\|^2.$$

To check the first item of the Opial's lemma, let us now assume that $z \in \operatorname{argmin}_{\mathcal{H}} f$ which is fixed. As a result, it follows

$$\frac{1}{2h}(\|x_k-z\|^2 - \|x_{k-1}-z\|^2) \leq \|y_k\|\|x_{k-1}-z\| + \frac{1}{2h}\|x_k-x_{k-1}\|^2. \tag{5.42}$$

By summing the above inequalities from $k=1$ to $N \geq 1$, we obtain

$$\frac{1}{2h}(\|x_N-z\|^2 - \|x_0-z\|^2) \leq \sum_{k=1}^{N}\|y_k\|\|x_{k-1}-z\| + \frac{1}{2h}\sum_{k=1}^{N}\|x_k-x_{k-1}\|^2. \tag{5.43}$$

Recall that we have already obtained $\sum_{k=1}^{\infty}\|x_k-x_{k-1}\|^2 < +\infty$ and $\sum_{k=1}^{\infty}\|y_k\| < +\infty$. Set $P = \sum_{k=1}^{\infty}\|y_k\| \geq 0$, $Q = \sum_{k=1}^{\infty}\|x_k-x_{k-1}\|^2 \geq 0$ and $m_n = \max_{0\leq i\leq n}\|x_i - z\|$. For all $n \geq 1$ and $1 \leq i \leq n$, we deduce from (5.43) that

$$\frac{1}{2h}(\|x_i-z\|^2 - \|x_0-z\|^2) \leq P.m_{i-1} + \frac{1}{2h}Q \leq P.m_n + \frac{1}{2h}Q.$$

It follows that for all $n \geq 1$, we have

$$m_n^2 - \|x_0-z\|^2 \leq 2hPm_n + Q,$$

or

$$m_n^2 - 2hPm_n - \|x_0-z\|^2 - Q \leq 0.$$

The above inequality implies that

$$m_n \leq hP + \sqrt{h^2P^2 + \|x_0-z\|^2 + Q} \quad \forall n \geq 1,$$

which means that $(m_n)_n$ is bounded, and hence the sequence $(\|x_k-z\|)_k$ is bounded.

Combining this boundedness property with (5.43), we can easily show that $(\|x_k - z\|)_k$ is a Cauchy sequence in $\mathbb{R}$, and hence converges. We have shown that $(x_k)_k$ fulfills the first item of the Opial's lemma.
Now, we turn to proving that $(x_k)_k$ also satisfies the second item of the Opial's lemma. To this end, take any subsequence $(x_{k_l})_l$ of $(x_k)_k$ and assume that $(x_{k_l})_l$ converges weakly to some $x \in \mathcal{H}$. Since f is convex, we have for all $z \in \mathcal{H}$

$$f(z) \geq f(x_{k_l}) + \langle \nabla f(x_{k_l}), z - x_{k_l} \rangle.$$

Let us pass to the lim inf as $l \to +\infty$ in the above inequality. Since $(\nabla f(x_{k_l}))_l$ converges strongly to 0 and $(x_{k_l})_l$ is bounded, we obtain

$$f(z) \geq \liminf_{l \to \infty} f(x_{k_l}).$$

Moreover, f is weakly lower semicontinuous, so the above inequality gives

$$f(z) \geq f(x).$$

Since z can be taken arbitrarily in $\mathcal{H}$, we deduce that $x \in \operatorname{argmin}_{\mathcal{H}} f$.
With all things considered, we apply the Opial's lemma to deduce that there exists $x_\infty \in \operatorname{argmin}_{\mathcal{H}} f$ such that $(x_k)_k$ converges weakly to x_∞ in $\mathcal{H}$.

6

A doubly nonlinear evolution system with threshold effects associated with dry friction

Contents

This chapter covers the material discussed in the submitted paper [9], which was produced in collaboration with S. Adly and H. Attouch

6.1 Introduction

In recent years, the interplay between continuous optimization and the theory of dynamical systems has resulted in significant advancements in the field of applied mathematics. The investigation of the long-term behavior of inertial dynamics, particularly within the context of a Hilbert space for convex differentiable optimization, has become a focal point. In this chapter, we delve into a new layer of complexity by considering threshold effects associated with dry friction in the framework of inertial dynamics. We lay our foundation on a doubly nonlinear first-order evolution equation that involves two potentials. The differentiable function f to be minimized interacts with the system's state via its gradient and the nonsmooth dry friction potential $\varphi = r\|\cdot\|$, $r > 0$, that operates on a linear combination of the velocity vector and the gradient of f through its convex subdifferential. These two potential components interplay to shape the dynamics of the system. In order to shed light on the centrality of $\nabla f(x)$, we adopt a dual formulation approach, featuring a Riemannian gradient structure, thus providing a deeper insight into the dynamics of the system. Building on the general acceleration method proposed by Attouch, Bot, and Nguyen [28], and recently extended by Adly and Attouch [6] to dry friction, our methodology incorporates time scaling and averaging of a first-order continuous differential equation. These techniques pave the way for obtaining fast convergence results for second-order time-evolution systems that include dry friction, asymptotic vanishing damping, and Hessian-driven damping in an implicit form. In this chapter, we develop these concepts, provide mathematical proofs in support of our results and illustrate this through numerical simulations. We believe that these new results can contribute to the understanding and development of accelerated gradient methods from the continuous time

perspective, potentially providing valuable insight into intricate optimization problems. Let us just briefly recall some facts about previous related works and at the same time highlight differences between those and our work in this chapter. First, acting as the basis upon which our current chapter is built is the work by Attouch, Bot, and Nguyen [28] on the acceleration of first order dynamics via the time scaling and averaging techniques. The authors in that paper provide a generic approach by which second order dynamics with improved convergence properties can be deduced from first order ones; what is notable is while one needs to develop a Lyapunov analysis for the convergence of the first order dynamic, the improved convergence properties of the resulting second order dynamic obtained by the time scaling and averaging techniques can be yielded solely by the differential and integral calculus. Making use of this acceleration approach, we, in this chapter, manage to speed up the convergence of a doubly nonlinear evolution equation that involves the presence of dry friction. However, the addition we make, which is not performed in the original paper [28], is that we further propose a dual approach to the initial evolution system by introducing a dual dynamic with the function variable being the gradient of the function to be minimized f. The study of the dual dynamic makes it possible to have a better understanding of the properties of the gradient of f. This dual dynamic, which has a Riemannian gradient structure, further yields a second order dynamic with accelerated convergence rates via the time scaling and averaging techniques. Considering the dual approach is, in fact, initiated by Adly and Attouch in [6] where they also study a doubly nonlinear evolution system involving dry friction which turns out to be a special case of our first order dynamic. What distinguishes our work with [6] is largely in the first order dynamic itself. Equipped with a slightly different dry friction term, our first order dynamic improves that of [6] in the sense that the limit point of the solution trajectory is now the exact critical point, not just an approximate one. This difference in the dynamic will be precisely indicated shortly.

Throughout this chapter, $\mathcal{H}$ is a real Hilbert space equipped with the scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $\| \cdot \|$. We first look at the first-order evolution equation.

$$\text{(DRYAD)} \quad \gamma(\dot{x}(t) + \beta \nabla f(x(t))) + \partial \varphi\Big(\dot{x}(t) + \beta \nabla f(x(t))\Big) + \nabla f(x(t)) \ni 0, \quad t \in [t_0, \infty)$$

that is a doubly nonlinear dynamic that involves two potentials.

We make the following standing assumptions on the two potentials f and φ.

$$\begin{cases} f : \mathcal{H} \to \mathbb{R} \text{ is a continuously differentiable function which is bounded from below.} \\ \nabla f \text{ is Lipschitz continuous on the bounded sets of } \mathcal{H}. \\ \varphi : \mathcal{H} \to \mathbb{R} \text{ satisfies } \varphi(x) = r\|x\| \text{ for some } r > 0 \text{ and } \gamma > 0,\ \beta \geq 0. \end{cases}$$

This doubly nonlinear differential inclusion contains the term $\partial\varphi\big(\dot{x}(t) + \beta\nabla f(x(t))\big)$ attached to dry friction, hence the abbreviation (DRYAD) for Dry friction Acting Doubly. The case $\beta = 0$ and $\gamma = 1$, was studied in [6]. It's worth noticing that the basic starting dynamic for the majority of gradient methods in optimization is the steepest descent method. The first potential, designated as f, affects the system's state via its gradient and is a differentiable function to be minimized. The velocity vector is affected by the second potential $\varphi = r\|\cdot\|$. The study of the associated dynamics' asymptotic behavior is significantly altered by the presence of this nonsmooth dry friction potential.
One distinctive characteristic of (DRYAD) is the inclusion of the dry friction term $\partial\varphi\Big(\dot{x}(t)+\beta\nabla f(x(t))\Big)$, which incorporates both the velocity vector and the gradient of f. This differentiation sets it apart from previously studied dynamics, where the dry friction term exclusively involves the velocity vector. Although seemingly straightforward, this modification significantly alters the dynamics in comparison to those investigated in [3,5–7]. An advantageous aspect of representing the dry friction term in this new form is that each trajectory generated by (DRYAD) converges towards a critical point of f, specifically a minimizer in the case of convex f. In fact, any stationary point x_∞ of the dynamic (DRYAD) satisfies $\partial\varphi(\beta\nabla f(x_\infty)) + (1+\gamma\beta)\nabla f(x_\infty) \ni 0$. This condition is equivalent to $\beta\nabla f(x_\infty) = \mathrm{prox}_{\frac{\beta}{1+\gamma\beta}\varphi}(0)$, which, in combination with the dry friction property $(\mathrm{DF})_r$, implies that $\nabla f(x_\infty) = 0$ if $\beta > 0$ (see Lemma 5.1). Thus, x_∞ corresponds to a critical point of f. In contrast, in the case $\beta = 0$, each trajectory generated by the dynamic converges towards an "approximate" critical point x_∞ of f, characterized by $-\nabla f(x_\infty) \in \partial\varphi(0)$.
To emphasize the role played by the gradient, we also examine the dual approach that involves the dual variable $g(x) = \nabla f(x)$, and the corresponding evolution reads

$$\nabla^2 f^*(g(t))\dot{g}(t) + \frac{\gamma\beta+1}{\gamma}g(t) - \frac{1}{\gamma}\mathrm{proj}_{B(0,r)}(g(t)) = 0,$$

thus making appear the Riemannian structure associated with the Hessian of the convex Fenchel conjugate function f^* (when this function is assumed of class C^2) associated with f. Here, $\mathrm{proj}_{B(0,r)}$ denotes the projection operator onto the closed ball $B(0,r)$. Our first investigation focuses on the convergence properties of the trajectories produced by the primal evolution system (DRYAD) and its dual.

Next, we leverage the universal acceleration approach developed by Attouch, Bot, and Nguyen [28], wherein they employ a time scaling technique on a first-order continuous differential equation and subsequently apply the method of averaging. These techniques

give a second-order evolution system when applied to (DRYAD)

$$\ddot{z}(s)+\frac{\alpha}{s}\dot{z}(s)+\frac{\gamma\beta+1}{\gamma}\nabla f\Big(z(s)+\frac{s}{\alpha-1}\dot{z}(s)\Big) \\ +\frac{1}{\gamma}\nabla\varphi_{\frac{s}{\gamma(\alpha-1)}}\Big(-\frac{s}{\gamma(\alpha-1)}\nabla f(z(s)+\frac{s}{\alpha-1}\dot{z}(s))\Big)=0,$$

that involves dry friction aspects (smoothly via the gradient of the Moreau envelope $\nabla\varphi_{\frac{s}{\gamma(\alpha-1)}}$ of φ), asymptotically vanishing viscous damping (which is closely related to Nesterov's accelerated gradient method), and a damping term that is driven by the Hessian of f in an implicit form. Doing the same for the dual dynamic, we obtain

$$\nabla^2 f^*\Big(w(s)+\frac{s}{\alpha-1}\dot{w}(s)\Big)\Big(\ddot{w}(s)+\frac{\alpha}{s}\dot{w}(s)\Big)+\nabla\Psi_\beta^*\Big(w(s)+\frac{s}{\alpha-1}\dot{w}(s)\Big)=0.$$

In the case of these inertial systems, there is no necessity to conduct a Lyapunov analysis due to the utilization of the scaling and averaging method. Instead, we exploit the convergence results of the first-order system (DRYAD) by employing techniques from differential and integral calculus. Consequently, we achieve fast convergence results for second-order time-evolution systems that incorporate dry friction, asymptotically vanishing viscous damping, and Hessian-driven damping in the implicit form.

6.1.1 Some historical facts

Let's discuss the function and significance of each damping term involved in our inertial dynamics.

Viscous friction

The term $\gamma\dot{x}(t)$ in (DRYAD) models the viscous damping with a positive coefficient $\gamma>0$. This is linked to the heavy ball with friction method of Polyak [125,126]. Precisely, in [125] Polyak introduced the Heavy Ball with Friction method, which is based on the following inertial system with a fixed viscous damping coefficient

$$\text{(HBF)}\qquad \ddot{x}(t)+\gamma\dot{x}(t)+\nabla f(x(t))=0.$$

The Heavy-Ball Method (HBF) ensures exponential convergence of $f(x(t))$ to $\min_{\mathcal{H}} f$ for a smooth strongly convex function f. The convergence rate of (HBF) for general convex functions is $\mathcal{O}(1/t)$, which isn't faster than the steepest descent approach. Su-Boyd-Candès' approach of introducing a vanishing viscous damping coefficient in [143], denoted by $\gamma(t)=\alpha/t$, where α is a positive parameter, made a substantial addition

to the field. The corresponding ordinary differential equation (ODE) known as the Su-Boyd-Candès dynamic represents a continuous surrogate of the Nesterov accelerated gradient (NAG) method and is given by

$$(\mathrm{AVD})_\alpha \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0.$$

We have the inversely quadratic convergence rate of the values $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}(1/t^2)$ for any trajectory $x(t)$ of $(\mathrm{AVD})_\alpha$ with $\alpha \geq 3$. The viscous damping coefficient $\frac{\alpha}{t}$ vanishes (tends to zero) as time t approaches infinity, hence the terminology Asymptotic Vanishing Damping. The convergence properties of the dynamic $(\mathrm{AVD})_\alpha$ have been the subject of many recent studies, see [20, 21, 30, 32, 35, 36, 41, 43, 44, 109, 143]. The case where the parameter $\alpha = 3$ is crucial since it matches Nesterov's historical algorithm. With the exception of the one dimensional case, where convergence of the trajectories has been demonstrated [36], the question of whether the trajectories converge in this case is still unanswered. According to Attouch-Chbani-Peypouquet-Redont [35], each trajectory weakly converges to a minimizer of f for values $\alpha > 3$. The corresponding algorithmic result was obtained by Chambolle-Dossal [65]. Furthermore, it has been proved in [41] and [109] that for $\alpha > 3$, the asymptotic convergence rate of the values is actually $o(1/t^2)$. Apidopoulos-Aujol-Dossal [24] and Attouch-Chbani-Riahi [36] investigated the subcritical situation where $\alpha < 3$ and showed that the convergence rate of the objective values is $\mathcal{O}(t^{-\frac{2\alpha}{3}})$. These rates are optimal, which means they can be reached or approached arbitrarily closely.

