Stochastic Majoration-Minimization Algorithms

Jean-Baptiste Fest

université PARIS-SACLAY

Algorithmes de Majoration-Minimisation Stochastiques

Stochastic Majorization-Minimization Algorithms

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580 Sciences et Technologie de l'Information et de la Communication (STIC)
Spécialité de doctorat: Sciences du Traitement du signal et des images
Graduate School: Sciences de l'ingénierie et des systèmes.
Référent : CentraleSupélec

Thèse préparée dans l'unité de recherche Centre de la Vision Numérique (Université Paris-Saclay, CentraleSupélec), sous la direction de **Émilie Chouzenoux**, Chargée de Recherche, Inria Saclay.

Thèse soutenue à Paris-Saclay, le 03 octobre 2023, par

Jean-Baptiste FEST

Composition du jury

Membres du jury avec voix délibérative

Jérôme MALICK Directeur de Recherche, LJK, CNRS, Université Grenoble Alpes	Rapporteur / Président du jury
Silvia VILLA Professeure, Università degli studi di Genova	Rapporteur
Claire BOYER Maître de Conférences, LPSM, Sorbonne Université	Examinatrice
Gersende FORT Directrice de Recherche, CNRS, Institut Mathématique de Toulouse	Examinatrice

THESE DE DOCTORAT

NNT : 2023UPAST134

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Algorithmes de Majoration-Minimisation stochastiques

Mots clés: Optimisation Stochastique et Non Convexe, Algorithmes MM, Problèmes Inverses, Apprentissage

Résumé: De nombreux problèmes rencontrés en optimisation différentiable sont mathématiquement associés à des fonctions de coût dont la minimisation requiert de travailler sur un espace de dimension particulièrement élevée. Certaines opérations numériques sont alors impossibles à réaliser et ainsi, de par l'occupation mémoire que nécessite leur mise en œuvre, l'utilisation de certains algorithmes, pourtant réputés efficaces, devient inenvisageable. Par ailleurs, en étant uniquement le fruit d'un modèle d'observation et/ou de connaissance, la fonction de coût en elle-même peut ne pas rendre pleinement compte des phénomènes physiques à décrire. Un traitement purement déterministe des données est alors insuffisant et l'élaboration de méthodes probabilistes devient indispensable. En réponse à cette double problématique, le travail de thèse présenté a pour objectif de construire des approches d'optimisation pour la résolution de problèmes de grandes tailles posés aussi bien dans les domaines du traitement d'images ou de l'apprentissage statistique. Notre point de départ s'articule autour d'une classe de méthode particulière, fondée sur le principe de Majoration-Minimisation (MM), réputée pour la robustesse des schémas numériques qu'elle est susceptible d'engendrer indépendamment de la convexité (ou non) du problème. Les contributions de ce travail de thèse sont fondées sur deux axes d'analyse. Une première partie s'attache à concevoir un nouveau schéma MM, quadratique, pour manipuler des données à grande échelle, en permettant idéalement d'exploiter pleinement les capacités intrinsèques des outils de calculs modernes. D'un point de vue théorique, nous établissons les propriétés de convergence associées à ce nouvel algorithme MM quadratique distribué sous des hypothèses raisonnables. Dans un second temps, nous proposons une extension stochastique de ce dernier en supposant inexact l'accès à certaines informations relatives à la fonction de coût, en particulier l'évaluation de son gradient. Les méthodes d'analyse asymptotiques nécessitent alors la mise ne place d'outils théoriques originaux. En particulier, l'obtention de certaines garanties asymptotiques, dans le cas non convexe, suppose la mise en évidence de résultats inédits qu'il devient indispensable d'étudier en détails.

Dans ce contexte, la dernière partie de ce travail de thèse se détache du cadre MM et est dédiée au développement d'une nouvelle méthodologie pour le raffinement de certains résultats de convergence autour des schémas stochastiques et toujours dans un cadre non nécessairement convexe. Nous nous appuyons plus spécifiquement sur les récents développements autour de la théorie de Kurdyka-Łojasiewicz (KL) pour l'optimisation déterministe. L'objectif en résultant est finalement de proposer une transposition de ces derniers dans le domaine stochastique à des fins d'application sur le plus grand nombre d'algorithmes possibles.

Title: Stochastic Majorization-Minimization Algorithms
Keywords: Stochastic and Non-Convex Optimization, MM Algorithms, Inverse Problems, Machine Learning

Abstract: Many problems encountered in differentiable optimization are mathematically associated with cost functions whose minimization requires working in a particularly high-dimensional space. Certain numerical operations are therefore impossible to carry out and, because of the amount of memory required to implement them, the use of certain algorithms, even though they are reputed to be efficient, becomes unthinkable. Furthermore, by being solely the result of an observation and/or knowledge model, the cost function in itself cannot fully account for the physical phenomena to be described. Purely deterministic data processing is therefore insufficient, and the development of probabilistic methods becomes essential. In response to this twofold problem, the long-term aim of this thesis work is to build a stochastic algorithm for solving large-scale problems in the fields of image processing and statistical learning.

Our starting point revolves around a particular class of method, based on the principle of Majorization-Minimization (MM), reputed for the robustness of the numerical schemes regardless of the convexity (or not) of the problem. Our contribution lies on two lines of analysis. The first part aims to design a new MM scheme, more specifically a quadratic scheme, to manipulate large-scale data, ideally enabling the intrinsic capabilities of modern computation tools. From a theoretical point of view, we establish the convergence properties associated with the proposed distributed quadratic MM algorithm under mild assumptions. Secondly, we propose a so-called stochastic extension of the latter by assuming that the access to certain information relating to the cost function, in particular the evaluation of its gradient, is prone to errors. We present an asymptotic analysis of the algorithm, by relying on theoretical tools relatively new in the probabilistic field. Obtaining deeper asymptotic guarantees supposes the demonstration of novel results that need to be studied in depth.

In this context, the last part of this thesis goes beyond the MM framework and is dedicated to the development of a new methodology for the strengthening of convergence results of stochastic schemes in general and in a non-necessarily convex framework. More specifically, we draw on recent developments in Kurdyka-Łojasiewicz (KL) theory for deterministic optimization. The resulting objective is finally to propose a transposition of the latter into the stochastic domain for application to the largest possible number of algorithms.

Remerciements

L'aboutissement de ce travail est le fruit d'une expérience professionnelle faite de rencontres, de partages et bien sûr d'amitiés. Aussi, je souhaite exprimer, à travers ces quelques lignes, la plus sincère gratitude à l'ensemble des personnes qui m'ont accompagné au cours de ces trois belles années d'odyssée scientifique.

Mes premiers remerciements s'adressent à mon encadrante et instigatrice de ce sujet de thèse, Émilie Chouzenoux. Je te suis tout d'abord reconnaissant pour m'avoir placé dans de rassurantes conditions en vue d'aborder ce travail de longue haleine, en particulier au cours des premiers mois quand bien même cette fichue pandémie nous initiait aux joies du distanciel. Malgré mes retours tardifs et parfois lacunaires, tu as continué à m'accorder ta confiance, pour mener au mieux les différents projets que tu souhaitais compléter. Certains de mes travaux n'auraient sans doute pas vu le jour sans le soutien appuyé d'Audrey Repetti. Aussi, je tiens à t'exprimer toute ma reconnaissance. Mon séjour à Edimbourg, parmi ton équipe, m'ont été extrêmement formateurs et ces quelques semaines ont constitué un véritable dépaysement. Au-delà-même de l'aspect purement scientifique, j'ai été touché par ta bienveillance au cours de nos nombreux échanges. Je tiens également à remercier l'ensemble des membres du jury, ayant consacré une partie de leur temps précieux à la lecture de ce manuscrit puis à l'évaluation de cette thèse Claire Boyer, Gersende Fort, Silvia Villa et Jérôme Malick. J'adresse, par ailleurs, une pensée toute particulière à Saïd Moussaoui, mon responsable d'option de spécialisation à l'école Centrale de Nantes, ayant pris la peine de m'orienter au cours de ma recherche de thèse et sans qui rien de tout cela n'aurait peut-être été possible.

J'adresse bien évidemment un grand Merci à l'ensemble des membres du CVN que j'ai eu la chance de rencontrer et de fréquenter tout au long de ces années. Je tiens en premier lieu, à remercier, Jana Dutrey ainsi que Céline Leroux, pour leur gentillesse et leur efficacité à gérer une grande partie des soucis administratifs auxquels j'ai été confrontés. Je salue également Jean-Christophe Pesquet, directeur du CVN, m'ayant accepté au sein de son équipe en vue de faire partie du projet scientifique OPIS. Bien évidemment, il m'est impossible de ne pas évoquer tous mes camarades et amis de l'Open-Space, les aînés Kavya, Sagar, Ana, Jhony, mes compagnons avec qui cette aventure a débuté, Marion, Mario, Gabriele, Théodore, Younes, Alexandre, Loïc, Mathieu, Simona, Claire, la jeune génération, Aymen, Aya, Clément, Alix ainsi que toutes les autres personnes avec qui j'ai eu la chance d'échanger. Une petite dédicace à Ségolène, ma complice de bureau, et avec qui j'ai partagé nombre de péripéties à Rome puis dans les Alpes suisses. Enfin, mention spéciale au camarade et entreprenant Thomas, pour toutes nos discussions et ses délicieux jus de pommes !

Je ne peux également oublier mes partenaires d'entraînement et surtout amis de l'US Métro, devenus indispensables à mon équilibre de vie.

Il m'est finalement impossible d'achever ces modestes mots sans évoquer ma famille, mes amis proches et de toujours. Je leur dois énormément, si ce n'est tout, et ce paragraphe, à lui-seul, ne suffirait à leur rendre véritablement hommage. Du fond du cœur, Merci pour tout ce que vous avez fait et été pour moi, ce travail est également vôtre.

Bercé alors par la fougue de l'adolescence,
Il était une fois un jeune homme de dix-huit ans,
Qui pour contempler le grand et bel Océan,
A voulu grimper le plus haut sommet de France.
A ses pieds, vêtu de deux simples petites toiles,
Toi seul Papy, pouvait caresser les étoiles.

Résumé français

Ce travail de thèse, mené dans le cadre du projet européen ERC MAJORIS, sous la direction de Emilie Chouzenoux, s'articule autour de l'optimisation différentiable et est dédié, sur le long-terme, à la résolution de problèmes de grande taille pour le traitement de données (e.g. signal, image).

Une majorité de phénomènes physiques se traduit mathématiquement par la minimisation, c'est à dire la recherche d'un point de minimum, d'une fonction de coût particulière, découlant d'un modèle d'observation ou d'approximation. Par exemple, en mécanique Newtonienne, les positions d'équilibre d'un objet, dans un référentiel donné, s'interprètent typiquement comme les extrema d'une énergie potentielle. Le tracé d'une droite de régression linéaire est, de même, caractérisé par un coefficient directeur et une ordonnée à l'origine rendant la plus faible possible une certaine erreur, issue de la donnée des points expérimentaux. Quand bien même l'établissement de la fonction de coût associée au problème étudié résulte généralement de choix ne dépendant pas du numéricien, la minimisation de cette dernière est une tâche qui lui est, à l'inverse, pleinement confiée.

Deux stratégies de résolutions peuvent alors être distinguées. La première, plus naturelle, consiste à exhiber une expression analytique, sous réserve d'existence, d'un point minimum, par l'intermédiaire d'outils purement analytiques (dérivation, résolution d'équations à la main, etc.). Toutefois, de telles opérations supposent de travailler exclusivement avec une fonction de coût possédant une structure relativement simple. En pratique, la complexité, toujours croissante, des modèles de connaissances ne permet pas l'applicabilité de cette approche pour l'immense majorité des problèmes d'optimisation rencontrés aujourd'hui. La seconde, par opposition, réside dans la construction d'un procédé numérique spécifique, un algorithme, visant à approcher au mieux, en un certain sens mathématique, une solution, c'est à dire un point de minimum de la fonction étudiée.