Dry friction

Following [3–5], we say that the potential function φ satisfies the dry friction property $(\mathrm{DF})_r$, $r > 0$, if the following properties are satisfied:

$$(\mathrm{DF})_r \quad \begin{cases} \varphi : \mathcal{H} \to \mathbb{R}_+ \text{ is convex continuous,} \\ \min_{\xi \in \mathcal{H}} \varphi(\xi) = \varphi(0) = 0, \\ \varphi(\xi) \geq r\|\xi\| \ \ \forall \xi \in \mathcal{H}. \end{cases}$$

The function $\varphi(x) = r\|x\|$, $r > 0$ is a model example of potential which satisfies the dry friction property, which will be used throughout this chapter. An important property associated with dry friction is stated in the lemma below (see [3–5] for further details).

Lemma 6.1 *Suppose that $\varphi : \mathcal{H} \to \mathbb{R}_+$ satisfies $(\mathrm{DF})_r$. Then one has $\overline{\mathbb{B}}(0, r) \subset \partial\varphi(0)$, and therefore*

$$\|x\| \leq \lambda r \Longrightarrow \mathrm{prox}_{\lambda\varphi}(x) = 0.$$

In the above formula, $\text{prox}_{\varphi} : \mathcal{H} \to \mathcal{H}$ denotes the proximal mapping associated with the convex function φ. Recall that, for any $x \in \mathcal{H}$, for any $\lambda > 0$

$$\text{prox}_{\lambda\varphi}(x) = \text{argmin}_{\xi\in\mathcal{H}} \left\{\lambda\varphi(\xi) + \tfrac{1}{2}\|x - \xi\|^2\right\}.$$

For a thorough background on convex analysis in Hilbert spaces, we refer to [51]. Lemma 6.1 establishes a thresholding property for the proximal operator associated with a dry friction potential.

Dry friction holds significant importance in the realm of mechanics as it induces stabilization of mechanical systems within finite time. This stands in contrast to viscous damping, which tends to produce numerous small oscillations asymptotically. Consequently, dry friction serves as an appealing tool for optimization purposes. Although the use of dry friction in optimization is a relatively recent topic, initial findings regarding the property of finite convergence under the influence of dry friction were obtained by Adly, Attouch, and Cabot [7]. Corresponding results for Partial Differential Equations have been established in [22, 68, 71, 132].

Hessian-driven damping

The combination of viscous friction with dry friction and Hessian driven damping has been considered by Adly and Attouch in [3–5]. The Hessian driven damping has a natural connection with the strong damping property in mechanics and physics, see [82]. It helps to control and attenuate the oscillation effects that occur naturally with inertial systems. Recent research has concentrated on the inertial dynamic

$$\text{(DIN)}_{\alpha,\beta} \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0,$$

which combines asymptotic vanishing damping with Hessian-driven damping. The corresponding algorithms involve a correcting term in the Nesterov accelerated gradient method which reduces the oscillatory aspects, see Attouch-Peypouquet-Redont [42], Attouch-Chbani-Fadili-Riahi [34], Shi-Du-Jordan-Su [138]. Related to this is the Inertial System with Implicit Hessian Damping

$$\text{(ISIHD)} \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f\big(x(t) + \beta(t)\dot{x}(t)\big) = 0,$$

considered by Alecsa-László-Pinta in [19], see also Attouch-Fadili-Kungurtsev [38] in the perturbed case. The justification for using the term “implicit" stems from the observation that through Taylor expansion (as $t \to \infty$ we obtain $\dot{x}(t) \to 0$) one has

$$\nabla f\big(x(t) + \beta(t)\dot{x}(t)\big) \approx \nabla f(x(t)) + \beta(t)\nabla^2 f(x(t))\dot{x}(t),$$

hence making the Hessian damping appear indirectly.

6.1.2 Contents

The structure of this chapter is as follows. In Section 6.2, we study the first order system (DRYAD), where we show the wellposedness of the system, energy estimates and some convergence properties. A dual approach to (DRYAD) is studied in Section 6.3. Then, in Section 6.4, we use the time scaling and averaging techniques for (DRYAD) to obtain an inertial dynamic with accelerated convergence results. We also employ these techniques for the dual dynamic of (DRYAD) in Section 6.5. We illustrate our theoretical results with some numerical examples in Section 6.6.

6.2 Study of the first order system (DRYAD)

6.2.1 Wellposedness, and energy estimates: f not necessarily convex

Recall that our approach is based on the dynamical system (DRYAD)

$$\text{(DRYAD)} \quad \gamma(\dot{x}(t)+\beta\nabla f(x(t)))+\partial\varphi\Big(\dot{x}(t)+\beta\nabla f(x(t))\Big)+\nabla f(x(t))\ni 0,$$

which is a doubly nonlinear evolution equation.
Let us first observe that the Cauchy problem associated with (DRYAD) is well posed. In fact, we can rewrite (DRYAD) as follows

$$(I+\frac{1}{\gamma}\partial\varphi)(\dot{x}(t)+\beta\nabla f(x(t)))\ni-\frac{1}{\gamma}\nabla f(x(t)).$$

This is equivalent to

$$\dot{x}(t)+\beta\nabla f(x(t))=\text{prox}_{\frac{1}{\gamma}\varphi}(-\frac{1}{\gamma}\nabla f(x(t))), \tag{6.1}$$

where $\text{prox}_{\frac{1}{\gamma}\varphi}$ denotes the proximal operator which is single-valued since $\varphi:\mathcal{H}\to\mathbb{R}$ is a convex function (hence $\partial\varphi$ is a maximally monotone operator). Set $T:\mathcal{H}\to\mathcal{H}$ defined by $T(y)=-\beta y+\text{prox}_{\frac{1}{\gamma}\varphi}(-\frac{1}{\gamma}y)$, Equation (6.1) can be cast under the form

$$\dot{x}(t)=T(\nabla f(x(t)))=F(x(t)),$$

where $F = T \circ \nabla f$. Since T is globally Lipschitz continuous and ∇f is Lipschitz continuous on the bounded sets, F is Lipschitz continuous on the bounded sets. This property guarantees the existence and uniqueness of a local solution to (DRYAD) according to the classical Cauchy-Lipschitz theorem.
To pass from a local to a global solution, we need the following energy estimates

$$\gamma \int_{t_0}^{t} \|y(s)\|^2 ds + r \int_{t_0}^{t} \|y(s)\| ds + \beta \int_{t_0}^{t} \|\nabla f(x(s))\|^2 ds + f(x(t)) \leq f(x(t_0)),$$

where $y(t) = \dot{x}(t) + \beta \nabla f(x(t))$.
This, combined with f being bounded below, classically implies the global existence property. In order to achieve this energy estimate, we proceed as follows. Taking the scalar product of (DRYAD) with $y(t)$, we obtain

$$\gamma \|y(t)\|^2 + \langle \partial \varphi(y(t)), y(t) \rangle + \langle \nabla f(x(t)), \dot{x}(t) + \beta \nabla f(x(t)) \rangle = 0$$

As a property of dry friction, we have

$$\langle \partial \varphi(y(t)), y(t) \rangle \geq r \|y(t)\|.$$

Therefore,

$$\gamma \|y(t)\|^2 + r\|y(t)\| + \beta \|\nabla f(x(t))\|^2 + \frac{d}{dt} f(x(t)) \leq 0.$$

Taking the integration from t_0 to t yields the energy estimate. We summarize what has just been shown in the following theorem.
Theorem 6.1 *Given an arbitrary $x_0 \in \mathcal{H}$, there exists a unique global solution trajectory $x : [t_0, \infty) \to \mathcal{H}$ such that $x(t_0) = x_0$ to the system* (DRYAD). *Furthermore, we have the following properties*

- $t \mapsto f(x(t))$ *is decreasing*
- $\displaystyle\int_{t_0}^{\infty} \|\dot{x}(t) + \beta \nabla f(x(t))\|^2 dt < \infty$
- $\displaystyle\int_{t_0}^{\infty} \|\dot{x}(t) + \beta \nabla f(x(t))\| dt < \infty$
- $\displaystyle\int_{t_0}^{\infty} \|\nabla f(x(t))\|^2 dt < \infty.$

Taking advantage of the following form of (DRYAD)

$$\dot{x}(t) + \beta \nabla f(x(t)) = \text{prox}_{\frac{1}{\gamma}\varphi}(-\frac{1}{\gamma} \nabla f(x(t))),$$

it is easy to show, using Lemma 6.1, that if $\|\nabla f(x(t))\| \leq r$ then the system (DRYAD) becomes the gradient flow (continuous steepest descent method), that is, $\dot{x}(t)+\beta\nabla f(x(t)) = 0$. We will see later that $\nabla f(x(t))$ tends to zero as t tends to $+\infty$. There is therefore a change in the nature of the dynamic after a certain time, going from a doubly nonlinear evolution equation to the gradient flow (without perturbation). Since the acceleration of gradient flow is well understood, it is interesting to examine the new dynamics and their convergence properties attached to our approach.
We initially examine the primal problem and subsequently explore its dual approach.

6.2.2 (DRYAD) seen as the perturbed gradient flow: f convex

Let us start from the equivalent formulation of (DRYAD) given by

$$\dot{x}(t) + \beta\nabla f(x(t)) = \text{prox}_{\frac{1}{\gamma}\varphi}(-\frac{1}{\gamma}\nabla f(x(t))).$$

Let us show that the right hand side of the above equality, defined by $g(t) := \text{prox}_{\frac{1}{\gamma}\varphi}(-\frac{1}{\gamma}\nabla f(x(t)))$, satisfies

$$\int_{t_0}^{\infty} \|g(t)\|dt < +\infty.$$

Indeed we have

$$\int_{t_0}^{\infty} \|g(t)\|dt = \int_{\|\nabla f(x(t))\|\leq r} \|g(t)\|dt + \int_{\|\nabla f(x(t))\|>r} \|g(t)\|dt.$$

On the set $\{\|\nabla f(x(t))\| \leq r\}$, we have according to Lemma 6.1

$$g(t) = \text{prox}_{\frac{1}{\gamma}\varphi}(-\frac{1}{\gamma}\nabla f(x(t))) = 0.$$

Hence,

$$\int_{t_0}^{\infty} \|g(t)\|dt = \int_{\|\nabla f(x(t))\|>r} \|g(t)\|dt.$$

Then note that the proximal mapping of φ is nonexpansive, and is equal to zero at zero. So we have

$$\|g(t)\| = \|\text{prox}_{\frac{1}{\gamma}\varphi}(-\frac{1}{\gamma}\nabla f(x(t)))\| \leq \frac{1}{\gamma}\|\nabla f(x(t))\|.$$

On the set $\{\|\nabla f(x(t))\| > r\}$, we deduce that

$$\|g(t)\| \leq \frac{1}{\gamma}\|\nabla f(x(t))\| \leq \frac{1}{\gamma r}\|\nabla f(x(t))\|^2.$$

Therefore,

$$\int_{t_0}^{\infty} \|g(t)\|dt = \int_{\|\nabla f(x(t))\|>r} \|g(t)\|dt \leq \frac{1}{\gamma r} \int_{\|\nabla f(x(t))\|>r} \|\nabla f(x(t))\|^2 dt.$$

Finally we get,

$$\int_{t_0}^{\infty} \|g(t)\|dt \leq \frac{1}{\gamma r} \int_{t_0}^{\infty} \|\nabla f(x(t))\|^2 dt.$$

According to Theorem 6.1 we have ,

$$\int_{t_0}^{\infty} \|\nabla f(x(t))\|^2 dt < +\infty.$$

Thus, (DRYAD) is the gradient flow with a right-hand side in $\mathbb{L}^1(t_0, +\infty)$, which classically preserves the convergence properties of the gradient flow. We refer to [62] for further details on this topic.

We have established that (DRYAD) can be regarded as a perturbed gradient flow with the perturbation belonging to $\mathbb{L}^1(t_0, +\infty)$. As a result, it exhibits the classical convergence properties observed in the gradient flow literature [62]. Additionally, as we will prove in the subsequent section, (DRYAD) eventually transforms into the unperturbed gradient flow after a finite time. Consequently, it inherits all the convergence rates detailed in Theorem 6.5.
Remark 6.1 It is an open question to obtain convergence rates for the perturbed gradient flow equation (6.7) under the sole assumption $\int_{t_0}^{+\infty} \|g(t)\| dt < +\infty$. We know that there is convergence of the trajectories, but in the above result, to get convergence rates, we need also to use the energy assumption $\int_{t_0}^{+\infty} t \|g(t)\|^2 dt < +\infty$.

6.3 A dual approach to (DRYAD)

Examining the dual dynamic of (DRYAD) can help us better comprehend the convergence properties of gradients, which is of fundamental importance in (DRYAD).
To begin with, let us recall our dynamical system

$$\gamma(\dot{x}(t) + \beta\nabla f(x(t))) + \partial\varphi(\dot{x}(t) + \beta\nabla f(x(t))) + \nabla f(x(t)) \ni 0.$$

Set $\Psi(x) = \frac{\gamma}{2}\|x\|^2 + \varphi(x)$. We have $\partial\Psi(x) = (\gamma I + \partial\varphi)(x)$. We transform the original system as follows

$$\begin{aligned}
A := \quad & \gamma(\dot{x}(t) + \beta\nabla f(x(t))) + \partial\varphi(\dot{x}(t) + \beta\nabla f(x(t))) + \nabla f(x(t)) \ni 0,\\
\iff \quad & -\nabla f(x(t)) \in \partial\Psi(\dot{x}(t) + \beta\nabla f(x(t)))\\
\iff \quad & \nabla f(x(t)) \in \partial\Psi(-\dot{x}(t) - \beta\nabla f(x(t))),\\
\iff \quad & -\dot{x}(t) - \beta\nabla f(x(t)) \in \partial\Psi^*(\nabla f(x(t)))\\
\iff \quad & \dot{x}(t) + \partial G(\nabla f(x(t))) \ni 0, \text{ where } G(x) = \frac{\beta}{2}\|x\|^2 + \Psi^*(x)\\
\iff \quad & -\dot{x}(t) \in \partial G(\nabla f(x(t)))\\
\iff \quad & \dot{x}(t) \in \partial G(-\nabla f(x(t)))\\
\iff \quad & -\nabla f(x(t)) \in \partial G^*(\dot{x}(t))\\
\iff \quad & \partial G^*(\dot{x}(t)) + \nabla f(x(t)) \ni 0
\end{aligned}$$

We have $G^* = (\frac{\beta}{2}\|\cdot\|^2 + \Psi^*)^*$ which is exactly the Moreau envelop of Ψ, denoted by Ψ_β. Hence, we have the dual dynamical system of (DRYAD) of the following form

$$\nabla\Psi_\beta(\dot{x}(t)) + \nabla f(x(t)) = 0.$$

Set $g(t) = \nabla f(x(t))$, and the idea is to transform this dynamical system such that the left hand side is a function of $g(t)$. To this end, we have

$$\begin{aligned}
\nabla\Psi_\beta(\dot{x}(t)) + g(t) = 0 &\iff -g(t) = \nabla\Psi_\beta(\dot{x}(t))\\
&\iff \dot{x}(t) = \nabla\Psi_\beta^*(-g(t)).
\end{aligned}$$

Assume that f is a convex function, it follows that

$$g(t) = \nabla f(x(t)) \iff x(t) \in \partial f^*(g(t)).$$

Therefore,

$$\frac{d}{dt}(\partial f^*(g(t))) - \nabla\Psi_\beta^*(-g(t)) \ni 0.$$

Since $\Psi_\beta = (\frac{\beta}{2}\|\cdot\|^2 + \Psi^*)^*$, we have $\Psi_\beta^* = \frac{\beta}{2}\|\cdot\|^2 + \Psi^*$ and hence $\nabla\Psi_\beta^*(x) = \beta x + \nabla\Psi^*(x)$. Let us recall that $\Psi(x) = \frac{\gamma}{2}\|x\|^2 + \varphi(x)$. Based on the specific form of the dry friction,

we can compute the Fenchel conjugate of Ψ as follows

$$\Psi^*(x) = \frac{1}{2\gamma}\text{dist}^2(x, B(0,r)).$$

Hence, we can compute its gradient $\nabla\Psi^*(x) = \frac{1}{\gamma}(x - \text{proj}_{B(0,r)}(x))$. Therefore

$$\nabla\Psi^*_\beta(x) = \beta x + \nabla\Psi^*(x) = \beta x + \frac{1}{\gamma}(x - \text{proj}_{B(0,r)}(x)) = \frac{\gamma\beta+1}{\gamma}x - \frac{1}{\gamma}\text{proj}_{B(0,r)}(x).$$

Plugging this into $\frac{d}{dt}(\partial f^*(g(t))) - \nabla\Psi^*_\beta(-g(t)) \ni 0$, we have

$$\frac{d}{dt}(\partial f^*(g(t))) + \frac{\gamma\beta+1}{\gamma}g(t) - \frac{1}{\gamma}\text{proj}_{B(0,r)}(g(t)) \ni 0. \tag{6.2}$$

Assume that f^* is of class C^2, this system can be equivalently written as

$$\text{(DDRYAD)} \qquad \nabla^2 f^*(g(t))\dot{g}(t) + \frac{\gamma\beta+1}{\gamma}g(t) - \frac{1}{\gamma}\text{proj}_{B(0,r)}(g(t)) = 0,$$

thus making appear the Riemannian structure associated with the Hessian of the convex function f^*.

Let us summarize the above results in the following statement, and establish the convergence rates of the gradients.

Theorem 6.2 *Let $x : [t_0, \infty) \to \mathcal{H}$ be a global solution trajectory of (DRYAD). Suppose that f is convex. Then $g(t) := \nabla f(x(t))$ is a solution trajectory of the generalized Riemannian flow.*

$$\frac{d}{dt}(\partial f^*(g(t))) + \frac{\gamma\beta+1}{\gamma}g(t) - \frac{1}{\gamma}\text{proj}_{B(0,r)}(g(t)) \ni 0.$$

Furthermore, the following convergence properties hold as $t \to \infty$

- *The function $t \mapsto D(g(t), 0)$ is decreasing where*

$$D(g(t), 0) = f^*(0) - f^*(g(t)) + \langle \nabla f^*(g(t)), g(t)\rangle$$

- $\|g(t)\| = \|\nabla f(x(t))\| = o\Big(\frac{1}{\sqrt{t}}\Big)$

Proof. First, we define the Bregman distance function

$$D(g(t), 0) = f^*(0) - f^*(g(t)) + \langle \nabla f^*(g(t)), g(t)\rangle.$$

We have $t \mapsto D(g(t), 0)$ is nonnegative because f^* is convex.

Let us derivate $t \mapsto D(g(t),0)$ to get

$$\begin{aligned}\frac{d}{dt}D(g(t),0) &= -\frac{d}{dt}f^*(g(t)) + \langle\frac{d}{dt}\nabla f^*(g(t)), g(t)\rangle + \langle\nabla f^*(g(t)), \dot{g}(t)\rangle \\ &= \langle\frac{d}{dt}\nabla f^*(g(t)), g(t)\rangle \\ &= -\frac{\gamma\beta+1}{\gamma}\|g(t)\|^2 + \frac{1}{\gamma}\langle\text{proj}_{B(0,r)}(g(t)), g(t)\rangle,\end{aligned}$$

where the last inequality comes from the dual inclusion (6.2).
Define $h(x) = \gamma\Psi^*(x) = \frac{1}{2}\text{dist}^2(x, B(0,r))$. We have h is a smooth and convex function. Therefore, according to the first order characteristic of convex functions we have

$$0 \geq h(g(t)) - \langle g(t) - \text{proj}_{B(0,r)}(g(t)), g(t)\rangle.$$

Therefore,

$$\langle\text{proj}_{B(0,r)}(g(t)), g(t)\rangle = \|g(t)\|^2 - \langle g(t) - \text{proj}_{B(0,r)}(g(t)), g(t)\rangle \leq \|g(t)\|^2 - h(g(t))$$

Using this inequality gives us the following estimate of $\frac{d}{dt}D(g(t),0)$

$$\frac{d}{dt}D(g(t),0) \leq -\beta\|g(t)\|^2 - \frac{1}{\gamma}h(g(t))$$

Considering the integral while acknowledging the nonnegativity of $D(g(t),0)$ yields

$$\int_{t_0}^{t} \beta\|g(s)\|^2 + \frac{1}{\gamma}h(g(s))ds \leq D(g(t_0),0)$$

Due to the nonnegativity of $\beta\|g(s)\|^2 + \frac{1}{\gamma}h(g(s))$, we can take the limit as t tends to infinity and obtain

$$\int_{t_0}^{\infty} \beta\|g(s)\|^2 + \frac{1}{\gamma}h(g(s))ds \leq D(g(t_0),0) < \infty$$

Let us now pass from this integral property to asymptotic properties. Recall that we have

$$\frac{d}{dt}(\partial f^*(g(t))) + \frac{\gamma\beta+1}{\gamma}g(t) - \frac{1}{\gamma}\text{proj}_{B(0,r)}(g(t)) \ni 0$$

Taking the scalar product of both sides of this inclusion with $\dot{g}(t)$, we obtain

$$\begin{aligned}
&\langle \frac{d}{dt}(\partial f^*(g(t))), \dot{g}(t)\rangle + \frac{\beta}{2}\frac{d}{dt}\|g(t)\|^2 + \frac{1}{\gamma}\langle g(t) - \mathrm{proj}_{B(0,r)}(g(t)), \dot{g}(t)\rangle = 0\\
&\Longleftrightarrow \langle \frac{d}{dt}(\partial f^*(g(t))), \dot{g}(t)\rangle + \frac{\beta}{2}\frac{d}{dt}\|g(t)\|^2 + \frac{1}{\gamma}\langle \dot{h}(g(t)), \dot{g}(t)\rangle = 0\\
&\Longleftrightarrow \langle \frac{d}{dt}(\partial f^*(g(t))), \dot{g}(t)\rangle + \frac{\beta}{2}\frac{d}{dt}\|g(t)\|^2 + \frac{1}{\gamma}\frac{d}{dt}h(g(t)) = 0
\end{aligned}$$

Due to the convexity of f^*, we have $\langle \frac{d}{dt}(\partial f^*(g(t))), \dot{g}(t)\rangle \geq 0$. Combining this with the last equality, we have

$$\frac{d}{dt}\Big(\frac{\beta}{2}\|g(t)\|^2 + \frac{1}{\gamma}h(g(t))\Big) \leq 0,$$

which means that the function $t \mapsto \frac{\beta}{2}\|g(t)\|^2 + \frac{1}{\gamma}h(g(t))$ is decreasing. On the other hand, we have

$$\int_{t_0}^{\infty} \frac{\beta}{2}\|g(s)\|^2 + \frac{1}{\gamma}h(g(s))ds \leq \int_{t_0}^{\infty} \beta\|g(s)\|^2 + \frac{1}{\gamma}h(g(s))ds < \infty.$$

Therefore we obtain that

$$\lim_{t\to\infty} t\Big(\frac{\beta}{2}\|g(t)\|^2 + \frac{1}{\gamma}h(g(t))\Big) = 0.$$

Or equivalently (due to the nonnegativity of the underlying functions and the fact that $h(g(t)) \leq \frac{1}{2}\|g(t)\|^2$)

$$\lim_{t\to\infty} t\|g(t)\|^2 = \lim_{t\to\infty} t\|\nabla f(x(t))\|^2 = 0, \text{ or } \|g(t)\| = \|\nabla f(x(t))\| = o\Big(\frac{1}{\sqrt{t}}\Big).$$

The last conclusion is obtained due to the following lemma (see [1, Lemma 5.2]).

Lemma 6.2 *Let $h : [t_0, \infty] \to \mathbb{R}^+$ be a nonincreasing function belonging to $\mathbb{L}^1([t_0, \infty])$. Then it holds that $\lim_{t\to+\infty} th(t) = 0$.*

The proof of Theorem 6.2 is thereby completed. ∎

As a consequence we have the following corollary.

Corollary 6.1 *Suppose that $f\colon \mathcal{H} \to \mathbb{R}$ is a convex differentiable function that satisfies $S = \mathrm{argmin} f \neq \emptyset$. Let $x\colon [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of (DRYAD). Then the following statements are satisfied:*

(i) *(convergence of gradients towards zero)* $\|\nabla f(x(t))\| = o\left(\frac{1}{\sqrt{t}}\right)$ *as* $t \to +\infty$.

(ii) *(integral estimate of the gradients)* $\displaystyle\int_{t_0}^{+\infty} t\left\|\nabla f\left(x\left(t\right)\right)\right\|^2 dt < +\infty$.

(iii) *(convergence of values)* $f\left(x\left(t\right)\right) - \inf_{\mathcal{H}} f = o\left(\frac{1}{t}\right)$ *as* $t \to +\infty$.

(iv) *The solution trajectory* $x(\cdot)$ *converges weakly as* $t \to +\infty$*, and its limit belongs to* $S = \operatorname{argmin} f$.

Proof. It follows from Theorem 6.2 that $\nabla f(x(t))$ approaches zero as t approaches infinity, precisely displayed in the first item. As previously stated, when $\|\nabla f(x(t))\| \leq r$, the system becomes the gradient flow. Consequently, there exists a $T > 0$ such that for all $t \geq T$, the system is the gradient flow. Therefore, any solution trajectory generated by (DRYAD) inherits all the convergence rates of the continuous steepest descent method listed in Theorem 6.5 (in the appendix) which includes the last three items of this theorem. ■

When $\mathcal{H}$ is a finite dimensional Euclidian space, we have the following corollary concerning the convergence property (DRYAD) when the objective function f is not necessarily convex but satisfies the Kurdyka–Lojasiewicz property. Let us recall some basic facts concerning the Kurdyka–Lojasiewicz property, which we briefly designate by (KL). No convexity assumption is made on the function f to be minimized. A function $f : \mathbb{R}^N \to \mathbb{R}$ satisfies the (KL) property if its values can be reparametrized in the neighborhood of each of its critical points so that the resulting function becomes sharp. This means that there exists a continuous, concave, increasing function θ such that for all u in a slice of f, we have

$$\|\nabla(\theta \circ f)(u)\| \geq 1.$$

The function θ captures the geometry of f around its critical points, and is called a desingularizing function; see [25, 26], for further details.

Corollary 6.2 *Suppose that* $f : \mathbb{R}^N \to \mathbb{R}$ *is a differentiable function that satisfies the (KL) property. Then, any bounded solution trajectory of (DRYAD) has a finite length and hence converges to a critical point of* f.

Proof. Since (DRYAD) arrives at the regime of the gradient flow from a sufficiently large time, the statement of the corollary follows from the result of Lojasiewicz [104] ■

Remark 6.2 Let us present a dual viewpoint on the finite time stabilization property. Since $\nabla f(x(t))$ converges to zero as $t \to \infty$, from a sufficiently large time T, we have $\nabla f(x(t)) \in B(0, r)$ and hence $g(t) - \operatorname{proj}_{B(0,r)}(g(t)) = 0$. As a result, the dual dynamical system becomes

$$\frac{d}{dt}(\partial f^*(g(t))) + \beta g(t) = 0,$$

or equivalently

$$\nabla^2 f^*(g(t))\dot{g}(t) + \beta g(t) = 0,$$

when f^* is assumed to be twice continuously differentiable.

6.4 Applying the time scaling and averaging techniques to (DRYAD)

Before going further, let us mention that these techniques of scaling and averaging were initiated by Attouch, Bot and Nguyen [28] for the gradient flow method. In our specific case, we have adapted these techniques to address our problem, and we now recapitulate only the essential elements required for our analysis.