Dans ce contexte, le travail que nous présentons dans ce manuscrit propose le développement de stratégies itératives pour la résolution de problème d'optimisation non contraints (sur un espace de Hilbert tout entier). La dimension de l'espace de recherche est également supposée suffisamment importante pour rendre difficilement envisageable le recours à des opérations usuelles de type inversion ou même produits matriciels, par les outils de calculs employés. Le problème d'optimisation est alors dit de grande taille ou grande échelle. Par ailleurs, nous nous plaçons dans une situation, en aval, pour laquelle, nous n'interférons nullement dans les choix de modélisation ; la fonction de coût du problème est supposée donnée et immuable. Nous lui attribuons cependant un caractère continûment différentiable et, de fait, une régularité suffisante, pour permettre l'utilisation de méthodes de type descente de gradient (Chapitre 2). La relative simplicité de ces dernières et leur interprétation géométrique, particulièrement intuitive, ont historiquement permis leur développement dès le dix-huitième siècle. Aujourd'hui encore de telles approches sont privilégiées pour traiter l'immense majorité des problèmes d'optimisation rencontrés.

Bien que ce travail de thèse soit principalement fondé sur la construction de stratégies de résolution de type descente de gradient, il présente la spécificité d'adopter un point de vue dit quadratique. Notre constat étant le suivant ; un algorithme de descente de gradient résulte de la minimisation, sur un espace donné, d'une fonction quadratique, approchant et se substituant à la fonction de coût, à l'itéré courant. Nous nous intéressons plus spécifiquement à la classe d'approximation reposant sur le

principe dit de Majoration-Minimisation (MM) ; la fonctionnelle quadratique, à minimiser, est alors construite de façon à majorer la fonction de coût initiale tout en coïncidant avec cette dernière au point d'intérêt. Notre choix est motivé par les garanties théoriques en résultant et mise en évidence au cours de ces vingt dernières années. En effet, de par sa construction, un schéma de type MM induit nécessairement une décroissance des évaluations de la fonction de coût associées aux itérés. L'algorithme en résultant possède ainsi des propriétés naturelles de robustesse, lui offrant certaines garanties de convergence, quand bien même cette fonction de coût possède une courbure au comportement particulièrement capricieux, sous-entendu non nécessairement convexe (Chapitre 3). Par ailleurs, la forme quadratique de la fonction d'approximation permet, en complément, la construction de règles de mise à jour structurellement simples facilitant l'incorporation de méthodes d'accélération de type sous-espace.

Toutefois, dans un cadre d'optimisation grande-échelle, de telles stratégies d'accélération à elles-seules ne suffisent pas à véritablement compenser les limitations générées par la dimension du problème, rendant toujours difficile, en machine, le stockage en mémoire de la plupart objets à manipuler. La nature des obstacles auxquels nous sommes confrontés nous incite ainsi à concevoir un schéma MM quadratique, plus souple, notamment à des fins d'implémentation parallèle. Plusieurs travaux, antérieurs à ce projet de thèse, ont en particulier proposé une nouvelle structure de gestion des données de type bloc distribuée ; une étape de mise à jour repose alors uniquement sur l'actualisation, via une procédure quadratique MM, avec accélération de sous-espace, d'un petit nombre de coordonnées. Le schéma en résultant, nommé BP3MG, possède ainsi une complexité grandement réduite, à itération fixé, par rapport à son homologue MM quadratique et présente l'avantage d'en conserver les propriétés de convergence sans ajout d'hypothèse supplémentaire sur la fonction de coût (Chapitre 4). Cependant ce même algorithme ne permet pas d'exploiter pleinement les potentielles capacités de la machine de calcul se chargeant de sa mise en œuvre, en particulier lorsqu'il s'agit d'un serveur multicœurs. De par la gestion cloisonnée et uniquement successive de ses différentes données (i.e. des groupes de coordonnées aux itérés courant), BP3MG ne peut en effet être véritablement implémenté de façon à pleinement exploiter les possibilités offertes par ce type d'architecture.

Le développement d'un algorithme inédit s'est ainsi avéré indispensable pour répondre au mieux à cette problématique. Plus précisément, le schéma conçu, BD3MG, est dérivé de BP3MG dans sa structure, mais possède une règle de mise à jour plus générique, favorisant un traitement asynchrone des données (Chapitre 5). Plusieurs processeurs sont alors mobilisés, travaillent de façon indépendante et effectuent chacun une itération de type MM quadratique sur un groupe de coordonnées spécifique. Ils sont enfin assignés en amont par un homologue commun, dit le maître. Les propriétés asymptotiques de cet algorithme sont étudiées dans un premier temps et toujours dans un cadre non convexe. La méthode d'analyse mise en place, moyennant certaines spécificités induites par la nature asynchrone de la gestion des données, peut en particulier s'interpréter comme une généralisation de celle proposée au cours l'étude de BP3MG. La deuxième partie de ce travail, de nature purement numérique, a vocation à tester les performances de BD3MG pour la résolution d'un problème inverse de type déconvolution, sur des données simulées puis réelles dans un contexte de microscopie biphotonique.

La première partie de ce projet de thèse propose en conséquence des développements pour une amélioration d'ordre structurel du schéma MM quadratique initial. De façon générale, c'est en effet la mise à jour même des itérés qui se voit remodelée afin de tenir compte des capacités intrinsèques des architectures de traitement de données. Un second axe d'analyse, jouant un rôle complémentaire à

celui décrit jusqu'à présent, s'articule directement autour de la fonction de coût en s'interrogeant sur l'exactitude de l'information qu'elle est susceptible de délivrer. D'une part, cette même fonction de coût est la représentation mathématique d'un modèle de connaissance et ne peut donc, à elle seule, suffire à décrire l'ensemble des phénomènes mis en jeu par le problème induit. D'autre part et toujours dans un contexte d'optimisation grande-échelle, certaines évaluations telles celles du gradient, du Hessien (dans le cas deux fois différentiable) voire de la fonction de coût elle-même, sont difficilement envisageables de par la place qu'elles occupent en mémoire machine. En réponse à ces deux limitations, l'utilisation d'un modèle de type "boîte noire" devient légitime ; les phénomènes physiques non identifiés et parfois la donnée exacte de la fonction de coût en résultant se retrouvent masqués par le prisme d'approximations obéissant à des lois probabilistes. Ce type de modélisation a plus particulièrement donné naissance, dans la littérature, au domaine de l'optimisation dite stochastique (Chapitre 7).

La problématique mise en évidence précédemment est finalement à l'origine de ce travail de thèse, à savoir étendre l'algorithme MM quadratique en vue de traiter des problèmes d'optimisation différentiable dans des cadres probabilistes et idéalement toujours non convexes. Face à la multitude des modèles "boîte noire" envisageables, nous considérons ici une incertitude au premier ordre, c'est à dire portant directement sur le gradient de la fonction de coût. Les fonctionnelles quadratiques d'approximation, découlant du procédé MM, s'en voient directement affectées et perdent, en particulier, la propriété de majoration. Le schéma MM quadratique associé, que nous nommons SABRINA, dispose alors d'une mise à jour de type gradient stochastique préconditionné (Chapitre 7) et obéit logiquement à une condition de descente stochastique, synonyme d'une certaine robustesse. Nous en déduisons des propriétés de convergence, certes élémentaires mais attestant toutefois de l'intérêt, au moins théorique, de l'utilisation d'un tel schéma. A des fins d'évaluation de performances, l'algorithme SABRINA est mis en œuvre pour la résolution d'un problème d'apprentissage de type classification binaire puis pour l'identification d'un noyau de bruit sur une image satellite bruitée.

La stratégie employée pour la mise en évidence des propriétés de convergence de SABRINA demeure cependant classique et, d'une certaine façon, les résultats ainsi obtenus ne sont que le reflet d'un algorithme stochastique finalement bien construit. En dehors de toute hypothèse de convexité, relier des comportements élémentaires (e.g. une limite nulle pour les évaluations du gradient et d'une autre finie pour celles de la fonction de coût) avec idéalement la convergence des itérés vers un point stationnaire constitue aujourd'hui un problème ouvert de la littérature stochastique. A l'inverse, de récents outils, utilisant la théorie de Kurdyka Łojasiewicz (KL), ont permis l'émergence d'une nouvelle catégorie de preuves pour démontrer la convergence globale (i.e. en itérés) de plusieurs algorithmes déterministes à partir d'une condition de descente. Ici, les difficultés rencontrées pour leur mise en application sont d'ordre techniques et engendrées par l'ajout du cadre d'étude stochastique que nous nous imposons. Partant de ce constat, nous proposons de nous détacher du schéma MM quadratique afin de développer, de façon plus générale, une méthodologie permettant d'obtenir des convergences de type presque sûr des itérés vers un point stationnaire en partant de propriétés issues typiquement d'une condition de descente conditionnelle (Chapitre 8). Notre objectif consiste plus précisément à étendre les approches initiées en optimisation déterministe pour mieux les adapter aux contraintes imposées par un contexte d'ordre probabiliste.

Une courte dernière partie a vocation à clôturer notre propos et propose quelques pistes de recherche à court et moyen termes associées à plusieurs aspects développés tout au long de ce projet.

Contents

General notations

Here we present some of the mathematical notations we will use throughout this manuscript. Some of them will also be reminded on the fly.

- $(\mathcal{H}, \langle ., . \rangle)$: any real finite dimensional Hilbert space endowed with a scalar product $\langle ., . \rangle$. If $\mathcal{H} = \mathbb{R}^d$ (with$d > 0$), $\langle ., . \rangle$ corresponds to the usual Euclidean scalar product.
- *Italic* style: any deterministic quantity.
- *Italic* and **Bold** style: any deterministic vector (in the sense of an element of $\mathcal{H}$) or linear operator,
- $\mathbf{I}$: the identity application whatever the space,
- $(\Omega, \mathcal{F}, \mathbb{P})$: any generic probability space provided with a σ-algebra $\mathcal{F}$ and a probability measure $\mathbb{P}$,
- Non-italic style: any random quantity,
- Non-italic and **Bold** style: any random vector (in the sense of an element of $\mathcal{H}$) or random linear operator.
- f or F: the cost function to minimize.
- ∇f or ∇F: (under existence) the gradient operator of the cost function to minimize.
- $\boldsymbol{x}_s$: any global minimizer of the cost function, i.e. any solution of the unconstrained optimization problem considered .
- $\boldsymbol{x}^*$: any stationary (or critical) point of the cost function (for which it is not possible to directly conclude whether it is a global minimizer or not).

List of useful accronyms

This list of the acronyms below is not exhaustive, but is intended to identify those that will be developed in this manuscript.

3MG	Majorization-Minimization Memory Gradient
B2MM	Block Majorization-Minimization
B2MS	Block Majorization-Minimization Subspace
BD3MG	Block Distributed Majorization-Minimization Memory Gradient
BP3MG	Block Parallel Majorization-Minimization Memory Gradient
KL	Kurdyka - Łojasiewicz
MG	Memory Gradient
MM	Majorization-Minimization
QMM	Quadratic Majorization-Minimization
SABRINA	Stochastic SuBspace MajoRIrization-MiNimization Algorithm
SMM	Subspace Majorization-Minimization
SQMM	Subspace Quadratic Majorization-Minimization

Chapter 1

General introduction

1.1 Context

This work focuses on the development of iterative algorithms for solving smooth unconstrained optimization problems. Mathematically, these consist in researching a minimizer of a so-called cost function within an entire given Hilbert space. In the situation where the cost function is regular enough to be considered as continuously differentiable, gradient descent approaches appear as the most natural resolution strategies. However, they imply to have access to a complete knowledge of first order information to be fully efficient.