6.4.1 Time scaling

The time scaling technique is in fact a change of variable $t = \tau(s)$, where τ is an increasing function from $\mathbb{R}^+$ to $\mathbb{R}^+$ which is continuously differentiable and which tends to ∞ when $s \to \infty$. Set

$$y(s) = x(\tau(s))$$

We have

$$\dot{y}(s) = \dot{\tau}(s)\dot{x}(\tau(s))$$

As a result, to obtain the corresponding dynamical system associated with the new trajectory $y(s)$, we respectively replace $x(t)$ and $\dot{x}(t)$ in the original system with $y(s)$ and $\frac{\dot{y}(s)}{\dot{\tau}(s)}$. To this end, we obtain

$$\gamma\Big(\frac{\dot{y}(s)}{\dot{\tau}(s)} + \beta\nabla f(y(s))\Big) + \partial\varphi\Big(\frac{\dot{y}(s)}{\dot{\tau}(s)} + \beta\nabla f(y(s))\Big) + \nabla f(y(s)) \ni 0.$$

Using the positive homogeneity of degree zero of $\partial\varphi$, this system can be simplified as

$$\dot{y}(s) + \frac{\dot{\tau}(s)}{\gamma}\partial\varphi\Big(\dot{y}(s) + \beta\dot{\tau}(s)\nabla f(y(s))\Big) + \frac{\gamma\beta + 1}{\gamma}\dot{\tau}(s)\nabla f(y(s)) \ni 0 \tag{6.3}$$

6.4.2 Averaging

Let us attach to $y(\cdot)$ the new function $z : [s_0, \infty] \to \mathcal{H}$ defined by

$$\dot{z}(s) + \frac{1}{\dot{\tau}(s)}(z(s) - y(s)) = 0,$$

with $z(s_0) = y(s_0) = x_0$ given in $\mathcal{H}$. Equivalently,

$$y(s) = z(s) + \dot{\tau}(s)\dot{z}(s).$$

By temporal derivation we have

$$\dot{y}(s) = (1 + \ddot{\tau}(s))\dot{z}(s) + \dot{\tau}(s)\ddot{z}(s)$$

After plugging $y(s)$ and $\dot{y}(s)$ by their expressions in terms of $z(t)$ into the dynamical system obtained from time scaling (6.3) and dividing both sides by $\dot{\tau}(s)$ we have

$$\ddot{z}(s) + \frac{\ddot{\tau}(s)+1}{\dot{\tau}(s)}\dot{z}(s) + \frac{\gamma\beta+1}{\gamma}\nabla f(z(s) + \dot{\tau}(s)\dot{z}(s)) + \frac{1}{\gamma}\partial\varphi(a(s)) \ni 0 \tag{6.4}$$

where $a(s) = \dot{y}(s) + \beta\dot{\tau}(s)\nabla f(y(s))$.
Let us return to the dynamical system (6.3) and express it using the notation $a(s)$. So we have

$$a(s) + \frac{\dot{\tau}(s)}{\gamma}\partial\varphi(a(s)) + \frac{\dot{\tau}(s)}{\gamma}\nabla f(y(s)) = 0$$

Therefore

$$\begin{aligned}
\partial\varphi(a(s)) &= -\frac{\gamma}{\dot{\tau}(s)}a(s) - \nabla f(y(s)) \\
&= -\frac{\gamma}{\dot{\tau}(s)}\Big(I + \frac{\dot{\tau}(s)}{\gamma}\partial\varphi\Big)^{-1}\Big(-\frac{\dot{\tau}(s)}{\gamma}\nabla f(y(s))\Big) - \nabla f(y(s)) \\
&= \frac{\gamma}{\dot{\tau}(s)}\Big[-\Big(I + \frac{\dot{\tau}(s)}{\gamma}\partial\varphi\Big)^{-1}\Big(-\frac{\dot{\tau}(s)}{\gamma}\nabla f(y(s))\Big) - \frac{\dot{\tau}(s)}{\gamma}\nabla f(y(s))\Big] \\
&= \nabla\varphi_{\frac{\dot{\tau}(s)}{\gamma}}(-\frac{\dot{\tau}(s)}{\gamma}\nabla f(y(s))) \\
&= \nabla\varphi_{\frac{\dot{\tau}(s)}{\gamma}}\Big(-\frac{\dot{\tau}(s)}{\gamma}\nabla f(z(s) + \dot{\tau}(s)\dot{z}(s))\Big)
\end{aligned}$$

Plugging this into the dynamical system (6.4) gives

$$\ddot{z}(s) + \frac{\ddot{\tau}(s)+1}{\dot{\tau}(s)}\dot{z}(s) + \frac{\gamma\beta+1}{\gamma}\nabla f(z(s)+\dot{\tau}(s)\dot{z}(s)) \tag{6.5}$$
$$+ \frac{1}{\gamma}\nabla\varphi_{\frac{\dot{\tau}(s)}{\gamma}}\Big(-\frac{\dot{\tau}(s)}{\gamma}\nabla f(z(s)+\dot{\tau}(s)\dot{z}(s))\Big) = 0$$
(6.6)

Finally we have obtained a second order dynamical system by doing time scaling and averaging from the original first order dynamic. In order to have fast convergence properties, we choose τ such that the viscous damping coefficient in this dynamical system asymptotically vanishes as follows

$$\frac{\ddot{\tau}(s)+1}{\dot{\tau}(s)} = \frac{\alpha}{s},$$

for some $\alpha > 1$. We can easily show that this is achieved by setting $\tau(s) = \frac{s^2}{2(\alpha-1)}$. Hence, the dynamical system (6.5) becomes the following which we will call (iDRYAD)

$$\ddot{z}(s) + \frac{\alpha}{s}\dot{z}(s) + \frac{\gamma\beta+1}{\gamma}\nabla f\Big(z(s)+\frac{s}{\alpha-1}\dot{z}(s)\Big)$$
$$+ \frac{1}{\gamma}\nabla\varphi_{\frac{s}{\gamma(\alpha-1)}}\Big(-\frac{s}{\gamma(\alpha-1)}\nabla f(z(s)+\frac{s}{\alpha-1}\dot{z}(s))\Big) = 0$$

The corresponding convergence properties for $z(s)$ with this specific choice of τ are captured in the following theorem

Theorem 6.3 *Let f be a convex smooth function whose gradient is Lipschitz continuous on the bounded sets and such that argminf is non empty. Assume $\alpha > 3$, let $z : [s_0, \infty] \to \mathcal{H}$ be a solution trajectory of*

$$\ddot{z}(s) + \frac{\alpha}{s}\dot{z}(s) + \frac{\gamma\beta+1}{\gamma}\nabla f\Big(z(s)+\frac{s}{\alpha-1}\dot{z}(s)\Big)$$
$$+ \frac{1}{\gamma}\nabla\varphi_{\frac{s}{\gamma(\alpha-1)}}\Big(-\frac{s}{\gamma(\alpha-1)}\nabla f(z(s)+\frac{s}{\alpha-1}\dot{z}(s))\Big) = 0$$

Then we have the following properties

- $f(z(s)) - \inf_{\mathcal{H}} f = \mathcal{O}(1/s^2)$
- $\|\nabla f(z(s))\| = \mathcal{O}(1/s)$
- $\displaystyle\int_{s_0}^{\infty} s^3\|\nabla f(z(s)+\frac{s}{\alpha-1}\dot{z}(s))\|^2 ds < \infty$
- *only assume that $\alpha > 1$, then $z(s)$ converges weakly and its limit belongs to argmin f*

Proof. Since our original system follows the steepest descent method after a finite time, the results are achieved according to [28]. Let us present the main lines. First,

we interpret the transition from y to z as an averaging process. Precisely, rewriting the relation between y and z we obtain

$$s\dot{z}(s) + (\alpha - 1)z(s) = (\alpha - 1)y(s)$$

Multiplying this equation with $s^{\alpha-2}$, we get

$$s^{\alpha-1}\dot{z}(s) + (\alpha - 1)s^{\alpha-2}z(s) = (\alpha - 1)s^{\alpha-2}y(s),$$

which is equivalent to

$$\frac{d}{ds}(s^{\alpha-1}z(s)) = (\alpha - 1)s^{s-2}y(s).$$

Integrating this equation from s_0 to s, we obtain

$$z(s) = \frac{s_0^{\alpha-1}}{s^{\alpha-1}}y(s_0) + \frac{\alpha - 1}{s^{\alpha-1}}\int_{s_0}^{s} u^{\alpha-2}y(u)du,$$

which can be written abstractly as

$$z(s) = \int_{s_0}^{s} y(u)d\mu_s(u),$$

where μ_s is the positive Radon measure $[s_0, s]$ defined by

$$\mu_s = \frac{s_0^{\alpha-1}}{s^{\alpha-1}}\delta_{s_0} + (\alpha - 1)\frac{u^{\alpha-2}}{s^{\alpha-1}}du,$$

where δ_{s_0} denotes the Dirac measure at s_0. Since μ_s is positive and has the integral over $[s_0, s]$ being 1, it is a probability measure. It is clear that $z(s)$ can be seen as the average of trajectory $y(\cdot)$ on $[s_0, s]$ with respect to μ_s.
For the first item of the theorem, we use the Lipschitz continuity of the gradient of f

$$f(z(s)) - \inf_{\mathcal{H}} f \leq f\Big(\int_{s_0}^{s} y(u)d\mu_s(u)\Big) - \inf_{\mathcal{H}} f + \mathcal{O}(1/s^2).$$

What remains is to show that

$$f\Big(\int_{s_0}^{s} y(u)d\mu_s(u)\Big) - \inf_{\mathcal{H}} f = \mathcal{O}(1/s^2),$$

which can be achieved by making use of the convexity of f, and hence the Jensen inequality.

For the second item, since $\nabla f(x_*) = 0$ and according to [115, Theorem 2.1.5], we have

$$\frac{1}{2L}\|f(z(s))\|^2 \leq f(z(s)) - \inf_{\mathcal{H}} f.$$

This inequality combined with the first item gives us the result.

For the third item, according to Corollary 6.1, we have for any solution trajectory of (DRYAD) that

$$\int_{t_0}^{+\infty} t \left\|\nabla f\left(x\left(t\right)\right)\right\|^2 dt < +\infty.$$

Making a change of variable associated with the time scaling step

$$t = \tau(s) = \frac{s^2}{2(\alpha - 1)},$$

we obtain

$$\int_{t_0}^{+\infty} s^3 \left\|\nabla f\left(y\left(s\right)\right)\right\|^2 dt < +\infty,$$

where $y(s) = x(\tau(s))$. In light of the averaging step, we replace $y(s)$ with $x(s) + \frac{s}{\alpha-1}\dot{x}(s)$ and obtain the result of this item.

For the last item which concerns the weak convergence of $z(s)$, we argue as follows. We know that the solution trajectory of the steepest descent dynamic converges weakly to a solution $x_* \in S = \operatorname{argmin} f$. This immediately gives that $y(s) = x(\tau(s))$ converges weakly to x_*. To pass from this result to the result of $z(\cdot)$, we use the interpretation of z as an average of y

$$z(s) = \frac{s_0^{\alpha-1}}{s^{\alpha-1}} y(s_0) + \frac{\alpha - 1}{s^{\alpha-1}} \int_{s_0}^{s} u^{\alpha-2} y(u) du.$$

In order to have the weak convergence of $z(\cdot)$, meaning $\langle z(s), v\rangle \to \langle x_*, v\rangle$ as $s \to \infty$ for all $v \in \mathcal{H}$, after elementary calculus, it is sufficient to require that if $a(\cdot)$ is a positive real valued function which satisfies $\lim_{r\to\infty} a(r) = 0$, then $\lim_{s\to\infty} A(s) = 0$, where

$$A(s) = \frac{\alpha - 1}{s^{\alpha-1}} \int_{s_0}^{s} r^{\alpha-2} a(r) dr,$$

which indeed can be proven to be true. ■

6.5 Applying the time scaling and averaging techniques to the dual system (DDRYAD)

Let us recall the dual dynamical system

$$\frac{d}{dt}(\partial f^*(g(t))) + \nabla\Psi^*_\beta(g(t)) \ni 0$$

Performing similarly to the case of (DRYAD), we can apply the time scaling and averaging to the dual system to obtain the following second order dynamic

$$\text{(iDDRYAD)} \qquad \nabla^2 f^*\Big(w(s) + \frac{s}{\alpha-1}\dot{w}(s)\Big)\Big(\ddot{w}(s) + \frac{\alpha}{s}\dot{w}(s)\Big) + \nabla\Psi^*_\beta\Big(w(s) + \frac{s}{\alpha-1}\dot{w}(s)\Big) = 0,$$

where the relation of $g(t)$ with $w(t)$ is as follows

$$\begin{cases} v(s) = g(\tau(s)), \\ v(s) = w(s) + \dot{\tau}(s) + \dot{w}(s) \end{cases}, \text{ with } \tau(s) = \frac{s^2}{2(\alpha-1)}$$

Here v is associated with the scaling step and w is associated with the averaging step. Recall that for the dual dynamic, given that f is convex we have

$$\lim_{t\to\infty} t\|g(t)\|^2 = 0$$

As $\tau(s) \to \infty$ as $s \to \infty$, we can replace t with $\tau(s)$ in the above limit to obtain

$$\lim_{s\to\infty} s\|v(s)\| = 0, \text{ or } \|v(s)\| = o(1/s)$$

The differential equation connecting v and w gives us the interpretation of w as an average of v as follows:

$$w(s) = \int_{s_0}^{s} v(u)d\mu_s(u),$$

where $\mu_s = \frac{s_0^{\alpha-1}}{s^{\alpha-1}}\delta_{s_0} + (\alpha-1)\frac{u^{\alpha-2}}{s^{\alpha-1}}du$ is a probability measure on $[s_0, s]$. Here δ_{s_0} denotes the Dirac measure at s_0.
According to the convexity of $\|\cdot\|$ and the Jensen inequality, we obtain that

$$\|w(s)\| = o(1/s)$$

The following theorem summarizes the results we just showed for this second order dual dynamic

Theorem 6.4 *Let $x : [t_0, \infty) \to \mathcal{H}$ be a global solution trajectory of (DRYAD). Suppose that f is convex. Then $g(t) := \nabla f(x(t))$ is the solution trajectory of the generalized Riemannian flow*

$$\frac{d}{dt}(\partial f^*(g(t))) + \frac{\gamma\beta + 1}{\gamma} g(t) - \frac{1}{\gamma} proj_{B(0,r)}(g(t)) \ni 0.$$

Set $\tau(s) = \frac{s^2}{2(\alpha-1)}$ with $\alpha > 1$, and $v(s) = g(\tau(s))$. Define w as the solution of the differential equation

$$\dot{w}(s) + \frac{1}{\dot{\tau}(s)}(w(s) - v(s)) = 0, \text{ with } w(s_0) = v(s_0) = x_0.$$

Then w satisfies the following inertial system

$$\nabla^2 f^*\Big(w(s) + \frac{s}{\alpha - 1}\dot{w}(s)\Big)\Big(\ddot{w}(s) + \frac{\alpha}{s}\dot{w}(s)\Big) + \nabla \Psi_\beta^*\Big(w(s) + \frac{s}{\alpha - 1}\dot{w}(s)\Big) = 0,$$

and we have $\|w(s)\| = o(1/s)$ as $s \to \infty$.

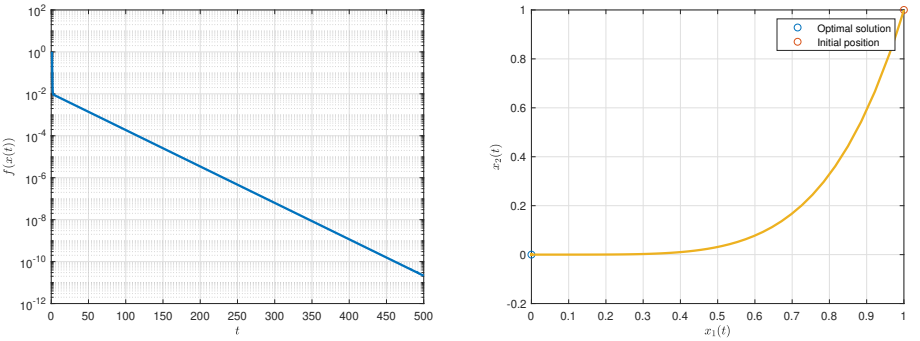
6.6 Numerical results

In this section, we will use adapted standard Runge-Kutta methods to solve numerically the involved continuous dynamics and conduct a series of numerical illustrative experiments to illustrate the theoretical results discussed in the previous sections.

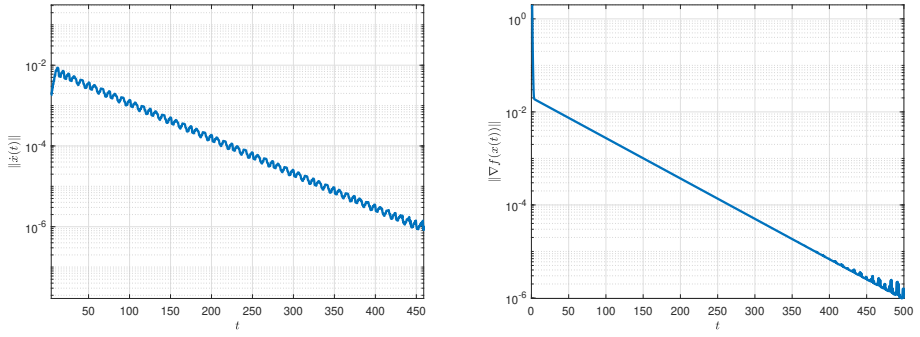
Example 6.1 Let us begin this section by considering an example to illustrate the dynamic (DRYAD) in dimension 2 in the case of a convex and quadratic function. More precisely, let us set $f(x_1, x_2) = ax_1^2 + bx_2^2$ with $0 \leq a < b$ and the initial condition $x(1) = (1, 1)$ and $\dot{x}(1) = (0, 0)$. Note that f is of the form $f(x) = \langle x, Qx\rangle$ with $Q = \text{diag}([a, b])$. We take $\varphi(x) = r\|x\|_2$, with $r = 0.1$.

In Figure 6.1 we illustrate the behaviors of several quantities associated with (DRYAD). Figure 6.1(a) shows the value of the objective function f along the trajectory as a function of time. We can see that the function value decreases over time which is in accordance with the theoretical result. Figure 6.1(b) shows the trajectory of the system starting from an initial position and finally ending up at the unique minimizer of f. The last two figures display the convergences towards zero of the velocity and gradient vectors.

Example 6.2 Let us now compare the two primal dynamics (DRYAD) and (iDRYAD) on a quadratic function in dimension 2.



(a) Behavior of the objective **(b)** Solution trajectory

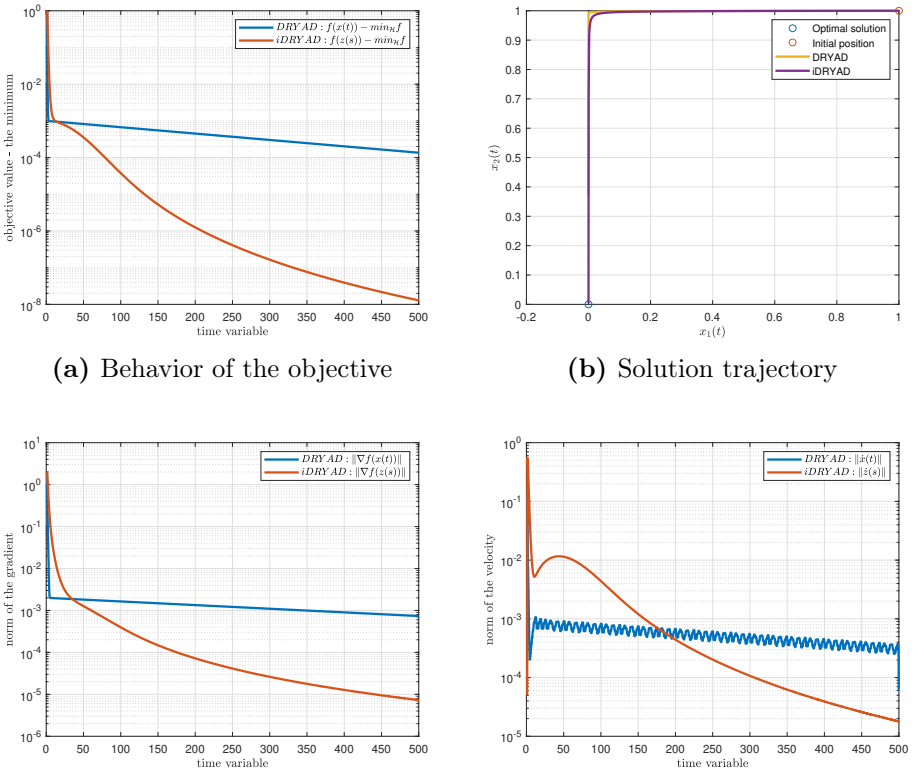


(c) Behavior of $\|\dot{x}(t)\|$ **(d)** Behavior of $\|\nabla f(x(t))\|$

Figure 6.1: Illustration of the convergence results of (DRYAD).

Similarly to Example 6.1, we consider 4 comparison criteria in Figure 6.2, namely the objective value, solution trajectories, norm of the gradient, and norm of the velocity. As can be seen from the figures, (iDRYAD) displays a superior performance compared to (DRYAD) which confirms our theoretical results.

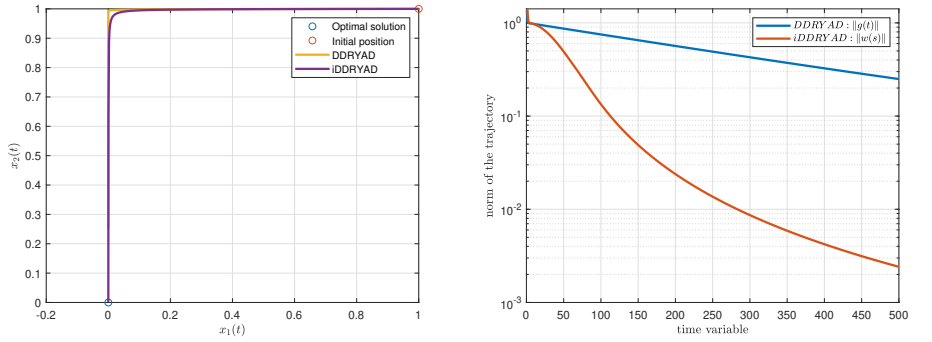
Example 6.3 We conduct the same numerical experiments as the previous example to compare the two dual dynamics (DDRYAD) and (iDDRYAD).
In this comparison, in addition to providing the solution trajectories of (DDRYAD) and (iDDRYAD) on the plane, we also present the evolutions of the norm of their trajectories, which are supposed to converge to zero according to the theoretical results. Clearly, (iDDRYAD) outperforms (DDRYAD).
Example 6.4 Let us conclude the numerical tests with this example where we bring together the 4 dynamics, namely (DRYAD), (iDRYAD), (DDRYAD), and (iDDRYAD) into one plot. To this end, we will display the norms of the gradients of the two primal dynamics' trajectories and the norms of the two dual dynamics' trajectories. We will see the evolutions of these 4 quantities which are supposed to converge to zero as time tends to infinity. We use a quadratic problem for this numerical test.



(a) Behavior of the objective **(b)** Solution trajectory

(c) Behavior of the gradient **(d)** Behavior of the velocity

Figure 6.2: Comparison between (DRYAD) and (iDRYAD)



(a) Solution trajectory **(b)** Behavior of norm of the trajectory

Figure 6.3: Comparison between (DDRYAD) and (iDDRYAD)

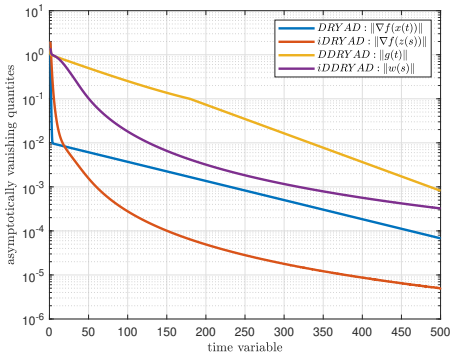


Figure 6.4: Comparing the four dynamics

If we look at Figure 6.4, in addition to the observation that the second order dynamics have faster convergences than the first order ones, which has been seen in previous examples, we can also say that the primal dynamics seem to have slightly better performances than the dual dynamics in terms of convergence to zero of their respective considered quantities

In summary, based on the above examples and the theoretical results, it is evident that inertial dynamics exhibit accelerated convergence rates for objective values, gradient norms and velocity vectors in both primal and dual contexts. These enhanced convergence properties present important advantages in the optimization field by facilitating more effective and efficient optimization processes. The ability to achieve faster convergence means that optimal or near-optimal solutions can be obtained more quickly, reducing computation time and improving resource utilization. These benefits not only improve overall optimization efficiency but also the performance of various optimization applications. Consequently, the enhanced convergence properties of inertial dynamics make them extremely valuable and desirable tools in the optimization field.

6.7 Conclusion

In this chapter, we study the long-time behavior of inertial dynamics with dry friction in a Hilbert setting for convex differentiable optimization problems. The analysis made use of the the time scaling and averaging techniques developed by Attouch, Bot and Nguyen [28] to accelerate first order dynamical systems. We initially study a doubly nonlinear first-order evolution equation, and subsequently adopt the mentioned acceleration method to obtain a second-order in time evolution system involving dry friction, asymptotically vanishing viscous damping, and a damping driven by the Hessian in the implicit form. The obtained accelerated convergence rates of the inertial dynamic do not require developing a Lyapunov analysis, but instead rely on the convergence results of the original first-order system and tools from differential and integral calculus. However, there are several questions

that require more research regarding this topic. One natural direction is to study the case where the objective function f is a nonsmooth convex function by replacing the gradient terms in (DRYAD) with the subdifferential of f. Another important area is to design from (DRYAD) the associated first order dynamic representing the minimization problem of additive functions where f is the sum of a smooth and a nonsmooth function. Regarding the dual formulation of (DRYAD), the double differentiability of the Legendre-Frenchel transform f^* of the convex function f plays a crucial role in the formulation of the dual dynamic approach. This poses an issue since this assumption may not be available in practice. The numerical experiments highlight the accelerated convergence properties of inertial dynamics as opposed to their first order counterparts. While we have focused on continuous-time scenarios in this chapter, it is important to explore the temporal discretization of these dynamics and examine the convergence properties of the associated algorithms. In addition, it would be interesting to carry out tests on various optimization problems at different scales. This research will enable us to gain a deeper understanding of these algorithms, their optimization efficiency and their applicability to a wider range of problem sizes.

6.8 Appendix

6.8.1 Asymptotic convergence rates for the perturbed gradient flow

Let us provide asymptotic convergence rates for the perturbed gradient flow, which we have relied on in previous sections.

Theorem 6.5 *Suppose that $f\colon \mathcal{H} \to \mathbb{R}$ is a convex differentiable function that satisfies $S = \operatorname{argmin} f \neq \emptyset$ and that has Lipschitz continuous gradient on bounded sets. Let $z\colon [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of*

$$\dot{z}(t) + \nabla f(z(t)) = g(t) \tag{6.7}$$

where $g\colon [t_0, +\infty[\to \mathcal{H}$ is such that

$$\int_{t_0}^{+\infty} \|g(t)\|\, dt < +\infty \text{ and } \int_{t_0}^{+\infty} t\, \|g(t)\|^2\, dt < +\infty. \tag{6.8}$$

Then the following statements are satisfied:

(i) *(convergence of gradients towards zero)* $\|\nabla f(z(t))\| = o\left(\frac{1}{\sqrt{t}}\right)$ *as* $t \to +\infty$.

(ii) *(integral estimate of the gradients)* $\int_{t_0}^{+\infty} t\, \|\nabla f(z(t))\|^2\, dt < +\infty$.

(iii) *(convergence of values)* $f(z(t)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{t}\right)$ *as* $t \to +\infty$.

(iv) *The solution trajectory* $z(\cdot)$ *converges weakly as* $t \to +\infty$*, and its limit belongs to* $S = \operatorname{argmin} f$.