We here place ourselves in the challenging context where such an information is too incomplete to apply the gradient-based usual methods. The first reason is computer-related and is classically induced by the dimension of the problem; the larger this dimension, the more restrictive the storage of certain objects (typically the gradient) in memory. This generates a natural trade-off between capturing the best information as possible and preserving reasonable machine complexities. In the context of signal processing or statistical learning, the management of huge quantities of data, that are growing inexorably with the development of the technological tools available, requires confronting such a situation in each of the problems encountered today. The second source takes it origin from the modeling choices themselves; the construction of a cost function is nothing but a setting in equations of the problem and therefore it bridges the gap between the physical observation world, from which it initially comes from, and the mathematical one attached to its resolution. The complexity of the observation/acquisition model and that of the cost function to be adopted therefore evolve in the same direction and when these are too high, they logically lead to the manipulation of partially known or approximate mathematical objects. As a result, the choice of a purely deterministic framework might be too limited to theoretically account with relevancy for typical observed phenomenons. It is then necessary to consider a setting in which the objects at stake can be modeled as black boxes from which only a few outputs are accessible to the user. On the mathematical point of view, such a framework is said to be probabilistic or stochastic.

1.2 Objectives identifications

Several lines of analysis will be conducted in this thesis as a response to the previous challenges. The first one consists in adapting the existing minimization algorithms by changing the way the data are handled. One possible approach is to use block processing of the different variables, possibly on different machines, while taking into account and controlling the resulting communication delays. This is generally referred to as asynchronous programming and, on the mathematical viewpoint, it generally implies to drastically change the structure of the associated minimization scheme. We then seek in a second step to free ourselves from the purely deterministic setting for a probabilistic one. Adopting such a framework requires the use of specific tools usually employed in the study of so-called random processes. The development of modern probability theory actually promoted the mathematical modeling of uncertainties resulting from a wide class of optimization problems and gave birth to stochastic approximation paradigm in the early 1960s. Somewhat less popularized during the next decades, the recent interest for this field has logically gone hand in hand with the increasing challenges involving large scale data manipulations, in particular in the field of machine learning.

Although this thesis is initially based on the construction of gradient-type descent methods, our starting point differs from the usual approach to the extent it revolves around the Majorization-Minimization (MM) principle and in particular, on quadratic MM schemes. Our motivation initially relies on the fact that many efficient existing schemes from the literature can be interpreted as the result of the minimization, given a certain space, of a surrogate quadratic application, which substitutes for (and approaches) the cost function, in a neighborhood of the current iterate. More specifically, quadratic MM resolution strategy results from the consideration of the specific class of quadratic tangent majorization approximations. Such a choice is all the more encouraging as many theoretical guarantees have been highlighted over the last years for the deterministic framework. The inner construction of any MM update necessarily implies the successive evaluations of the cost function to be decreasing. The resulting algorithm therefore becomes particularly robust to deal with functions whose curvatures properties are difficult to handle and in particular the non-convex ones. In addition, the quadratic structure of the approximations of F promotes a simple closed-form and easily parallelizable updates.

In light of those, the presented thesis is intended to develop two different aspects of differentiable optimization relative to quadratic MM approaches. First, remaining in a deterministic framework, the very structure of quadratic MM scheme is rethought so as for the latter to be a better legitimate candidate capable of meeting the requirements of "efficient" programming at a large scale. Our medium-term objective is to obtain a parallel version of the MM algorithm able to operate in a asynchronous manner. Our second approach consists in extending the existing MM results to a stochastic setting. We particularly pay attention to deal with non-necessary convex cost function in a way to preserve as best as possible the primary versatility of MM algorithms. The second part of the thesis seeks to develop a range of theoretical tools for convergence analysis in the stochastic context, that can be reused for MM algorithms and even beyond. More generally, our investigations lead us to move away from the purely MM framework by constructing new theoretical results for non-convex stochastic optimization.

1.3 Outlines

We provide here a short description of each of the chapters contained in this manuscript. With the exception of this one, there are eight of them that can be divided into two categories regarding the two main objectives of this thesis. Chapters 2-5 are relative to the deterministic part of our work while Chapters 6-8 concern the stochastic one. Figure 1.1 illustrates the dependency of the various chapters in this manuscript.

Chapter 2 is a short literature review on unconstrained and differentiable deterministic optimization. We favour a pedagogical approach by introducing to the reader a reminder of the usual tools and existing algorithms used in such a field. Some points may notably be supplemented in usual academic books as [19] or [178]. Two specific features shall be highlighted. On the one hand, we propose an original approach based on the descent condition notion to better understand the usual convergence proofs strategies. One the other hand, we also introduce the reader to the Kurdyka-Lojasiewicz (KL) theory, that we then we use all along this manuscript as a cornerstone approach to overcome the limitations imposed by a non-convex framework.

Chapter 3 provides a summary of existing results and methods around the so-called quadratic MM principle. Once the notion of (quadratic) majorant has been introduced, we present, in order of increasing complexity, the main schemes that have been developed in the literature in recent years. Some theoretical results that are already known or that have been overlooked are also presented for the sake of understanding and to justify the legitimacy of such approaches. In particular, we introduce the subspace technique, which shall be incorporated in MM algorithm to deal with large scale problems.

Chapter 4 can be considered as our first primary contribution and is dedicated to the asymptotical study of an extended class of quadratic MM based algorithms. Although the use of block parallel versions have already been partially experimented to deal with large scale optimization problems over the last years, no asymptotical guarantees have yet been genuinely proposed in the non-convex setting. In this chapter, we thus promote a complete theoretical investigation for a generic Block parallel MM Subspace (B2MS) scheme for which the cost function under consideration is not supposed to be convex. More specifically, we establish convergence results both global (convergence to a stationary point) and local (convergence speed estimation) by using KL theory.

Chapter 5 exhibits our main contribution to the use of deterministic quadratic MM methods in a large scale optimization setting. B2MS algorithm has the advantage of greatly reducing the complexity of the update steps, but its non-distributed structure generally does not allow it to process the received information to converge in a reasonable number of iterations in high dimension. Nevertheless, the latter algorithm remains an interesting starting point to build a new MM-based scheme whose structure is flexible enough to allow distributed data processing on several autonomous machine cores while allowing the arrival of different updates on a first-in, first-out basis. This results in a distributed version of B2MS we call BD3MG (Block Distributed MM Memory Gradient) for the specific choice of memory gradient subspaces. Once BD3MG scheme is presented, an in-depth study of its convergence properties, through the establishment of a descent inequality, is conducted and several numerical experiments are carried out to prove its efficiency in high dimension on both academic and real data examples.

Chapter 6 can be seen as the stochastic counterpart of Chapter 2. The diversity of existing methods

today has led us to make choices of presentation that we hope to be as little restrictive as possible around stochastic gradient type schemes. For the sake of analogy, we also try to keep the theoretical approach adopted in Chapter 2 via the descent notion but this time adapted to a probabilistic framework. We insist once again on the pedagogical aspect of this chapter, the objective of the latter being to familiarise the reader with the notions that we will discuss in the rest of this manuscript. In particular, we try to highlight the difficulties that a stochastic setting implies for the use, as such, of the non-convex convergence tools from the KL theory.

Chapter 7 is dedicated to the introduction of our stochastic MM scheme we call SABRINA. Although the structure of the latter is based on those of the usual subspace MM quadratic, its interest lies in its ability to remain stable and efficient using only an estimate of the gradient. Through the construction of this new scheme, our objective is thus to recover the initial robustness of the usual MM deterministic scheme for the study of not necessarily convex cost functions. We establish various properties of convergence by showing in particular that SABRINA satisfies a stochastic descent condition. Various numerical tests around binary classification problems are conducted for performance evaluations as well as comparisons with the usual stochastic gradient algorithms from the literature.

Chapter 8 is of theoretical nature and aims to provide novel answers to the problem raised in Chapter 6, namely a new version of the generic KL inequality in a way to be technically consistent with a stochastic context. The objective of this work is to provide a proof methodology based on this new mathematical tool in a way to conclude on the almost sure convergence of a given scheme without convexity hypothesis. The medium-term objective is to apply the latter methodology to refine the convergence guarantees of as many stochastic algorithms as possible.

Chapter 9 finally gives the conclusion by both summarizing the work presented in this manuscript and proposing various medium an long term research perspectives.

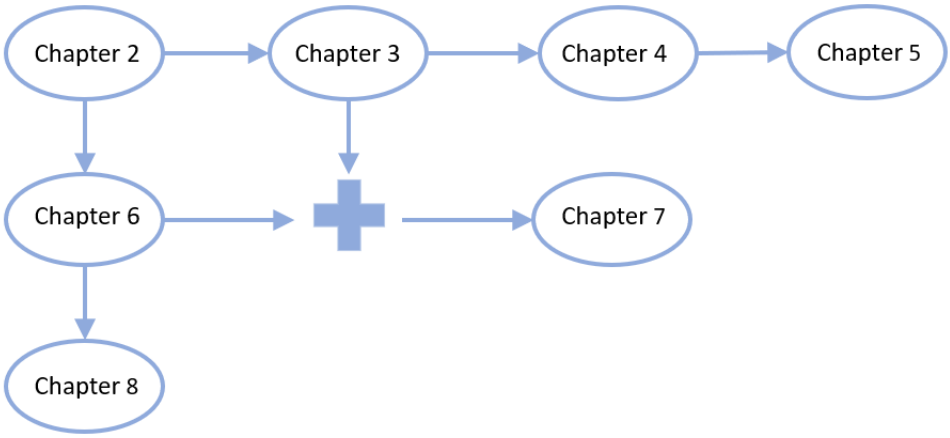


Figure 1.1: Chapters Dependency

1.4 List of publications and contributions

Several chapters of this manuscript are based on publications that I have submitted and sometimes even already published over the last years. My complete publication list, as well as the related chapters, is provided herebelow:

Journal papers and preprints

1 J.-B. Fest and E. Chouzenoux. *Convergence Analysis of Block Majorize-Minimize Subspace Approach*, submitted in 2022 to Optimization Letters (now in minor revision). https://hal-lara.archives-ouvertes.fr/hal-03920026/ [Chapter 4]

2 M. Chalvidal, E. Chouzenoux, J.-B. Fest and C. Lefort. *Block delayed Majorize-Minimize subspace algorithm for large scale image restoration*, Inverse Problems, Special Issue on Optimisation and Learning Methods for Inverse Problems in Microscopy, volume 39, number 4, pp. 044002, 2023. [Chapter 5]

3 E. Chouzenoux and J.-B. Fest. *Sabrina: A stochastic subspace majorization-minimization algorithm*, Journal of Optimization Theory and Applications, volume 195, pages 919-952, 2022. [Chapter 7]

4 E. Chouzenoux, J.-B. Fest and A. Repetti. *A Kurdyka-Lojasiewicz property for stochastic optimization algorithms in a non-convex setting*, https://arxiv.org/abs/2302.06447. [Chapter 8]

International conferences with proceedings

1 J.-B. Fest and E. Chouzenoux. *Stochastic Majorize-Minimize Subspace Algorithm with Application to Binary Classification.* In Proceedings of the 29th European Signal Processing Conference (EUSIPCO 2021), Dublin, Ireland (virtual), August 23-27 2021.

2 J.-B. Fest, A. Repetti and E. Chouzenoux. *A new non-convex framework to improve asymptotical knowledge on generic stochastic gradient descent.* In Proceedings of 2023 IEEE International Workshop on Machine Learning and Image Processing (MLSP), Rome, Italy 17-20 September 2023.

Invited talks

1 J.-B. Fest and E. Chouzenoux. *Majorization-Minimization algorithms: New convergence results in a stochastic setting*, PGMO DAYS, Saclay, France, 30th December 2021.

2 J.-B. Fest and E. Chouzenoux. *Stochastic Subspace Majorization-Minimization Algorithm*, SIAM Conference on Imaging Science (IS22), Berlin, Germany (virtual), 25th March 2022.

3 J.-B. Fest and E. Chouzenoux *Block Delayed MM Subspace Algorithm for Large Scale Image Restoration*, BASP Conference, Villars-sur-Ollon, Switzerland, 10th February 2023.

Table 1.1 below summarizes the themes covered by each chapter of the manuscript.

	Chapter 4	Chapter 5	Chapter 7	Chapter 8
MM algorithms	✓	✓	✓	
Convex optimization			✓	
Non-convex Optimization	✓	✓	✓	✓
Block coordinate optimization	✓	✓		
Distributed optimization		✓		
KL Theory	✓	✓		✓
Stochastic optimization			✓	✓
Stochastic KL theory				✓

Table 1.1: List of the main topics of the manuscript and their distribution by chapter.

Chapter 2

Deterministic differentiable optimization: a general overview

Contents

2.1 Introduction

Unconstrained (Euclidean) optimization is the mathematical field dedicated to find the minimizers of $f : \mathcal{H} \to \mathbb{R}$, a real-valued function defined over a real finite dimensional Hilbert space $(\mathcal{H}, \langle ., . \rangle)$. The common generic formulation of the problem reads in the literature as

$$\text{Find } \boldsymbol{x}_s \in \mathcal{H} \text{ s.t. } f(\boldsymbol{x}_s) \leq f(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathcal{H}. \tag{2.1}$$

A minimizer of f is then defined as a point of $\mathcal{H}$ solution of (2.1). Assuming that f is a continuously differentiable function classically ensures the existence of its associated gradient application $\nabla f : \mathcal{H} \to \mathcal{H}$ as well as its continuity. Beyond its formal role as a representative of the derivative of f, ∇f is physically directly linked to the notion of slope also and thus naturally provides some crucial information on the curvature of f. One of the first fundamental result giving a necessary condition for a point $\boldsymbol{x} \in \mathcal{H}$ to be a minimizer of f is the following:

Theorem 2.1. *(First order optimality condition) Assuming f differentiable, any solution $\boldsymbol{x}^*$ of (2.1) satisfies the Euler equation*

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0}. \tag{2.2}$$

Any vector satisfying (2.2) is named a *stationary point of f*. In concrete terms, the stationary points are those for which the slope of f cancels out and therefore indicate a potential changing of its variations. Theorem 2.1 thus guarantees that any minimizer of f cancels out its slope. Its proof, classically using the notion of directional derivatives, appears in several reference books such as [19]. Only under the differentiability of f, the most common approach to solve (2.1) first consists in focusing on the alternative problem (2.2) reducing to the solution of an equation involving only the gradient operator. Two strategies then emerge for its resolution.

The first one consists in solving (2.2) in an exact manner by only using mathematical operations. However, in the vast majority of the problems encountered, such an operation is generally unthinkable either because of lack of mathematical knowledge or the complexity of the calculation operations involved. The main limiting factor actually remains the dimension of $\mathcal{H}$. The more $\dim(\mathcal{H})$ is large, the more the complexity of (2.2) and therefore that of (2.1). This phenomenon has more generally been known since few decades in all the fields of optimization, beyond the differentiable case, as the curse of dimentionality [16]. One simple example to understand how the space dimension affects the number of operations to obtain the stationary points of f is to consider the situation where ∇f is affine. Solving problem (2.2) becomes equivalent to compute the solution of a linear system and generally requires $\dim(\mathcal{H})^3$ operations as an order of magnitude. Although such a number would remains reasonable regarding academical examples, the great majority of practical minimization problems are encountered in high-dimensional study spaces.

The second strategy historically that emerged to overcome the limiting factors is known as iterative resolution [178]. It consists in generating a sequence $(\boldsymbol{x}_k)_{k \in \mathbb{N}}$, recursively, in order to approach one solution of (2.2) in some mathematical sense. In other words, this strategy is based on the conception of an algorithm in order to ideally asymptotically obtain a stationary point of f. This leads to several lines of analysis. On the first hand, the governing scheme, i.e. the bunch of all necessary operations

giving $\boldsymbol{x}_{k+1}$ from a step $k \in \mathbb{N}$ (where $\boldsymbol{x}_k, \ldots, \boldsymbol{x}_0$ have already been determined) does not require too high computational complexity, otherwise it runs into the same obstacle as the first strategy. On the other hand, obtaining convergences guarantees for the algorithm requires the use of specific mathematical tools and therefore constitutes the major part of the theoretical work with which the numericist is confronted. It is generally carried out in two stages, first, an asymptotical analysis to, in the ideal case, prove the convergence of the iterates to a stationary point, and second by looking more specifically at the speed of convergence of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ or at least of a related alternative sequence, which is physically similar to an accuracy analysis. On a mathematical aspect, this latter study generally amounts to investigate on the evolution, in order of magnitude, of an easily interpretable metric, chosen regarding the results provided by the asymptotical analysis.

In order to present the main principles of unconstrained and differentiable optimization in a pedagogical way, our chapter is divided into five sections. Section 2.2 introduces gradient descent minimization methods from their relatively intuitive nature. In particular, we try to keep a visual and geometrical approach for the reader's understanding. Section 2.3 is more technical in nature and highlights the need for mathematical tools to judge the quality of a minimization algorithm. This requires, in particular, the use of classical analytical assumptions that we will detail in section 2.4. Among these, the convexity assumption is generally considered to be the most powerful as it allows for efficient technical shortcuts to obtain particularly accurate convergence results. However, it remains relatively restrictive and cannot be systematically verified by the cost function under consideration. In such a context, obtaining sufficiently accurate convergence results in the absence of convexity requires the development of new theoretical tools. We provide a description of these in section 2.5. While the notion of quasi-convexity remains relatively popular, the theory based on the use of the so-called Kurdyka-Łojasiewicz property is a recent area of differentiable optimization and gives interesting research perspectives for this thesis.

2.2 The class of descent methods

The most encountered iterative scheme for the research of stationary points and then minimizers of f (i.e. the resolution of (2.2) and (2.1)) are grouped in the category of so-called descent methods. Their general structure is

$$\begin{aligned} &\boldsymbol{x}_0 \in \mathcal{H}, \\ &(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k. \end{aligned} \tag{2.3}$$

Descent shemes thus rely on the construction of two sequences ; directions $(\boldsymbol{d}_k)_{k\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ and stepsizes $(\alpha_k)_{k\in\mathbb{N}} \in \left(\mathbb{R}_+^*\right)^{\mathbb{N}}$ playing the role of adjustment factors. More specifically, directions $(\boldsymbol{d}_k)_{k\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ are said to be descent directions ; they have to be, in some physical sense, compatible with a minimizer research by promoting a global decreasing of the sequence of the iterates evaluation, i.e. $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$. The rest of this section introduces a general overview of the strategies from literature to construct sequences $(\boldsymbol{d}_k)_{k\in\mathbb{N}}, (\alpha_k)_{k\in\mathbb{N}}$ in a way to obtain a solution of (2.2).

2.2.1 Descent direction computation

The keypoint for building an interesting direction $\boldsymbol{d}_k$ at a given iteration $k \in \mathbb{N}$ is the second order Taylor's Formula applied to $\boldsymbol{x}_k$. It gives an overview of f in a neighborhood of $\boldsymbol{x}_k$. By dispensing with writing the order 2 residue and taking a generic $\boldsymbol{h} \in \mathcal{H}$ to have $\boldsymbol{x}_k + \boldsymbol{h}$ close to $\boldsymbol{x}_k$, approximation $f(\boldsymbol{x}_k + \boldsymbol{h}) \simeq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{h} \rangle$ remains valid and encourages $f(\boldsymbol{x}_k + \boldsymbol{h}) \leq f(\boldsymbol{x}_k)$ under the condition $\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{h} \rangle \leq 0$. As sequence $(\alpha_k)_{k \in \mathbb{N}}$ is positive, this naturally leads to the following definition of a descent direction:

Definition 2.1. *(Descent direction) Let $\boldsymbol{x} \in \mathcal{H}$. A vector $\boldsymbol{d} \in \mathcal{H}$ is said to be a descent direction at point $\boldsymbol{x}$ if it satisfies:*

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle < 0. \tag{2.4}$$

The rest of this subsection gives a small bunch of methods from the literature dedicated to build a relevant descent direction (at a given point $\boldsymbol{x}_k,\ k \in \mathbb{N}$).

2.2.1.1 *Steepest descent*

In the light of Definition 2.1, the most natural $(\boldsymbol{d}_k)_{k \in \mathbb{N}}$ sequence to adopt seems to be $(-\nabla f(\boldsymbol{x}_k))_{k \in \mathbb{N}}$, the opposite of the successive gradients. Several justifications are legitimate. One is physical and only keeps in mind the resolution, in the long-run, of (2.1); to approach a minimizer point without any prior information on the function f, the simplest and intuitive strategy is actually to head in the opposite direction to the slope (see figure 2.1). Another one is purely mathematical and focuses on the resolution of (2.2); any stationary point of f is a fixed-point of the recursive scheme $\boldsymbol{x}_{k+1} = J_t(\boldsymbol{x}_k)$ for all $k \in \mathbb{N}$ considering the mapping $J_t : \boldsymbol{x} \in \mathcal{H} \mapsto \boldsymbol{x} - t\nabla f(\boldsymbol{x})$ $(t > 0)$ [32, chapter 2.3]. Methods using $-\nabla f(\boldsymbol{x}_k)$ as descent direction at any iteration $k \in \mathbb{N}$ are grouped under the so-call gradient descent (or steepest descent) class of schemes.

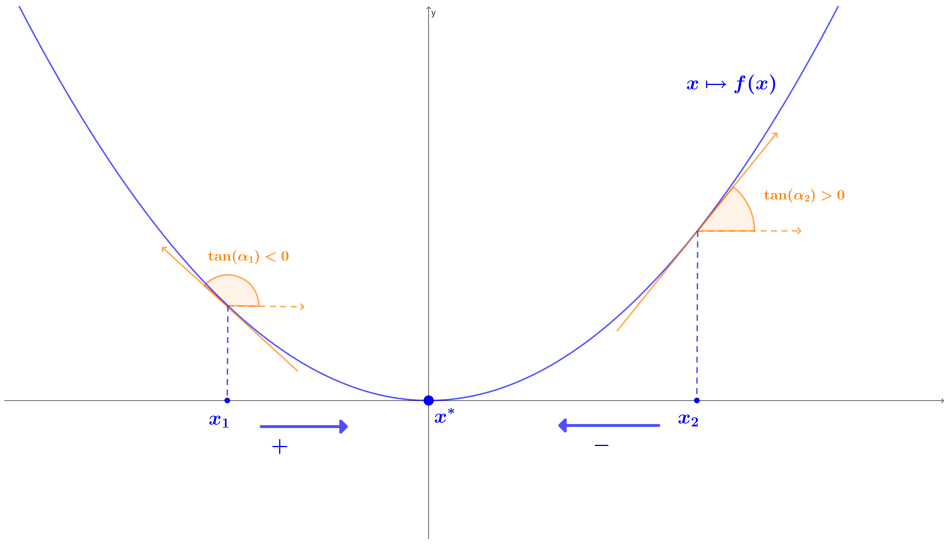


Figure 2.1: An illustration of steepest decent in dimension 1. The new direction is always taken as opposite to the sign the slope at the point considered ; the slope associated to $\boldsymbol{x}_1$ (resp $\boldsymbol{x}_2$) being negative (resp. positive), the new iterate is sought by moving positively (resp. negatively) along the abscissa.