Proof. For the sake of completeness, let us recall some of the arguments used in the asymptotic analysis of the perturbed steepest descent system when the perturbation g satisfies (6.8). Given $z_* \in S$, let $T > t_0$ be fixed and for every $t_0 \leq t \leq T$ consider

$$\mathcal{E}_T(t) := t\left(f(z(t)) - \inf_{\mathcal{H}} f\right) + \frac{1}{2}\|z(t) - z_*\|^2 + \int_t^T \langle z(\tau) - z_* + \tau \dot{z}(\tau), g(\tau)\rangle \, d\tau.$$

Differentiating $\mathcal{E}_T$ gives for all $T \geq t \geq t_0$

$$\begin{aligned}\frac{d}{dt}\mathcal{E}_T(t) &= f(z(t)) - \inf_{\mathcal{H}} f + t\langle \nabla f(z(t)), \dot{z}(t)\rangle + \langle z(t) - z_*, \dot{z}(t) - g(t)\rangle - t\langle \dot{z}(t), g(t)\rangle \\ &= f(z(t)) - \inf_{\mathcal{H}} f - t\|\dot{z}(t)\|^2 - \langle z(t) - z_*, \nabla f(z(t))\rangle \\ &\leq -t\|\dot{z}(t)\|^2, \end{aligned} \tag{6.9}$$

where the second equality comes from (6.7), and the last inequality follows from the convexity of f. By integration from t_0 to t, we deduce that

$$\begin{aligned} & t\left(f(z(t)) - \inf_{\mathcal{H}} f\right) + \frac{1}{2}\|z(t) - z_*\|^2 \\ \leq\ & C + \int_{t_0}^t \langle z(\tau) - z_*, g(\tau)\rangle \, d\tau + \int_{t_0}^t \tau\langle \dot{z}(\tau), g(\tau)\rangle \, d\tau - \int_{t_0}^t \tau\|\dot{z}(\tau)\|^2 \, d\tau \\ \leq\ & C + \int_{t_0}^t \|z(\tau) - z_*\|\,\|g(\tau)\| \, d\tau + \frac{1}{2}\int_{t_0}^{+\infty} \tau\|g(\tau)\|^2 \, d\tau - \frac{1}{2}\int_{t_0}^t \tau\|\dot{z}(\tau)\|^2 \, d\tau. \end{aligned} \tag{6.10}$$

We obtain the following estimate (recall that C denotes a generic constant), satisfied for all $t \geq t_0$

$$\frac{1}{2}\|z(t) - z_*\|^2 \leq C + \int_{t_0}^t \|z(\tau) - z_*\|\,\|g(\tau)\| \, d\tau.$$

According to Gronwall Lemma, we conclude that for all $t \geq t_0$

$$\|z(\tau) - z_*\| \leq \sqrt{2C} + \int_{t_0}^t \|g(\tau)\| \, d\tau \leq \sqrt{2C} + \int_{t_0}^{+\infty} \|g(\tau)\| \, d\tau < +\infty.$$

The trajectory is therefore bounded. Using this property and (6.8) allows us to as-

sert from (6.10) that

$$t\left(f\left(z\left(t\right)\right)-\inf_{\mathcal{H}} f\right)+\frac{1}{2}\left\|z\left(t\right)-z_{*}\right\|^{2}+\frac{1}{2}\int_{t_0}^{t}\tau\left\|\dot{z}\left(\tau\right)\right\|^{2}d\tau\leq C. \tag{6.11}$$

The above estimate does not depend on T, so it is satisfied for all $t\geq t_0$. It immediately gives the convergence rate of the values for the solution trajectories of the perturbed (SD)

$$f(z(t))-\inf_{\mathcal{H}} f\leq\frac{C}{t}.$$

Since $t\left(f\left(z\left(t\right)\right)-\inf_{\mathcal{H}} f\right)+\frac{1}{2}\left\|z\left(t\right)-z_{*}\right\|^{2}\geq 0$ for $t\geq t_0$, letting t goes to $+\infty$ in (6.11), we get

$$\int_{t_0}^{+\infty}t\left\|\dot{z}\left(t\right)\right\|^{2}dt<+\infty.$$

According to the constitutive equation (6.7) we get

$$\int_{t_0}^{+\infty}t\left\|\nabla f(z(t)\right\|^{2}dt\leq 2\int_{t_0}^{+\infty}t\left\|\dot{z}\left(t\right)\right\|^{2}dt+2\int_{t_0}^{+\infty}t\left\|g\left(t\right)\right\|^{2}dt<+\infty.$$

Let us differentiate the anchor function, which is another classical ingredient of the Lyapunov analysis

$$\begin{aligned}\frac{d}{dt}\left(\frac{1}{2}\left\|z\left(t\right)-z_{*}\right\|^{2}\right)&=\left\langle z\left(t\right)-z_{*},\dot{z}\left(t\right)\right\rangle\\&=-\left\langle z\left(t\right)-z_{*},\nabla f\left(z\left(t\right)\right)\right\rangle-\left\langle z\left(t\right)-z_{*},g\left(t\right)\right\rangle\\&\leq-\left(f\left(z\left(t\right)\right)-\inf_{\mathcal{H}} f\right)+\sup_{t\geq t_0}\left\|z\left(t\right)-z_{*}\right\|\cdot\left\|g\left(t\right)\right\| \qquad (6.12)\\&\leq\sup_{t\geq t_0}\left\|z\left(t\right)-z_{*}\right\|\cdot\left\|g\left(t\right)\right\|. \qquad (6.13)\end{aligned}$$

Recall that the trajectory $z(\cdot)$ is bounded. According to assumption (6.8), we deduce that the right hand side of (6.13) belongs to $\mathbb{L}^{1}\left([t_0,+\infty[\right)$. Therefore, from [1, Lemma 5.1] we obtain that

$$\lim_{t\to+\infty}\left\|z\left(t\right)-z_{*}\right\|^{2}\in\mathbb{R}\ \text{ exists}$$

and so $\lim_{t\to+\infty}\left\|z\left(t\right)-z_{*}\right\|\in\mathbb{R}$ does. In other words, the first condition of Opial's lemma is fulfilled. Furthermore, since $\lim_{t\to+\infty}f\left(z\left(t\right)\right)=\inf_{\mathcal{H}} f$ and f is convex and weakly lower semicontinuous, the second condition of Opial's lemma is also fulfilled. This gives the

weak convergence of the trajectory $z(t)$ as $t \to +\infty$ to an element in $S = \operatorname{argmin} f$. Now let us show that in fact

$$\lim_{t\to+\infty} t\left(f\left(z\left(t\right)\right) - \inf_{\mathcal{H}} f\right) = 0,$$

meaning that the convergence rate of $f\left(z\left(t\right)\right) - \inf_{\mathcal{H}} f$ is actually $o\left(1/t\right)$. To see this, we integrate (6.12) from t_0 to $t > t_0$ and then let t converge to $+\infty$. This yields

$$\int_{t_0}^{+\infty} \frac{1}{t} t\left(f\left(z\left(t\right)\right) - \inf_{\mathcal{H}} f\right) dt = \int_{t_0}^{+\infty} \left(f\left(z\left(t\right)\right) - \inf_{\mathcal{H}} f\right) dt < +\infty \tag{6.14}$$

and thus $\liminf_{t\to+\infty} t\left(f\left(z\left(t\right)\right) - \inf_{\mathcal{H}} f\right) = 0$. It remains to show that this limit exists. To this end we compute the time derivative of $t\left(f\left(z\left(t\right)\right) - \inf_{\mathcal{H}} f\right)$ and apply once again [1, Lemma 5.1]

$$\begin{aligned}
\frac{d}{dt}\left(t\left(f\left(z\left(t\right)\right) - \inf_{\mathcal{H}} f\right)\right) &= f\left(z\left(t\right)\right) - \inf_{\mathcal{H}} f + t\left\langle \nabla f\left(z\left(t\right)\right), \dot{z}\left(t\right)\right\rangle \\
&= f\left(z\left(t\right)\right) - \inf_{\mathcal{H}} f - t\left\|\dot{z}\left(t\right)\right\|^2 - t\left\langle g\left(t\right), \dot{z}\left(t\right)\right\rangle \\
&\leq f\left(z\left(t\right)\right) - \inf_{\mathcal{H}} f + \frac{1}{4} t\left\|g\left(t\right)\right\|^2.
\end{aligned}$$

Statement (iii) follows from assumption (6.8) and (6.14).
Let L be the Lipschitz constant of ∇f on a ball containing the trajectory $z(\cdot)$. It follows from [34, Lemma 1] that for every $t \geq t_0$

$$0 \leq \frac{t}{2L}\left\|\nabla f(z(t))\right\|^2 \leq t(f(z(t)) - \inf_{\mathcal{H}} f),$$

which implies that $\lim_{t\to+\infty} t\left\|\nabla f(z(t))\right\|^2 = 0$ and proves (i). ∎

7

Conclusion and perspectives

This thesis encompasses two primary subjects. The first subject explores the Pareto eigenvalue complementarity problem alongside its corresponding inverse problem, while the second delves into first-order optimization from the perspective of continuous dynamical systems.

Specifically, in the exploration of the first topic, given a matrix of size $n \times n$, we investigate the following system:

$$K \ni x \perp (\lambda x - Ax) \in K^*, \tag{7.1}$$

where K represents the nonnegative orthant, a closed convex cone in $\mathbb{R}^n$, and K^* denotes the positive dual cone of K, defined by

$$K^* = \{y \in \mathbb{R}^n : \langle y, x\rangle \geq 0 \quad \forall x \in K\}.$$

Any $\lambda \in \mathbb{R}$ that satisfies this condition with some non-zero λ is termed a Pareto eigenvalue of A. The exponential growth in the number of Pareto eigenvalues of a matrix with its size underscores the challenge of identifying all Pareto eigenvalues for medium to large-scale matrices. we propose the utilization of interior-point methods to address Pareto eigenvalue complementarity problems. Here, we demonstrate the efficacy of a nonparametric interior point method (NPIPM) and an adapted Mehrotra predictor-corrector method (MPCM), a widely acknowledged primal-dual interior point method. These methods exhibit efficiency across various problem instances, including those with real-world data. Additionally, we also present a proposed nonparametric smoothing method that demonstrates notable strength in resolving Pareto eigenvalue complementarity problems. We establish conditions ensuring the nonsingularity of the Jacobian matrix associated with our algorithms, which is particularly beneficial in the realm of Newton's method. In Chapter 4, we extend our study to the inverse problem, where a set of distinct real numbers is provided, and the task is to identify a matrix attaining these reals as Pareto eigenvalues. The exploration of these topics not only contributes to advancing the understanding of Pareto eigenvalue complementarity problems but also sheds light on novel methodologies for addressing such challenges within the realm of mathematical optimization.

Continuing with the discussion on the second subject, focusing on first-order optimization from the perspective of dynamical systems, Chapter 5 delves into the convergence properties of various proximal gradient inertial algorithms. These algorithms are discretized from a non-regular dynamical system described by:

$$\ddot{x}(t) + \gamma\dot{x}(t) + \partial\varphi\Big(\dot{x}(t) + \beta\nabla f(x(t))\Big) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) \ni 0,$$

which incorporates elements of dry friction, viscous damping, and Hessian-driven damping. Notably, the unique feature of this dynamic lies in the novel form of the dry friction term, where the action of dry friction is on a weighted sum of the velocity and the gradient, as opposed to solely the velocity. This innovative formulation enables our algorithms to converge to the exact critical point of the function being minimized, thus surpassing previous methodologies. Theoretical analyses underscore the robustness of our algorithms against perturbations or errors, a characteristic substantiated by numerical tests.
In Chapter 6, we extend our exploration to a doubly nonlinear evolution equation featuring two potentials:

$$\gamma(\dot{x}(t) + \beta\nabla f(x(t))) + \partial\varphi\big(\dot{x}(t) + \beta\nabla f(x(t))\big) + \nabla f(x(t)) \ni 0.$$

This formulation represents a generalization of the gradient flow, incorporating the presence of dry friction. Our investigation initially focuses on studying the convergence properties of this dynamic. Subsequently, we leverage the generic acceleration technique pioneered by Attouch, Bot, and Nguyen, known as time scaling and averaging. This technique enables the transformation of the original evolution equation into an inertial dynamic featuring dry friction and an implicit Hessian-driven damping term. Notably, our analysis demonstrates that this dynamic exhibits accelerated convergence properties. Furthermore, numerical experiments underscore the superior performance of inertial systems over their first-order counterparts.

There are several open questions and potential avenues for advancement stemming from these works.
For Chapter 3, further investigation into alternative interior point methods presents an intriguing prospect. Comparing their performances against the MPCM, NPIPM, and SM methodologies in the context of eigenvalue complementarity problems could provide valuable insights into the relative strengths and weaknesses of different approaches. Moreover, the exploration of inexact Newton methods within our framework holds promise for resolving large-scale problems more efficiently. Also, exploring the possibility of incorporating global convergence techniques into our algorithms represents a promising area of research.
For Chapter 5, a notable area of interest centers around the exploration of algorithms for general composite optimization problems, as well as their related stochastic variants. The idea of studying stochastic variants is suggested by the robustness of our proposed algorithms against perturbations and errors. It would also be interesting to extend the analyses for the following dynamic

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \partial\varphi\Big(\dot{x}(t) + \beta\nabla f(x(t))\Big) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) \ni 0,$$

where we replace the fixed vicious damping coefficient by an asymptotic vanishing term of the form $\frac{\alpha}{t}$. This concept is inspired by the work of Su, Boyd, and Candès [143] where they introduce a continuous representation of the Nesterov accelerated gradient method using this type of viscous damping. By incorporating this principle, we aim to explore the potential of deriving new dynamics that exhibit accelerated convergence properties.
As for Chapter 6, several intriguing questions arise. Firstly, there is a need to explore scenarios where the objective function f is a nonsmooth nonconvex function. This exploration could shed light on the behavior of the dynamical system in optimizing functions characterized by nonsmoothness and nonconvexity. Another question is how to design from (DRYAD) the associated first-order dynamic representing the minimization problem of additive functions where f is the sum of a smooth and a nonsmooth function. Last but not least, understanding how to discretize the resulting inertial systems to achieve fast optimization algorithms is of paramount importance. This is however not trivial and likely to require extensive effort. Exploring these open questions and directions for progression promises to advance the field of optimization and dynamical systems, paving the way for innovative techniques and solutions to challenging optimization problems.

8

Bibliography

Contents

References

[1] B. Abbas, H. Attouch, B.F. Svaiter, *Newton-like dynamics and forward–backward methods for structured monotone inclusions in Hilbert spaces*, J. Optim. Theory Appl., 161(2) (2014), pp. 331–360.

[2] S. Adly, *A variational approach to nonsmooth dynamics: applications in unilateral mechanics and electronics*, Springer Briefs in Mathematics, 2017.

[3] S. Adly, H. Attouch, *Finite convergence of proximal-gradient inertial algorithms combining dry friction with Hessian-driven damping*, SIAM J. Optim., 30(3) (2020), pp. 2134–2162.

[4] S. Adly, H. Attouch, *Finite time stabilization of continuous inertial dynamics combining dry friction with Hessian-driven damping*, J. Conv. Anal., 28(2) (2021), pp. 281–310.

[5] S. Adly, H. Attouch, *First-order inertial algorithms involving dry friction damping*, Math. Program., (2021), Ser. A, 193(1) (2022), pp. 405–445.

[6] S. Adly, H. Attouch, *Accelerated dynamics with dry friction via time scaling and averaging of doubly nonlinear evolution equations,* Nonlinear Anal. Hybrid Syst., 50 (2023).