2.2.1.2 *Newton/Quasi-Newton methods*

Many strategies to find a relevant descent direction have been proposed over the last decades. The most famous one remains the class of Newton's methods in the case when f is a twice continuously differentiable function. Historically, Newton's algorithm was developed by Isaac Newton (1643-1727) and Joseph Raphson (1648-1715) to find the roots of any polynomial application (see figure 2.2). Newton's methods simply adapt this strategy to ∇f in a way of solving (2.2). At every iteration $k \in \mathbb{N}$, the new iterate $\boldsymbol{x}_{k+1}$ is defined as a zero of the tangent t_k of ∇f from $\boldsymbol{x}_k$. The advantage of such an approach lies in fact the Taylor's formulas [178] easily give access to the tangent application at any point $x \in \mathcal{H}$. $\boldsymbol{x}_{k+1}$ is thus a solution of the sub-problem

$$\text{Find} \quad \boldsymbol{x} \in \mathcal{H} \quad \text{s.t} \quad t_k(\boldsymbol{x}) := \nabla f(\boldsymbol{x}_k) + \nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k) = 0, \tag{2.5}$$

where $\nabla^2 f(\boldsymbol{x}_k)$ denotes the Hessian matrix of f at point $\boldsymbol{x}_k$. In the case where the Hessian of f is definite positive at any point of $\mathcal{H}$ (or at least at all the iterates), $\boldsymbol{x}_{k+1}$ possesses a closed form (i.e. computable only using the four basic arithmetic operations) which directly conducts to the Newton's update formula:

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \nabla^2 f(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k). \tag{2.6}$$

Equation (2.6) has the advantage of naturally considering $\boldsymbol{d}_k = -\nabla^2 f(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k)$ as a new descent direction at any iteration $k \in \mathbb{N}$. Nevertheless and as mentioned in section 2.1, the dimension of many problems encountered in optimization remains too high to consider an inversion operation for the Hessian (complexity of $\mathcal{O}\left(\dim(\mathcal{H})^3\right)$ for $\nabla^2 f$ of general form). To overcome such an issue the most common strategy consists in building successive approximations of $\nabla^2 f$ or its inverse at every step, i.e. $\boldsymbol{d}_k = -\boldsymbol{H}_k \nabla f(\boldsymbol{x}_k)$ for all $k \in \mathbb{N}$ where $\boldsymbol{H}_k \simeq \nabla^2 f(\boldsymbol{x}_k)^{-1}$ regarding certain criteria. It is necessary to keep in mind that the first goal of such approximation is always to dodge the too demanding inversion of the Hessian operator. The current way to proceed is to generate $(\boldsymbol{H}_k)_{k\in\mathbb{N}}$ recursively by only making small rank modifications. Moreover, in order to get as close as possible to the structure of the Hessian and to recover some of the curvature properties, $(\boldsymbol{H}_k)_{k>0}$ is constructed to be symmetric positive definite and also to verify the same order 2 secant condition as $(\nabla^2 f)^{-1}$ (obtained through Taylor's formula):

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{H}_{k+1} \left(\nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k)\right) = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k. \tag{2.7}$$

Without going into too much technical details, let us just mention the most popular update formulas. The first one is called the Symmetric Rank 1 (SR1) method [191]. Starting from a given $k \in \mathbb{N}$, the new approximation $\boldsymbol{H}_{k+1}$ is of the form of (notation $\otimes$ denotes the usual tensor product) $\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \sigma_k \boldsymbol{v}_k \otimes \boldsymbol{v}_k$, where $(\sigma_k, \boldsymbol{v}_k) \in (0, +\infty) \times \mathcal{H}$ is chosen so as to have, as mentioned previously, $\boldsymbol{H}_{k+1}$ a definite positive operator verifying (2.7). An explicit formula for $\boldsymbol{v}_k$ is known (see [84, 178] for example). The second strategy corresponds to a rank 2 update. It was developed in the early 70s by Broyden, Fletcher, Goldfarb and Shanno, and carries the initial of its designers (BFGS) [40, 99, 230]:

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{H}_{k+1} = \left(\boldsymbol{I} - \frac{\boldsymbol{s}_k \otimes \boldsymbol{y}_k}{\langle \boldsymbol{y}_k, \boldsymbol{s}_k \rangle}\right) \boldsymbol{H}_k \left(\boldsymbol{I} - \frac{\boldsymbol{y}_k \otimes \boldsymbol{s}_k}{\langle \boldsymbol{y}_k, \boldsymbol{s}_k \rangle}\right) + \frac{\boldsymbol{s}_k \otimes \boldsymbol{s}_k}{\langle \boldsymbol{y}_k, \boldsymbol{s}_k \rangle}, \tag{2.8}$$

where for all $k \in \mathbb{N}$ we denote $\boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ and $\boldsymbol{y}_k = \nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k)$. For computational purpose, formula (2.8) is interesting as it can easily be approximated using low machine memory.

The resulting algorithm L-BFGS of [156] is an efficient approach for non-linear optimization at a large scale. Among the rank 2 update formula, we can also mention those of Davidon-Fletcher-Powell (DFP) [75, 100] linked with BFGS by a duality relation. More generally, the mathematical approaches which consists in computing a descent direction using an approximation of the Hessian or its inverse are gathered under the name of Quasi-Newton methods.

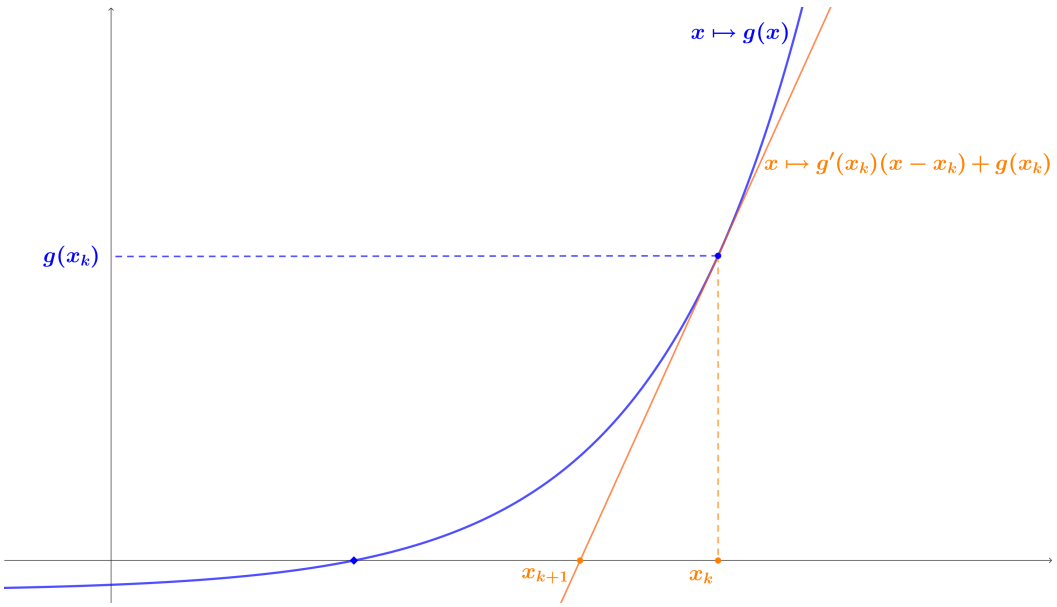


Figure 2.2: A graphical illustration of Newton's method in dimension 1 applied to a generic function $g : \mathbb{R} \to \mathbb{R}$. For any $k \in \mathbb{N}$, $\boldsymbol{x}_{k+1}$ is simply sought as the point cancelling the tangent of g at $\boldsymbol{x}_k$.

2.2.1.3 *Orthogonalization-based approaches*

The advantage of working on a finite dimensional Hilbert is the existence of a finite attached orthogonal base. To such an extent, another strategy consists in building successive directions so that they provide such a base for $\mathcal{H}$. In an ideal situation, $(\boldsymbol{d}_k)_{k\in\mathbb{N}}$ is thus constructed in a way to obtain $\mathcal{H} = Vect\left(\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{\dim(\mathcal{H})}\right)$. Moreover, if the coordinates of $\boldsymbol{x}^*$, a solution of (2.2), are computable regarding this base, then the induced algorithm may converge in a finite number of iterations. The most accessible way to build an orthogonal base of direction consists relies on the Gram-Schmidt process. However, the natural scalar product $\langle ., . \rangle$ is not necessarily used to the extent it does not take account of the structure of the problem.

Typically, when ∇f has an affine structure of the form $\nabla f : \boldsymbol{x} \in \mathcal{H} \mapsto \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$ with $\boldsymbol{A}$ a symmetric positive operator and $\boldsymbol{b}$ a given vector of $\mathcal{H}$, it logically seems more interesting to work with the associated scalar product $\langle ., . \rangle_{\boldsymbol{A}} = \langle ., \boldsymbol{A}. \rangle$. This specific situation is notably at the root of the Conjugate Gradient (GC) scheme [119]. The directions are updated according the alternative orthogonalization

process

$$\begin{aligned}
&\boldsymbol{d}_0 = -\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0 \ (= -\nabla f(\boldsymbol{x}_0)),\\
&(\forall k \in \mathbb{N}) \quad \boldsymbol{r}_{k+1} = A\boldsymbol{x}_{k+1} - \boldsymbol{b} \ (= \nabla f(\boldsymbol{x}_{k+1})),\ \boldsymbol{d}_{k+1} = -\boldsymbol{r}_{k+1} + \frac{\langle \boldsymbol{r}_{k+1}, \boldsymbol{d}_k \rangle_{\boldsymbol{A}}}{\|\boldsymbol{d}_k\|_{\boldsymbol{A}}^2}\boldsymbol{d}_k,
\end{aligned} \tag{2.9}$$

so as to verify the conjugate relation $\langle \boldsymbol{d}_k, \boldsymbol{d}_\ell \rangle_{\boldsymbol{A}} = 0$ for all $k, \ell \in \{0, \dots, \dim(\mathcal{H})\}^2$ and $k \neq \ell$.

The GC method has also been extended to the case where ∇f is not necessarily affine. This results in the following update scheme

$$\begin{aligned}
&\boldsymbol{d}_0 = -\nabla f(\boldsymbol{x}_0),\\
&(\forall k \in \mathbb{N}) \quad \boldsymbol{d}_{k+1} = -\nabla f(\boldsymbol{x}_{k+1}) + \beta_k \boldsymbol{d}_k,
\end{aligned} \tag{2.10}$$

where $(\beta_k)_{k\in\mathbb{N}}$ is usually a ratio of scalar products involving either the past gradient evaluations or the previous descent directions. Few choices of $(\beta_k)_{k\in\mathbb{N}}$ can be found in literature, we can especially mentioned those of Fletcher-Reeves (FR), Polak-Ribière (PR) or Hestenes -Stiefel (HS) [101, 187, 119].

2.2.2 Overview on step-size computation strategies

Proceeding in a similar way as previously, we here introduce the most usual class of methods to build a stepsize sequence $(\alpha_k)_{k\in\mathbb{N}}$ in a way to obtain interesting convergence guarantees for the minimization algorithm, i.e. to asymptotically find a solution of the Euler equation (2.2) or at least to have a decreasing of $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$. The induced problematic for which we here try to give some responses can be formulated as follow ; for a given iteration $k \in \mathbb{N}$ and assuming than $\boldsymbol{x}_0, \dots, \boldsymbol{x}_k$ and the new descent direction $\boldsymbol{d}_{k+1}$ have already been computed, how to obtain the new stepsize α_s as "best" as possible regarding all the information we possess either on the function or on the past process ?