[7] S. Adly, H. Attouch, A. Cabot, *Finite time stabilization of nonlinear oscillators subject to dry friction*, in Nonsmooth Mechanics and Analysis, Adv. Mech. Math. 12 (2006), Springer, New York, pp. 289–304.

[8] S. Adly, H. Attouch, M.H. Le, *First order inertial optimization algorithms with threshold effects associated with dry friction*, Comput. Optim. Appl., 86 (2023), pp. 801–843.

[9] S. Adly, H. Attouch, M.H. Le, *A doubly nonlinear evolution system with threshold effects associated with dry friction* (to appear in JOTA).

[10] S. Adly, B. Brogliato, B.K. Le, *Well-posednesss, robustness and stability analysis of a set-valued controller for Lagrangian systems*, SIAM J. Control Optim., 51 (2013), pp. 1592–1614.

[11] S. Adly, D. Goeleven, *A stability theory for second order nonsmooth dynamical systems with application to friction problems*, Journal de Mathématiques Pures et Appliquées, 83 (2004), pp. 17–51.

[12] S. Adly, D. Goeleven, *A nonsmooth approach for the modeling of a mechanical rotary drilling system with friction*, Evol. Equ. Control Theory, (2020), pp. 915–934.

[13] S. Adly, M. Haddou, M.H. Le, *Interior point methods for solving Pareto eigenvalue complementarity problems*, Optim. Methods Softw., 38(3) (2023), pp 543–569.

[14] S. Adly, M.H. Le, *Solving inverse Pareto eigenvalue problems*, Optim. Lett., 17(4) (2023), pp. 829–849.

[15] S. Adly, H. Rammal, *A new method for solving Pareto eigenvalue complementarity problems*, Comput. Optim. Appl., 55 (2013), pp. 703–731.

[16] S. Adly, H. Rammal, *A New Method for Solving Second-Order Cone Eigenvalue Complementarity Problems*, J. Optim. Theory Appl., 165 (2015), pp. 563–585.

[17] S. Adly, A. Seeger, *A nonsmooth algorithm for cone-constrained eigenvalue problems*, Comput. Optim. Appl., 49 (2011), pp. 299–318.

[18] L.X. Ahn, *Dynamics of mechanical systems with Coulomb friction*, Springer-Verlag, Berlin, 2003.

[19] C.D. Alecsa, S. László, T. Pinta, *An extension of the second order dynamical system that models Nesterov's convex gradient method*, Appl. Math. Optim., (2020), https://doi.org/10.1007/s00245-020-09692-1.

[20] F. Alvarez, *On the minimizing property of a second-order dissipative system in Hilbert spaces*, SIAM J. Control Optim., 38(4) (2000), pp. 1102–1119.

[21] F. Álvarez, H. Attouch, J. Bolte, P. Redont, *A second-order gradient-like dissipative dynamical system with Hessian-driven damping*, J. Math. Pures Appl., 81 (2002), pp. 747–779.

[22] H. Amann, J. I. Díaz, *A note on the dynamics of an oscillator in the presence of strong friction*, Nonlinear Anal., 55 (2003), pp. 209–216.

[23] G. Amontons, *On the Resistance Originating in Machines*, Proceedings of the French Royal Academy of Sciences, (1699), pp. 206–222.

[24] V. Apidopoulos, J.-F. Aujol, Ch. Dossal, *Convergence rate of inertial Forward-Backward algorithm beyond Nesterov's rule*, Math. Program., 180 (2020), pp. 137–156.

[25] H. Attouch, J. Bolte, P. Redont, A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Lojasiewicz inequality*, Math. Oper. Res., 35(2) (2010), pp. 438–457.

[26] H. Attouch, J. Bolte, B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, regularized Gauss-Seidel methods*, Math. Program., 137(1) (2013), pp. 91–129.

[27] H. Attouch, R.I. Boţ, E.R. Csetnek, *Fast optimization via inertial dynamics with closed-loop damping*, J. Eur. Math. Soc., 25(5) (2023), pp. 1985–2056.

[28] H. Attouch, R.I. Bot, D.-K. Nguyen, *Fast convex optimization via time scale and averaging of the steepest descent,* arXiv:2208.08260v2 [math.OC] 30 Aug 2022.

[29] H. Attouch, G. Buttazzo, G. Michaille, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*, 2nd ed., MOS/SIAM Ser. Optim. 17, SIAM, Philadelphia, 2014.

[30] H. Attouch, A. Cabot, *Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity*, J. Differential Equations, 263 (2017), pp. 5412–5458.

[31] H. Attouch, A. Cabot, *Convergence rates of inertial forward-backward algorithms*, SIAM J. Optim., 28(1) (2018), pp. 849–874.

[32] H. Attouch, A. Cabot, Z. Chbani, H. Riahi, *Rate of convergence of inertial gradient dynamics with time-dependent viscous damping coefficient*, Evol. Equ. Control Theory, 7(3) (2018), pp. 353–371.

[33] H. Attouch, A. Cabot, Z. Chbani, H. Riahi, *Accelerated forward-backward algorithms with perturbations. Application to Tikhonov regularization*, JOTA, 179(1) (2018), pp. 1–36 .

[34] H. Attouch, Z. Chbani, J. Fadili, H. Riahi, *First-order optimization algorithms via inertial systems with Hessian driven damping*, Math. Program., 193 (2020), pp. 113–155.

[35] H. Attouch, Z. Chbani, J. Peypouquet, P. Redont, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity*, Math. Program. Ser. B 168 (2018), pp. 123–175.

[36] H. Attouch, Z. Chbani, H. Riahi, *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case* $\alpha \leq 3$, ESAIM Control Optim. Calc. Var., 25 (2019), pp. 1–34.

[37] H. Attouch, Z. Chbani, H. Riahi, *Fast proximal methods via time scaling of damped inertial dynamics*, SIAM J. Optim., 29(3) (2019), pp. 2227–2256.

[38] H. Attouch, J. Fadili, V. Kungurtsev, *On the effect of perturbations, errors in first-order optimization methods with inertia and Hessian driven damping*, Evol. Equ. Control Theory, 12 (2023), pp. 71–117.

[39] H. Attouch, X. Goudou, P. Redont, *The heavy ball with friction method. The continuous dynamical system, global exploration of the local minima of a real-valued function by asymptotical analysis of a dissipative dynamical system*, Commun. Contemp. Math., 2(1) (2000), pp. 1–34.

[40] H. Attouch, P.E. Maingé, P. Redont, *A second-order differential system with Hessian-driven damping; Application to non-elastic shock laws*, Differ. Equ. and Appl., 4(1) (2012), pp. 27–65.

[41] H. Attouch, J. Peypouquet, *The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than* $1/k^2$, SIAM J. Optim., 26(3) (2016), pp. 1824–1834.

[42] H. Attouch, J. Peypouquet, P. Redont, *Fast convex minimization via inertial dynamics with Hessian driven damping*, J. Differential Equations, 261 (2016), pp. 5734–5783.

[43] H. Attouch, M. Théra, *A general duality principle for the sum of two operators*, J. Convex Anal., 3 (1996), pp. 1–24.

[44] J.-F. Aujol, Ch. Dossal, *Stability of over-relaxations for the Forward–Backward algorithm, application to FISTA*, SIAM J. Optim., 25 (2015), pp. 2408–2433.

[45] J.-F. Aujol, Ch. Dossal, *Optimal rate of convergence of an ODE associated to the Fast Gradient Descent schemes for $b > 0$*, 2017, https://hal.inria.fr/hal-01547251v2.

[46] J.-F. Aujol, Ch. Dossal, G. Fort, E. Moulines, *Rates of Convergence of Perturbed FISTA-based algorithms*, 2019, hal-02182949.

[47] J.-F. Aujol, Ch. Dossal, A. Rondepierre, *Convergence rates of the Heavy-Ball method for quasi-strongly convex optimization*, 2021, hal-02545245v2.

[48] F. Bach, *Statistical machine learning and convex optimization*, StatMathAppli 2017, Fréjus - September 2017.

[49] B. Baji, A. Cabot, *An inertial proximal algorithm with dry friction: finite convergence results*, Set-Valued Anal., 9(1) (2006), pp. 1–23.

[50] M. Balti, R. May, *Asymptotic for the perturbed heavy ball system with vanishing damping term*, Evol. Equ. Control Theory, 6 (2017), pp. 177–186.

[51] H. Bauschke, P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert spaces*, CMS Books in Math., Springer, New York, 2011.

[52] A. Beck, M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[53] Y. Bello-Cruz, M. L. N. Gonalves, N. Krislock, *On inexact accelerated proximal gradient methods with relative error rules*, preprint arXiv:2005.03766, (2020).

[54] R. F. Boisvert, R. Pozo, K. Remington, R. F. Barrett, J. J. Dongarra, *Matrix market: a web resource for test matrix collections*. In: The Quality of Numerical Software: Assessment and Enhancement, pp. 125–137. Chapman and Hall, London (1997).

[55] J. Bolte, A. Daniilidis, O. Ley, L. Mazet, *Characterizations of Lojasiewicz inequalities: subgradient flows, talweg, convexity*, Trans. Amer. Math. Soc., 362 (6) (2010), pp. 3319–3363.

[56] J. Bolte, S. Sabach, M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program., 146 (1-2) (2014), pp. 459-494.

[57] F. Bonnans, *Optimisation continue: cours et problèmes corrigés*, Mathématiques appliquées pour le Master, Dunod, 2006.

[58] R.I. Boţ, E. R. Csetnek, *Second order forward-backward dynamical systems for monotone inclusion problems*, SIAM J. Control Optim., 54(3) (2016), pp. 1423-1443.

[59] R. I. Boţ, E. R. Csetnek, S.C. László, *A second order dynamical approach with variable damping to nonconvex smooth minimization*, Appl. Anal., 99(3) (2020), pp. 361–378.

[60] R.I. Boţ, E. R. Csetnek, S.C. Laszló, *Tikhonov regularization of a second order dynamical system with Hessian damping*, Math. Program., 189 (2021), pp. 151–186.

[61] C. P. Brás, , J. J. Júdice, H. D. Sherali, *On the solution of the inverse eigenvalue complementarity problem*, J. Optim. Theory Appl., 162 (2014), pp. 88–106.

[62] H. Brézis, *Opérateurs maximaux monotones dans les espaces de Hilbert et équations d'évolution*, Lecture Notes 5, North Holland, 1972.

[63] B. Brogliato, *Nonsmooth mechanics. Models, dynamics and control.* Third edition. Communications and Control Engineering Series. Springer, 2016.

[64] C. Castera, J. Bolte, C. Févotte, E. Pauwels, *An inertial Newton algorithm for deep learning*, J. Mach. Learn. Res., 22 (2021), pp. 1–31.

[65] A. Chambolle, Ch. Dossal, *On the convergence of the iterates of the Fast Iterative Shrinkage Thresholding Algorithm*, J. Optim. Theory Appl., 166 (2015), pp. 968–982.

[66] A. Chambolle, T. Pock, *An introduction to continuous optimization for imaging*, Acta Numer., 25 (2016), pp. 161–319.

[67] F. H. Clarke, *Optimization and Nonsmooth Analysis.* Wiley, New York (1983). Reprinted by, SIAM, Philadelphia, PA, 1990.

[68] P. Colli, A. Visintin, *On a class of doubly nonlinear evolution equations*, Comm. Partial Differential Equations, 15(5) (1990), pp. 737–756.

[69] C.A. Coulomb, *Théorie des machines simples, en ayant egard au frottement de leurs parties, et a la roideur dews cordages*, Mem. Math Phys., Paris, X (1785), pp. 161–342.

[70] L. Da Vinci, *The Notebooks*, Ed. Jean Paul Richter, New York: Dover Pub. Inc., 1970.

[71] J. I. Díaz, A. Liñán, *On the asymptotic behavior of a damped oscillator under a sublinear friction term*, Rev. R. Acad. Cien. Serie A. Mat., 95(1) (2001), pp. 155–160.

[72] E. D. Dolan and J. J. Moré, *Benchmarking Optimization Software with Performance Profiles*, Math. Program., 91 (2002), pp. 201–213.

[73] M. C. Ferris, J. S. Pang, *Engineering and economic applications of complementarity problems*, SIAM Rev., 39 (1997), pp. 669–713.

[74] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.

[75] M. Fukushima, J. Júdice, W. Oliveira, V. Sessa, *A sequential partial linearization algorithm for the symmetric eigenvalue complementarity problem*, Comput. Optim. Appl., 77 (2020), pp. 711–728.

[76] P. Gajardo, A. Seeger, *Reconstructing a matrix from a partial sampling of Pareto eigenvalues*, Comput. Optim. Appl., 51 (2012), pp. 1119–1135.

[77] P. Gajardo, A. Seeger, *Solving inverse cone-constrained eigenvalue problems*, Numer. Math., 123 (2013), pp. 309–331.

[78] E. Ghadimi, H. R. Feyzmahdavian, M. Johansson, *Global convergence of the heavy-ball method for convex optimization*, in 2015 European Control Conference, July 2015, pp. 310–315.

[79] G. H. Golub, H. A. Van der Vorst, *Eigenvalue computation in the 20th century*, J. Comput. Appl. Math., 123 (2000), pp. 35–65.

[80] G. Gorni, *Conjugation and second-order properties of convex functions*, J. Math. Anal. Appl., 158 (1991), pp. 293–315.

[81] M. Haddou, P. Maheux , *Smoothing Methods for Nonlinear Complementarity Problems*, J. Optim. Theory Appl., 160 (2014), pp. 711–729.

[82] A. Haraux, M. Ghisi, M. Gambino, *Local and global smoothing effects for some linear hyperbolic equations with a strong dissipation*, Trans. Amer. Math. Soc., 368 (2016), pp. 2039–2079.

[83] A. Haraux, M. A. Jendoubi, *Convergence of solutions of second-order gradient-like systems with analytic nonlinearities*, J. Differential Equations, 144(2) (1998), pp. 313–320.

[84] A. Haraux, M. A. Jendoubi, *The convergence problem for dissipative autonomous systems*, Classical methods and recent advances, January 30, Springer, 2015.

[85] J.S. He, C. Li, J.H. Wang, *Newton's method for underdetermined systems of equations under the γ-condition*, Numer. Funct. Anal. Optim., 28 (2007), pp. 663–679.

[86] J.-B. Hiriat-Urruty, *How to Regularize a Difference of Convex Functions*, J. Math. Anal. Appl., 162 (1991), pp. 196–209.

[87] A.Ioffe, *An invitation to tame optimization*, SIAM J. Optim., 19(4) (2009), pp. 1894–1917.