2.2.2.1 *On the necessity of using inexact researches*

The first element of response is to consider the terminology of term "best" in a mathematical aspect. From a purely analytical point of view and to the extent we seek to minimize f, the "best" stepsize α logically seems to be the one which maximizes the decay, i.e. which is a solution of

$$\text{Find } \alpha_s > 0 \text{ s.t. } f(\boldsymbol{x}_k + \alpha_s \boldsymbol{d}_k) \leq f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) \text{ for all } \alpha > 0. \tag{2.11}$$

The approach consisting of finding α_s to verify (2.11) is named the exact step-size research. Solving (2.11) thus conducts to investigate a minimization sub-problem and finally requires to use one of the two strategies mentioned in section 2.1; a direct calculation of one solution only relying on the mathematical properties of f or the construction of a specific search algorithm. In general, none of these two strategies is adopted to the extent that they relatively time consuming [178]. The most common way of getting around such an obstacle is based on the use of an alternative approach known as the inexact stepsize research.

2.2.2.2 *Verifying the decreasing*

Instead of seeking for a solution of (2.11) at any cost, it is required for the new stepsize α_s to ensure a so-call *descent condition*, less restrictive than (2.11) and simply of the form of

$$f(\boldsymbol{x}_k + \alpha_s \boldsymbol{d}_k) \leq f(\boldsymbol{x}_k) - \rho_k(\alpha_s), \tag{2.12}$$

where $\rho_k > 0$ is a reduction coefficient. The most popular one is those proposed by *Goldstein-Armijo* [6, 178]; $\rho_k : \alpha > 0 \mapsto -c_1\alpha\langle\nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle$ given a fixed constant $c_1 \in (0,1)$. Coefficient $c_1\alpha$ acts as a control factor on the process decay. However, Goldstein-Armijo descent condition as such is not restrictive enough to be applied as a practical update's rule; the Taylor's Formula actually ensures it is verified for any $\alpha > 0$ small enough as long as $\langle\nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle < 0$:

$$f(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k) \underset{\alpha\to 0^+}{=} f(\boldsymbol{x}_k) + \underbrace{\alpha\langle\nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle}_{\text{negative and linear in } \alpha} + o(\alpha) \underset{\alpha\to 0^+}{\sim} f(\boldsymbol{x}_k) + \alpha\langle\nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle. \tag{2.13}$$

Typically, when the convergence of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is established under $(\alpha_k)_{k\in\mathbb{N}}$ as choice of stepsize, it still holds if $(\alpha_k)_{k\in\mathbb{N}}$ is replaced by $(\tilde{\alpha}_k)_{k\in\mathbb{N}}$, a smaller one (i.e. such that $\tilde{\alpha}_k \leq \alpha_k$ for all $k \in \mathbb{N}$). On a practical point of view, the limitations are rather based on the evolution speed of the algorithm. The smaller α_s is, the less incidence $\boldsymbol{d}_k$ has on the optimization process. The choice of a good stepsize thus induces a trade-off. On the one hand, the decay of $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ can only be ensured for relatively reasonable choices of control values; taking a stepsize as large as possible remains incompatible with the Goldstein-Armijo descent condition (otherwise it would imply that f is not lower-bounded and de facto the non existence of a solution for (2.1)). On the other hand, a too small stepsize tends to compromise the convergence speed of the algorithm, making it difficult to use for practical applications despite a certain stability.

2.2.2.3 *Wolfe conditions*

As a response of the dilemma mentioned in the previous paragraph, a second condition, relative to the curvature of f regarding the process, generally completes (2.12). It imposes for the new stepsize $\alpha_s > 0$ to be chosen so that $\boldsymbol{d}_k$ is a weaker descent direction with respect to the resulting gradient $\nabla f(\boldsymbol{x}_k + \alpha_s\boldsymbol{d}_k)$:

$$\langle\nabla f(\boldsymbol{x}_k + \alpha_s\boldsymbol{d}_k), \boldsymbol{d}_k\rangle \geq c_2\langle\nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle, \tag{2.14}$$

considering some $c_2 \in (0,1)$. Relation (2.14) is called in the literature the *curvature condition* [178]. The latter mostly works in symbiosis with the descent condition considering the rule of Goldstein-Armijo. Their combination finally leads to the *Wolfe conditions* [237, 178] ; the stepsize sequence $(\alpha_k)_{k\in\mathbb{N}}$ is built so as to verify

$$(\forall k \in \mathbb{N}) \quad f(\boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k) \leq f(\boldsymbol{x}_k) - c_1\alpha_k\langle\nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle \tag{2.15}$$

$$\langle\nabla f(\boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k), \boldsymbol{d}_k\rangle \geq c_2\langle\nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle \tag{2.16}$$

Constants c_1, c_2 are chosen to satisfy $0 < c_2 < c_1 < 1$. Although, it is always possible to find $(\alpha_k)_{k\in\mathbb{N}}$ satisfying Goldstein-Armijo inequality (2.15) (under descent condition (2.4) for the directions), its

counterpart (2.16) is generally more rarely met and also depends on the curvature properties of the function f. The differentiablity alone cannot guarantee the existence of a stepsize able to verify these two conditions at any time [178, 32]. Further stepsize rules also exist as those of *Goldstein-Price* [32] which imposes the Goldstein-Armijo condition as well as an additional lower-bound constraint on the process decay. The research of α_s to verify Wolfe conditions or one of their derivatives is in practice carried out via dichotomous processes also called *backtracking* approaches. Many examples of these methods have been developed [19, 32] and are based on the principle described on figure 2.3.

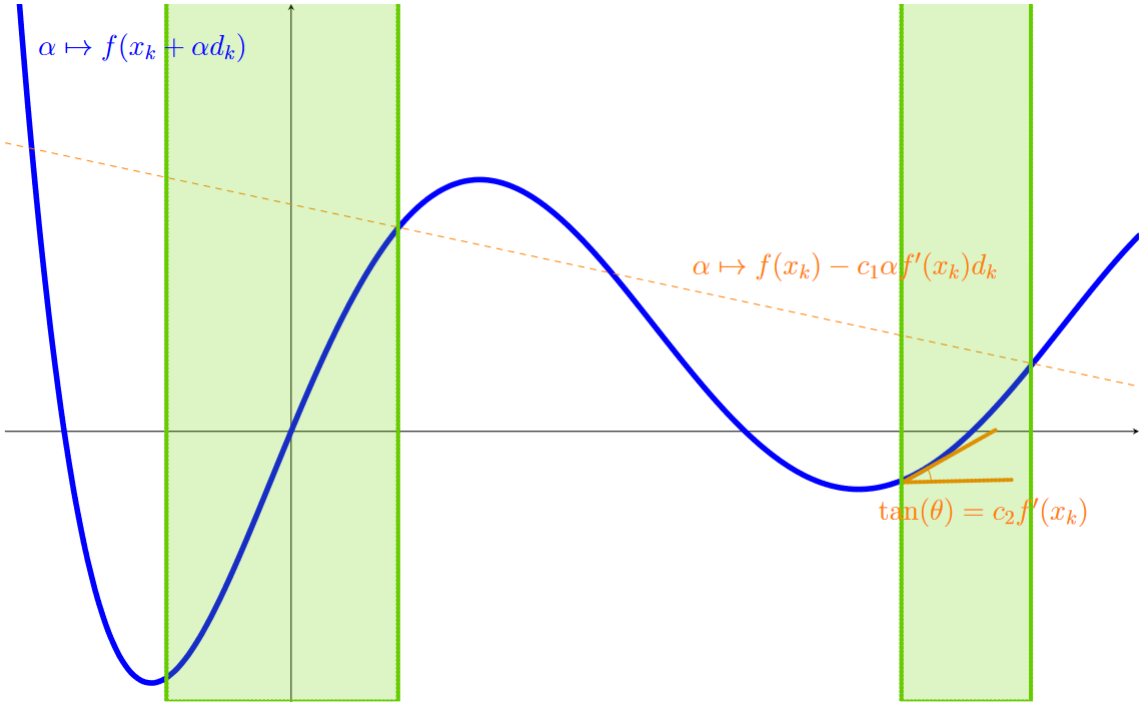


Figure 2.3: An example of stepsize reserach in dimension 1 so as to satisfy the two Wolfe conditions. The green areas representing the set of admissibles values for α_s.

2.3 Mathematical tools for asymptotic analysis

The aim of section 2.2 was to give the reader an overview of existing minimization algorithms in differentiable optimization using a geometric/graphical approach to explain their construction. Here we change our point of view, this time starting from an optimization algorithm. Our objective is to present the various theoretical tools that allow us to evaluate its capacity to approach (or not) a solution of the initial problem (2.1) or at least of the Euler's equation (2.2).

Throughout this section, we consider an algorithm generating a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ for which we want to evaluate the ability to construct a solution of (2.2) or even ideally of the minimization problem (2.1). "Evaluate" here means the establishment of specific mathematical criteria that may or may not be verified by $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$.

2.3.1 The ideal objective of global convergence

Our starting point is the following: saying that sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ approaches a stationary point (of f) mathematical means that it has a limit, solution of (2.2). There exists $\boldsymbol{x}^* \in \mathcal{H}$ such that:

$$\boldsymbol{x}_k \underset{k\to+\infty}{\longrightarrow} \boldsymbol{x}^* \quad \text{and} \quad \nabla f(\boldsymbol{x}^*) = 0. \tag{2.17}$$

This behaviour is generally difficult to obtain to the extent it requires a mathematical background more or less complex which highly depends on the information available on f. Sufficient strong curvature properties obviously favour the establishment of such a result. In practise, the cost function f is derived from a model of various kinds (physical, economic, etc.) that cannot be modified. It is therefore necessary to accept it as such, i.e. with or without some properties. Admitting additional assumptions on f to typically prove the global convergence (2.17) of sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ can mathematically be acceptable, but may severely compromise the use of the associated algorithm, depending on the context. In the case where f remains relatively generic (i.e. differentiable but without "much other requirements"), it is often necessary to deal with less precise convergence results.

2.3.2 Descent condition

Given a function f of any form, the first step relative to the convergence analysis of an optimization scheme is often based on the research of a monotony induced by the process at stake [188]. In a minimization context, it consists in finding alternative sequence derived from $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ with a decreasing behavior. If such a sequence exists, it heuristically implies that the initial scheme, by guaranteeing the decay of a certain quantity of interest all along the iterates, has a certain consistency. We thus define the fundamental and general notion of (l, r)- *descent condition.*

Definition 2.2. *(Descent condition) Let $l : \mathcal{H}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ be an application (defined on the space of sequences of $\mathcal{H}$ and returning sequences of real numbers) and $r = (r_k)_{k\in\mathbb{N}}$ a sequence of real non-negative numbers. A sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ verifies a (l, r)-descent condition if there exists $k_0 \in \mathbb{N}$ for which $(l_k)_{k\in\mathbb{N}} = l\left((\boldsymbol{x}_k)_{k\in\mathbb{N}}\right)$ satisfies the descent inequality*

$$(\forall k \geq k_0) \quad l_{k+1} \leq l_k - r_k. \tag{2.18}$$

- *In the following, we will commonly speak of a Lyapunov application to name l (relative to (l, r)-descent condition) and of a residual sequence to mention r (relative to (l, r)-descent condition).*
- *$(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is said to verify a simple descent condition (relative to f) if there exists a residual $r \in (\mathbb{R}_+)^{\mathbb{N}}$ for which $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ verifies a (l, r)-descent condition regarding $l : (\boldsymbol{y}_k)_{k\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}} \mapsto (f(\boldsymbol{y}_k))_{k\in\mathbb{N}}$, i.e. $l_k = f(\boldsymbol{x}_k)$ for any integer k starting from a certain rank.*

The Goldstein-Armijo relation is one academical example of a simple descent condition we have already met in the previous section (see paragraph 2.2.2.2). Subject to certain assumptions on f, any sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ generated from a gradient descent scheme (2.3) and for which the Wolfe conditions (2.15), (2.16) are valid is inclined to satisfy a simple descent condition (see subsection 2.4.2). More generally, in the situation where the decay of sequence $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ is not ensured, calculation tricks can

enable to obtain a (l, r)-descent condition [188, chapter 2]. Our Lyapunov designation for application l initially comes from the link between our optimization field and the dynamical systems theory. Relation (2.18) can then be seen as a the discrete version of an energy dissipation equation [160]. The reader may find various examples of Lyapunov applications in [235]. Below is a classical example of (l, r)-descent condition which is not simple descent condition.

Example 2.1. *Let us consider a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$. We assume that we are able to prove the existence of positive sequences $(\alpha_k)_{k\in\mathbb{N}}, (\beta_k)_{k\in\mathbb{N}}$ which do not depend on $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ as well as a non-negative function $g : \mathcal{H} \mapsto \mathbb{R}_+$ such that*

$$(\forall k \in \mathbb{N}) \quad f(\boldsymbol{x}_{k+1}) \leq (1 + \alpha_{k+1}) f(\boldsymbol{x}_k) - g(\boldsymbol{x}_k) + \beta_{k+1}. \tag{2.19}$$

Such an inequality can be seen as a relaxed form of a simple descent condition for which the decay of $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ is compromised by additional terms $(\alpha_k)_{k\in\mathbb{N}}, (\beta_k)_{k\in\mathbb{N}}$. However, the following relation can be turned into a (l, r)-descent condition using the following calculation tricks. Denoting $p_k = \prod_{i=0}^{k}(1+\alpha_i)$ for all $k \in \mathbb{N}$ and (2.19) is actually equivalent to

$$(\forall k \in \mathbb{N}) \quad p_{k+1}^{-1} f(\boldsymbol{x}_{k+1}) \leq p_{k+1}^{-1}(1 + \alpha_{k+1}) f(\boldsymbol{x}_k) - p_{k+1}^{-1} g(\boldsymbol{x}_k) + p_{k+1}^{-1}\beta_{k+1}. \tag{2.20}$$

With $p_{k+1}^{-1}\beta_{k+1} = \sum_{i=0}^{k+1} p_i^{-1}\beta_i - \sum_{i=0}^{k} p_i^{-1}\beta_i$, the latter equation can be rewritten as

$$(\forall k \in \mathbb{N}) \quad p_{k+1}^{-1} f(\boldsymbol{x}_{k+1}) - \sum_{i=0}^{k+1} p_i^{-1}\beta_i \leq p_{k+1}^{-1}(1 + \alpha_{k+1}) f(\boldsymbol{x}_k) - \sum_{i=0}^{k} p_i^{-1}\beta_i - p_{k+1}^{-1} g(\boldsymbol{x}_k), \tag{2.21}$$

and, using $p_{k+1}^{-1}(1 + \alpha_{k+1}) = p_k^{-1}$, this also gives

$$(\forall k \in \mathbb{N}) \quad \left(p_{k+1}^{-1} f(\boldsymbol{x}_{k+1}) - \sum_{i=0}^{k+1} p_i^{-1}\beta_i \right) \leq \left(p_k^{-1} f(\boldsymbol{x}_k) - \sum_{i=0}^{k} p_i^{-1}\beta_i \right) - p_{k+1}^{-1} g(\boldsymbol{x}_k). \tag{2.22}$$

Finally, $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ verifies a (l, r)-descent condition taking

$$l : (\boldsymbol{x}_k)_{k\in\mathbb{N}} \mapsto \left(p_k^{-1} f(\boldsymbol{x}_k) - \sum_{i=0}^{k} p_i^{-1}\beta_i \right)_{k\in\mathbb{N}} \quad \text{and } (r_k)_{k\in\mathbb{N}} = \left(p_{k+1}^{-1} g(\boldsymbol{x}_k)\right)_{k\in\mathbb{N}}. \tag{2.23}$$

More complex cases are typically encountered for algorithms whose update makes use, at a given $k \geq \tau$ ($\tau > 0$), of not only $\boldsymbol{x}_k$ but also of a bunch of past iterates $\boldsymbol{x}_{k-1}, \ldots, \boldsymbol{x}_{k-\tau}$. Such a situation generally conducts to considerate $k_0 \geq \tau$, $l : (\boldsymbol{y}_k)_{k\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}} \mapsto (h(\boldsymbol{y}_k, \boldsymbol{y}_{k-1}, \ldots, \boldsymbol{y}_{k-\tau}))_{k\geq k_0}$ where $h : \mathcal{H}^{\tau+1} \to \mathbb{R}$ often corresponds to a sum of f with functions having simple structures (typically a linear combination of norms). Variable τ, in general, is associated to a delay in the process and is typically encountered in distributed algorithms [227, 76].

In addition of ensuring that sequence $(l_k)_{k\in\mathbb{N}}$ is a decreasing one, inequality (2.18) ensures its convergence to a finite limit as soon as l is a lower-bounded function (as a direct consequence of the basic monotone convergence theorem). In the situation where we can build a simple descent condition and assuming that f admits at least one minimizer, the associated descent inequality directly leads to the convergence of the evaluations $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ to a real limit. As it stands, it is difficult to directly act on the value of the limit obtained since a monotonicity-based proof of convergence is purely non-constructive. However, the interest of the (l, r)- descent condition is not limited to such a use. The residual process r is likely to provide valuable additional information, as we explain hereafter.

2.3.3 Residual analysis

Sequence $r = (r_k)_{k\in\mathbb{N}}$ takes relatively varied forms depending on the structure of the studied scheme, and is willing to precious information. As mentioned previously, the lower-boundedness of l is relatively common (and even desired) and guarantees the convergence of $(l_k)_{k\in\mathbb{N}}$ to a finite limit. If such a situation occurs, denoting by $l_{\inf}$ the minimal value of l, and using the decay of the derived sequence, descent inequality (2.18) leads to

$$(\forall k \geq k_0) \quad \sum_{i=k_0}^{k} r_i \leq l_{k_0} - l_{k+1} \leq l_{k_0} - l_{\inf} \in \mathbb{R}. \tag{2.24}$$

Insofar as the upper-bound of (2.24) remains independant from the iterations, the positivity of $(r_k)_{k\in\mathbb{N}}$ ensures the convergence of the associated serie, i.e. $\sum_{k=0}^{+\infty} r_k < +\infty$. A particularly interesting and relatively common case is when the order of magnitude of $(r_k)_{k\in\mathbb{N}}$ is comparable to that of the gradient sequence $(\nabla f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$, i.e. there exist $n > 1$ and a bounded sequence $(\alpha_k)_{k\in\mathbb{N}} \in (0, +\infty)^{\mathbb{N}}$ such that

$$(\forall k \geq k_0) \quad r_k = \alpha_k \|\nabla f(\boldsymbol{x}_k)\|^n. \tag{2.25}$$

We then need to distinguish between two scenarios. The first and easiest one corresponds to the situation where $(\alpha_k)_{k\in\mathbb{N}}$ is bounded below by a positive constant. In such a case, with $\sum_{k=0}^{+\infty} r_k < +\infty$, (2.25) directly ensures that $\sum_{k=0}^{+\infty} \|\nabla f(\boldsymbol{x}_k)\|^n < +\infty$ and so the convergence of $(\nabla f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ to zero is obtained. If $(\alpha_k)_{k\in\mathbb{N}}$ converges to zero, we cannot lead to the same conclusion without additional assumptions. However, if $(\alpha_k)_{k\in\mathbb{N}}$ is adopted prior to the minimization scheme, i.e. fixed in advance, it can be constructed in order to facilitate the emergence of certain asymptotical behaviours. In this context, the typical strategy is to choose $(\alpha_k)_{k\in\mathbb{N}}$ so as to verify the following property.

Proposition 2.1. *Let $(r_k)_{k\in\mathbb{N}}$ a summable sequence satisfying inequality (2.25) for given $k_0 \in \mathbb{N}$ and $n > 1$. If $(\alpha_k)_{k\in\mathbb{N}}$ is non-summable, i.e. $\sum_{k=0}^{+\infty} \alpha_k = +\infty$, then:*

$$\liminf_{k\to+\infty} \|\nabla f(\boldsymbol{x}_k)\| = 0. \tag{2.26}$$

Proof. Let us assume the contrary. There thus exist $\varepsilon > 0$ and a rank $K \in \mathbb{N}$ such that $\|\nabla f(\boldsymbol{x}_k)\| \geq \varepsilon$ for any $k \geq K$. Relation (2.25) then leads to $r_k \geq \alpha_k \varepsilon^n$ considering every k greater than K. As $(\alpha_k)_{k\in\mathbb{N}}$ is non-summable and positive, the latter inequality guarantees that this is also the case for sequence $(r_k)_{k\in\mathbb{N}}$, which contradicts the initial statement. □

Following the definition of the limit inferior, any sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ verifying the conditions of Proposition 2.1 thus possesses, at least, a subsequence for which the gradient evaluations sequence $(\nabla f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to zero. As such, Proposition 2.1 can be seen as a generalisation of the case where $(\nabla f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to zero given a constant stepsize.

2.3.4 Making the link between descent condition and global convergence

Although, the limit inferior criterion (2.26) is generally verified for many existing schemes in the literature, it remains relatively inaccurate. On its own, such a condition does not guarantee the existence of a limit or even of an accumulation point for the sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$. To get closer to a global convergence behaviour (typically starting from (2.26)), one possibility consists in establishing the boundedness of the iterates and invoking classical topological results (in finite dimension).

Proposition 2.2. *Let $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ be a bounded sequence of $\mathcal{H}^{\mathbb{N}}$ for which we denote by χ^∞ its set of accumulation points. Subject to f being continuously differentiable, we have the followings properties:*

(i) *χ^∞ is non-empty and denoting by* dist*, the distance application,* $\operatorname{dist}(\boldsymbol{x}_k, \chi^\infty) \underset{k\to+\infty}{\longrightarrow} 0$,

(ii) *If $(\nabla f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to zero, χ^∞ is contained in* $\operatorname{zer}\nabla f$*, the set of stationary points of f,*

(iii) *If $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ satisfies* (2.26)*, χ^∞ contains at least one point of* $\operatorname{zer}\nabla f$,

(iv) *If $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ satisfies* (2.26) *and $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to a finite limit, the latter necessary belongs to $f(\operatorname{zer}\nabla f)$.*

Proof. (i) The fact that χ^∞ is non-empty is directly due to the boundedness of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ in finite dimension. By contradiction, if the sequence at stake does not converge to 0, then there exist $\varepsilon > 0$ and a subsequence $(\boldsymbol{x}_{\psi_1(k)})_{k\in\mathbb{N}}$ such that

$$(\forall k \in \mathbb{N}) \quad \operatorname{dist}\left(\boldsymbol{x}_{\psi_1(k)}, \chi^\infty\right) > \epsilon. \tag{2.27}$$

Since $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is bounded, $\left(\boldsymbol{x}_{\psi_1(k)}\right)_{k\in\mathbb{N}}$ is also bounded. With $\mathcal{H}$ a finite dimensional space, the set of cluster points of $\left(\boldsymbol{x}_{\psi_1(k)}\right)_{k\in\mathbb{N}}$ is non-empty and is basically included in χ^∞. Therefore, there exists another subsequence $\left(\boldsymbol{x}_{(\psi_1\circ\psi_2)(k)}\right)_{k\in\mathbb{N}}$ and $\boldsymbol{x}'_\infty \in \chi^\infty$ such that $\left\|\boldsymbol{x}_{(\psi_1\circ\psi_2)(k)} - \boldsymbol{x}'_\infty\right\| \underset{k\to+\infty}{\longrightarrow} 0$. Hence, $\operatorname{dist}\left(\boldsymbol{x}_{(\psi_1\circ\psi_2)(k)}, \chi^\infty\right) \underset{k\to+\infty}{\longrightarrow} 0$, which is contradictory to (2.27) and thus concludes the proof.