[88] A. N. Iusem, J. J. Júdice, V. Sessa, P. Sarabando, *Splitting methods for the Eigenvalue Complementarity Problem*, Optim. Methods Softw., 34 (2019), pp. 1184–1212.

[89] M. Jean, J.J. Moreau, *Unilateraly and dry friction in the dynamics of rigid body collections*, Proc. Contact Mechanics Int. Symp., (Ed. Curnier A.), Presses Polytechniques et Universitaires Romandes (1992), pp. 31–48.

[90] J. J. Júdice, M. Raydan, S. S. Rosa, S. A. Santos, *On the solution of the symmetric eigenvalue complementarity problem by the spectral projected gradient algorithm*, Numer. Algor., 47 (2008), pp. 391–407.

[91] J. J. Júdice, H. D. Sherali, I. Ribeiro, *The eigenvalue complementarity problem*, Comput. Optim. Appl., 37 (2007), pp. 139–156.

[92] J. J. Júdice, H. D. Sherali, I.M. Ribeiro, S.S. Rosa, *On the asymmetric eigenvalue complementarity problem*, Optim. Methods Softw., 24 (2009), pp. 549–568.

[93] C. Kanzow, H. Kleinmichel, *A New Class of Semismooth Newton-Type Methods for Nonlinear Complementarity Problems*, Comput. Optim. Appl., 11 (1998), pp. 227–251.

[94] N. Karmarkar, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[95] D. Kim, *Accelerated proximal point method for maximally monotone operators*, Math. Program., (2021).

[96] H.A. Le Thi, M. Moeini, T. Pham Dinh et al, *A DC programming approach for solving the symmetric Eigenvalue Complementarity Problem*, Comput. Optim. Appl., 51 (2012), pp. 1097–1117.

[97] H.A. Le Thi and T. Pham Dinh, *The DC (difference of convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems*, Ann. Oper. Res., 133 (2005), pp. 23–48.

[98] C. Lemarechal, C. Sagastizábal, *Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries.* SIAM J. Optim., 7 (1997), pp. 367–385.

[99] C. E. Lemke, J. T. Howson, *Equilibrium points of bimatrix games*, SIAM J. Appl. Math., 12 (1964), pp. 413–423.

[100] J. Liang, J. Fadili, G. Peyré, *Local linear convergence of forward-backward under partial smoothness*, Advances in Neural Information Processing Systems, 2014, pp. 1970–1978.

[101] T. Lin, M.I. Jordan, *A control-theoretic perspective on optimal high-order optimization*, (2019), arXiv:1912.07168v1.

[102] C. Ling, H. He, L. Qi, *On the cone eigenvalue complementarity problem for higher-order tensors*, Comput. Optim. Appl., 63 (2016), pp. 143–168.

[103] Z. Liu, J. Tang, *A new smoothing-type algorithm for nonlinear weighted complementarity problem*, J. Appl. Math. Comput., 64 (2020), pp. 215–226.

[104] S. Lojasiewicz, *Sur les trajectoires de gradient d'une fonction analytique*, Seminari di Geometria 1982-1983. Universita di Bologna, Dipartemento di Matematica, pp 115–117, 1984.

[105] J. A. C. Martins, S. Barbarin, M. Raous, A. Pinto da Costa, *Dynamic stability of finite dimensional linearly elastic systems with unilateral contact and Coulomb friction*, Comput. Methods Appl. Mech. Eng., 177 (1999), pp. 289–328.

[106] J. A. C. Martins, A. Pinto da Costa, *Stability of finite-dimensional nonlinear elastic systems with unilateral contact and friction*, Int. J. Solids Struct., 37 (2000), pp. 2519–2564.

[107] J. A. C. Martins, A. Pinto da Costa, *Computation of bifurcations and instabilities in some frictional contact problems*, in: European Conference on Computational Mechanics: ECCM (2001).

[108] J. A. C. Martins, A. Pinto da Costa, *Bifurcations and instabilities in frictional contact problems: theoretical relations, computational methods and numerical results*, in: European Congress on Computational Methods In Applied Sciences and Engineering: ECCOMAS (2004).

[109] R. May, *Asymptotic for a second-order evolution equation with convex potential and vanishing damping term*, Turkish Journal of Math., 41(3) (2017), pp. 681–685.

[110] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.

[111] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim. 15 (1977), pp. 957–972.

[112] J.J. Moreau, *Unilateral contact and dry friction in finite freedom dynamics*, Non-Smooth Mechanics and Applications, 302 of CISM Courses and Lectures (Eds. Moreau J.J. and Panagiotopoulos P.D.), 1–82, Springer, Wien, 1988.

[113] A. J. Morin, *New Friction Experiments carried out at Metz in 1831-1833*, Proceedings of the French Royal Academy of Sciences, (1833), pp. 1–128.

[114] Y. Nesterov, *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Soviet Math. Dokl., 27 (1983), pp. 372–376.

[115] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Appl. Optim. 87, Kluwer, Boston, MA, 2004.

[116] Y. S. Niu, T. Pham Dinh, H. A. Le Thi, J. J. Júdice, *Efficient DC programming approaches for the asymmetric eigenvalue complementarity problem*, Optim. Methods Softw., 28 (2012), pp. 812–829.

[117] J. M. Ortega, W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variable*, Academic Press, New York, 1970.

[118] P.D. Panagiotopoulos, *Non-Convex Superpotentials in the Sense of F.H. Clarke and Applications*, Mech. Res. Comm., (1981), pp. 335–340.

[119] J.S. Pang, F. Facchinei, *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Operations Research, vol. 2. Springer, New York, 2003.

[120] J. Peypouquet, S. Sorin, *Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time*, J. Convex Anal, 17 (3–4) (2010), pp. 1113–1163.

[121] T. Pham Dinh, E.B. Souad, *Algorithms for solving a class of nonconvex optimization problems. Methods of subgradients*. J.B. Hiriart-Urruty (ed.) Fermat Days 85: Mathematics for Optimization, North-Holland Math. Stud., 129 (1986), pp. 249–271.

[122] A. Pinto da Costa, I. Figueiredo, J. Júdice, J. Martins, *A complementarity eigenproblem in the stability analysis of finite dimensional elastic systems with frictional contact*. In: M. Ferris, J.S. Pang, O. Mangasarian (eds.) Complementarity: Applications, Algorithms and Extensions, pp. 67–83. Kluwer, New York (2001).

[123] A. Pinto da Costa, J. A. C. Martins , I. N. Figueiredo and J. J. Júdice, *The directional instability problem in systems with frictional contacts*, Comput. Methods Appl. Mech. Eng., 193 (2004), pp. 357–384.

[124] A. Pinto da Costa, A. Seeger, *Cone-constrained eigenvalue problems: theory and algorithms*, Comput. Optim. Appl., 45 (2010), pp. 25–57.

[125] B.T. Polyak, *Some methods of speeding up the convergence of iterative methods*, Z. Vylist Math. Fiz., 4 (1964), pp. 1–17.

[126] B.T. Polyak, *Introduction to optimization*. New York: Optimization Software. (1987).

[127] F. A. Potra, Y. Ye, *Interior-point methods for nonlinear complementarity problems*, J. Optim. Theory Appl., 88 (1996), pp. 617–642.

[128] L. Qi, J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.

[129] M. Queiroz, J. J. Júdice, C. Humes, *The symmetric eigenvalue complementarity problem*, Math. Comput., 73 (2004), pp. 1849–1863.

[130] O. Reynolds, *On the Theory of Lubrication and its application to Mr. Beauchamp Tower's experiments, including an experimental determination of the viscosity of olive oil*, Phil. Trans. Royal Soc., vol., (1886), pp.157-234.

[131] R.T. Rockafellar, R. Wets, *Variational analysis*, Springer, Berlin, 1998.

[132] R. Rossi, A. Mielke, G. Savaré, *A metric approach to a class of doubly nonlinear evolution equations and applications*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (5), (2008), pp. 97–169.

[133] M. Salahi, J. Peng, T. Terlaky, *On Mehrotra-Type Predictor-Corrector Algorithms*, SIAM J. Optim., 18 (2008), pp. 1377–1397.

[134] M. Schmidt, N. Le Roux, F. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*. In: NIPS'11—25, (2011), Granada. HAL inria-00618152v3.

[135] A. Seeger, *Eigenvalue Analysis of Equilibrium Processes Defined by Linear Complementarity Conditions*, Linear Algebra Appl., 292 (1999), pp. 1–14.

[136] A. Seeger, J. Vicente-Pérez, *On cardinality of Pareto spectra*, Electron. J. Linear Algebra, 22 (2011), pp. 758–766.

[137] A. Seeger, J. Vicente-Pérez, *Inverse eigenvalue problems for linear complementarity systems*, Linear Algebra App., 435 (2011), pp. 3029–3044.

[138] B. Shi, S. S. Du, M. I. Jordan, W. J. Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, Math. Program., 195(1) (2022), pp. 79–148.

[139] J. W. Siegel, *Accelerated first-order methods: Differential equations and Lyapunov functions*, arXiv: Optimization and Control (math.OC):1903.05671v1 (2019).

[140] M.V. Solodov, S.K. Zavriev, *Error stability properties of generalized gradient-type algorithms*, J. Optim. Theory Appl., 98 (1998), pp. 663—680.

[141] L. Song, Y. Gao, *A smoothing Levenberg-Marquardt method for nonlinear complementarity problems*, Numer. Algor., 79 (2018), pp. 1305–1321.

[142] R. Stribeck, *Die Wesentlichen Eigenschaften der Gleit-und Rollenlager*, Z. Verein. Deut. Ing. 46(38) (1902), pp. 1341–1348.

[143] W. Su, S. Boyd, E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method*, J. Mach. Learn. Res., 17 (2016), pp. 1–43.

[144] J. Toland, *Duality in nonconvex optimization*, J. Math. Anal. Appl., 66 (1978), pp. 399–415.

[145] L. van den Dries, *Tame Topology and o-Minimal Structures*, London Mathematical Society, Lecture Note Series, Vol. 248., Cambridge University Press, Cambridge, UK, (1998).

[146] H.A. Van der Vorst, G.H. Golub, *150 years old and still alive: Eigenproblems.* In: The State of the Art in Numerical Analysis, Institute of Mathematics and its Applications, vol. 52, pp. 93–119. Oxford University Press, New York (1997).

[147] S. Villa, S. Salzo, L. Baldassarres, A. Verri, *Accelerated and inexact forward–backward*, SIAM J. Optim., 23 (2013), pp. 1607—1633.

[148] D.T.S. Vu, I. Ben Gharbia, M. Haddou, Q. H. Tran, *A new approach for solving nonlinear algebraic systems with complementarity conditions. Application to compositional multiphase equilibrium problems*, Math. Comput. Simulation, (2021), pp. 1243–1274.

[149] H.F. Walker, L.T. Watson, *Least-change secant update methods for underdetermined systems*, SIAM J. Numer. Anal., 27 (1990), pp. 1227–1262.

[150] Z. Zhou, Y. Peng, *The locally Chen–Harker–Kanzow–Smale smoothing functions for mixed complementarity problems*, J. Glob. Optim., 74 (2019), pp. 169–193.

[151] J. Zhu, B. Hao, *A new smoothing method for solving nonlinear complementarity problems*, Open Math., 17 (2019), pp. 104–119.

Publications

[1] S. Adly, M. Haddou, M.H. Le, *Interior point methods for solving Pareto eigenvalue complementarity problems*, Optim. Methods Softw. 38(3) (2023), pp. 543–569.

[2] S. Adly, M.H. Le, *Solving inverse Pareto eigenvalue problems*, Optim. Lett. 17(4) (2023), pp. 829–849.

[3] S. Adly, H. Attouch, M.H. Le, *First order inertial optimization algorithms with threshold effects associated with dry friction*, Comput. Optim. Appl., 86 (2023), pp. 801–843.

[4] S. Adly, H. Attouch, M.H. Le, *A doubly nonlinear evolution system with threshold effects associated with dry friction*, 2023 (to appear in JOTA).

Études mathématiques et numériques de la complémentarité aux valeurs propres et des problèmes d'accélération dans l'optimisation du premier ordre

Résumé : Dans cette thèse, j'explore deux sujets clés. Premièrement, je m'intéresse à l'étude mathématique et numérique du problème de complémentarité des valeurs propres de Pareto et de sa contrepartie inverse. Notre approche utilise des méthodes de points intérieurs, complétées par une technique de lissage non paramétrique. L'efficacité des méthodologies proposées est soulignée par un ensemble d'expériences numériques. En mettant l'accent sur l'optimisation continue, nous adoptons une perspective de systèmes dynamiques. Plus précisément, nous étudions divers algorithmes inertiels à gradient proximal, discrétisés à partir d'un système dynamique inertiel non régulier comportant des éléments de frottement sec et d'amortissement piloté par le Hessien. En outre, nous examinons une équation d'évolution doublement non linéaire régie par deux potentiels, ainsi que l'accélération de sa convergence par l'application de techniques de mise à l'échelle temporelle et de calcul de la moyenne, ce qui se traduit par une dynamique inertielle comportant un frottement sec et un amortissement implicite induit par le hessien. Les tests numériques corroborent la performance supérieure des systèmes inertiels par rapport à leurs homologues du premier ordre, ce qui correspond aux résultats théoriques.

Mots clés : Problèmes de complémentarité, méthodes des points intérieurs, optimisation du premier ordre, optimisation convexe, algorithmes inertiels, systèmes dynamiques.

Mathematical and numerical studies of eigenvalue complementarity problems and acceleration methods in first-order optimization

Abstract: In this thesis, I explore two key topics. Firstly, I delve into the mathematical and numerical study of the Pareto eigenvalue complementarity problem and its inverse counterpart. Our approach employs interior point methods, supplemented by a non-parametric smoothing technique. The efficacy of these proposed methodologies is underscored through an array of numerical experiments. Shifting our focus to continuous optimization, we adopt a dynamical systems perspective. Specifically, we study various proximal gradient inertial algorithms, discretized from a non-regular inertial dynamical system featuring elements of dry friction and Hessian-driven damping. Additionally, we examine a doubly nonlinear evolution equation governed by two potentials, and its convergence acceleration through the application of time scaling and averaging techniques, which results in inertial dynamics featuring dry friction and implicit Hessian-driven damping. The numerical tests corroborate the superior performance of inertial systems over their first-order counterparts, aligning with the theoretical results.

Keywords: Complementarity problems, interior point methods, first-order optimization, convex optimization, inertial algorithms, dynamical systems.

Univ. de Limoges, CNRS, XLIM, UMR 7252, F-87000 Limoges, France