(ii) Let $\boldsymbol{x}_\infty$ be a point of χ^∞ and $\left(\boldsymbol{x}_{\psi(k)}\right)_{k\in\mathbb{N}}$ be one associated subsequence. The convergence of $(\nabla f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ to zero directly gives

$$\nabla f\left(\boldsymbol{x}_{\psi(k)}\right) \underset{k\to+\infty}{\longrightarrow} \mathbf{0}, \tag{2.28}$$

and the continuity of the gradient finally ensures that $\nabla f(\boldsymbol{x}_\infty) = \mathbf{0}$.

(iii) Following the definition of the limit inferior, there exists a subsequence $\left(\boldsymbol{x}_{\psi_1(k)}\right)_{k\in\mathbb{N}}$ for which

$$\nabla f\left(\boldsymbol{x}_{\psi_1(k)}\right) \underset{k\to+\infty}{\longrightarrow} \mathbf{0}. \tag{2.29}$$

Moreover, the boundedness of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ and thus those of $(\boldsymbol{x}_{\psi_1(k)})_{k\in\mathbb{N}}$ ensures the existence of a subsequence $(\boldsymbol{x}_{(\psi_1\circ\psi_2)(k)})_{k\in\mathbb{N}}$ converging to a vector $\boldsymbol{x}_\infty \in \mathcal{H}$ and basically $\boldsymbol{x}_\infty \in \chi^\infty$. On the one hand, as the gradient of f is continuous, it follows that

$$\nabla f\left(\boldsymbol{x}_{(\psi_1\circ\psi_2)(k)}\right) \underset{k\to+\infty}{\longrightarrow} \nabla f(\boldsymbol{x}_\infty). \tag{2.30}$$

On the other hand, (2.29) also directly leads to

$$\nabla f\left(\boldsymbol{x}_{(\psi_1\circ\psi_2)(k)}\right) \underset{k\to+\infty}{\longrightarrow} \mathbf{0}. \tag{2.31}$$

The uniqueness of the limit finally guarantees that $\nabla f(\boldsymbol{x}_\infty) = \mathbf{0}$.

(iv) Let us denote f_∞, the limit of $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$, then, continuing the previous proof, we have

$$f\left(\boldsymbol{x}_{(\psi_1\circ\psi_2)(k)}\right) \underset{k\to+\infty}{\longrightarrow} f_\infty. \tag{2.32}$$

As a consequence of f continuity and the convergence $\left(\boldsymbol{x}_{(\psi_1\circ\psi_2)(k)}\right)_{k\in\mathbb{N}}$ to $\boldsymbol{x}_\infty \in \operatorname{zer}(\nabla f)$, we also deduce that

$$f\left(\boldsymbol{x}_{(\psi_1\circ\psi_2)(k)}\right) \underset{k\to+\infty}{\longrightarrow} f(\boldsymbol{x}_\infty). \tag{2.33}$$

This conducts to $f_\infty = f(\boldsymbol{x}_\infty)$ and ends the proof. □

Although Proposition 2.2 remains weaker than the global convergence (2.17), it gives better information on the asymptotic behavior of the minimization process. In particular, in the case where f admits a unique stationary point and if the initial minimization problem (2.1) is feasible, then this same stationary point is the unique minimizer of f. Proposition 2.2 (iv) then ensures that sequence $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to the minimal value of f. Concluding on global convergence, i.e. on those of the iterates $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$, here consists in proving that χ^∞ is reduced to a unique element which is usually difficult. To do so, a useful strategy is the use of a connectedness argument when both the boundedness $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ and the convergence of differences $(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)_{k\in\mathbb{N}}$ to zero are guaranteed. It is based on the following theorem whose proof is given by Ostrowski [182, Theorem 26.1] (see figure 2.4 for one geometrical interpretation).

Theorem 2.2 (Ostrowki)**.** *Let $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ be a bounded sequence of $\mathcal{H}^{\mathbb{N}}$ for which $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k \underset{k\to+\infty}{\longrightarrow} \mathbf{0}$. Then, χ^∞, the set of accumulation points of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is non-empty, closed, and connex.*

As a consequence, in a continuously differentiable landscape, any bounded sequence of iterates $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ satisfying (2.26) and the convergence of $(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)_{k\in\mathbb{N}}$ to zero, has a connex set of accumulation points and, from Proposition 2.2 (iii), the latter contains one stationary point of f. Moreover, if χ^∞ turns out to be a finite set, we are able to conclude on the global convergence of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ by the following keypoint. Since any finite non-empty connex set is a singleton, it immediately follows that χ^∞ possesses a stationary point of f as unique element, and so $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ converges to it. Of course this ideal case is not so easily verifiable in practice and highly depends on the assumptions made on f.

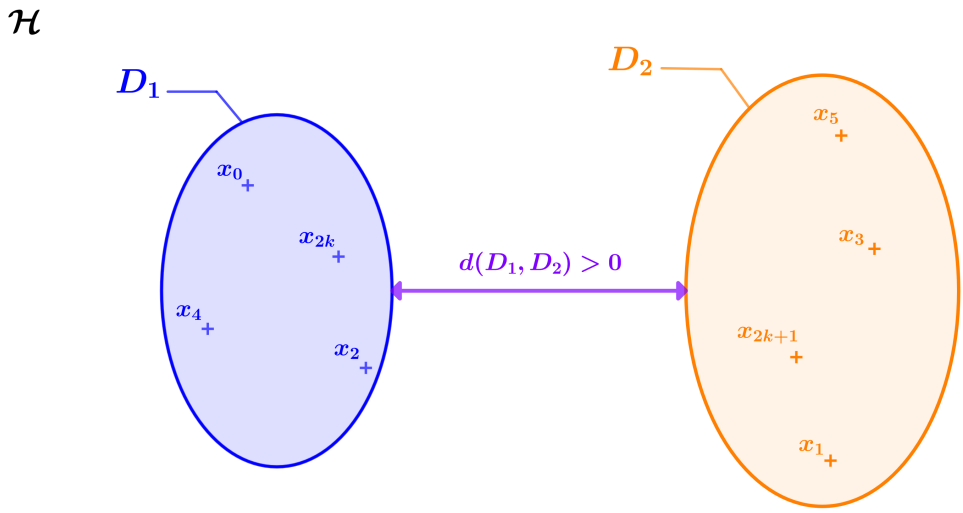


Figure 2.4: One situation where the Ostrowski's theorem does not apply. The even terms of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ belong to compact domain D_1 while the odd ones are contained in D_2 another compact separated from its counterpart by a positive distance $d(D_1, D_2)$. $D_1 \cup D_2$ is not connex as it made up of two separated "islands". It follows that $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| \geq d(D_1, D_2)$ for all $k \in \mathbb{N}$ and, consequently, the difference of terms sequence cannot converge to zero. Moreover, $(\boldsymbol{x}_{2k})_{k\in\mathbb{N}}$ and $(\boldsymbol{x}_{2k+1})_{k\in\mathbb{N}}$ being bounded, D_1 and D_2 each contain at least one accumulation point of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$. Such a situation thus forces χ^∞ to be composed of at least two separated closed set finally making it non-connex.

2.3.5 About local convergence

Even if the global convergence of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is not guaranteed, we have seen that a relatively well-constructed minimization scheme ensures at least one descent condition and thus the convergence of a certain derived process (subsection 2.3.2). Typically, under a simple descent condition, if the shape of f promotes the uniqueness of the solution of (2.1) as well as the boundedness of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$, Proposition 2.2 (iv) is able to ensure the convergence of its evaluations to f_{min}, the f minimum value. To obtain a more precise behaviour on the algorithm, the research for a convergence rate is appreciated as it remains an evaluation criterion which can also be used to highlight the interest of the scheme when compared to other methods. The construction of convergence rates is a very large domain for which we here cannot present all aspects. We nevertheless give key notions to the reader. Larger overviews can be found in [178, 32] as well as a detailed one in [181].

The general principle can be summarized as follows. Let $(m_k)_{k\in\mathbb{N}}$ be a sequence of $\mathbb{R}_+^{\mathbb{N}}$ which supposed to converge to zero. $(m_k)_{k\in\mathbb{N}}$ describes the evolution of a quantity of interest related to a minimization sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ and so depends on the asymptotical properties the latter is able to verify. In the case where $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ and f satisfy the same conditions as in the last example mentioned (see first paragraph), $(m_k)_{k\in\mathbb{N}} = (f(\boldsymbol{x}_k) - f_{min})_{k\in\mathbb{N}}$ appears as the most natural choice of metric to quantify. Even stronger, if sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ itself turns out to converge to a point $\boldsymbol{x}_\infty \in \mathcal{H}$, the choice of $(m_k)_{k\in\mathbb{N}} = (\|\boldsymbol{x}_k - \boldsymbol{x}_\infty\|)_{k\in\mathbb{N}}$ makes perfectly sense.

Generally speaking, building a convergence rate for $(m_k)_{k\in\mathbb{N}}$ consists in finding a positive sequence $(u_k)_{k\in\mathbb{N}}$, of simpler structure, so as to verify

$$m_k \underset{k\to+\infty}{=} \mathcal{O}(u_k). \tag{2.34}$$

The reason for keeping the structure of u_k as simple as possible is to facilitate interpretation and, in particular, evaluation of the algorithm's performance. In others words, it is necessary to keep comparison criteria as readable as possible. Typically, having $(u_k)_{k\in\mathbb{N}}$ of Riemann type $(n^{-\alpha})_{k\in\mathbb{N}}$ $(\alpha > 0)$ or, even better, of a geometrical form $(\epsilon^n)_{k\in\mathbb{N}}$ $(0 < \epsilon < 1)$ are particularly well-suitable.

When a behavior of the form (2.34) with a simple form for $(u_k)_{k\in\mathbb{N}}$ is too difficult to obtain, a common way of getting around such an issue consists in finding a recursive majorization of the form of

$$(\forall k \geq k_1) \quad m_{k+1} \leq \hat{u}_k \; m_k^p, \tag{2.35}$$

for a certain $p \geq 1$, $(\hat{u}_k)_{k\in\mathbb{N}}$ being a positive sequence from a rank $k_1 \in \mathbb{N}$. We end this subsection by mentioning the most encountered regimes:

- If $p = 1$ and $(\hat{u}_k) \in \mathbb{R}_+$ is a constant lying in $(0, 1)$, the convergence is said to be *(Quotient) Q-linear*.
- In the situation where $p = 1$ and $(\hat{u}_k) \in \mathbb{R}_+$ tends to zero, the convergence is said to be *super-linear*.
- When $p = 1$ and $(\hat{u}_k) \in \mathbb{R}_+$ remains constant, the convergence is said to be *quadratic*.

In the literature, p is found under the name of *order of convergence*, the larger it is, the faster the convergence of sequence $(m_k)_{k\in\mathbb{N}}$. In the vast majority of situations and without any prior information on f, the value of p remains small. In particular, only a few schemes are known to perform a quadratic convergence rate; the most famous one corresponding to the Newton's method [178] previoulsy described (see subsection 2.2.1.2). However, the use of convergence rates to illustrate performances turns out to be limited to a mathematical use as it does not take into account the practical complexity of iterations. For example, the Newton's method typically requires to proceed to an inversion at every step and is thus not suited considering a very high dimensional space research. The use of a Quasi-Newton method instead, with at most a super-linear rate of convergence [178] (therefore lower) remains often more efficient from a numerical point of view. To get the best possible idea of an algorithm's performance in practice (and not only in an optimization setting), a general complexity analysis is generally recommended. In such a context, we are then interested in the total number of elementary operations required to achieve a certain level of precision $\epsilon > 0$ in the criterion $(m_k)_{k\in\mathbb{N}}$.

2.3.6 Synthesis

Figure 2.5 concludes this section by giving the reader an overview, in schematic form, of the general strategy presented for conducting the asymptotic study of an optimization scheme in a deterministic and differentiable framework.

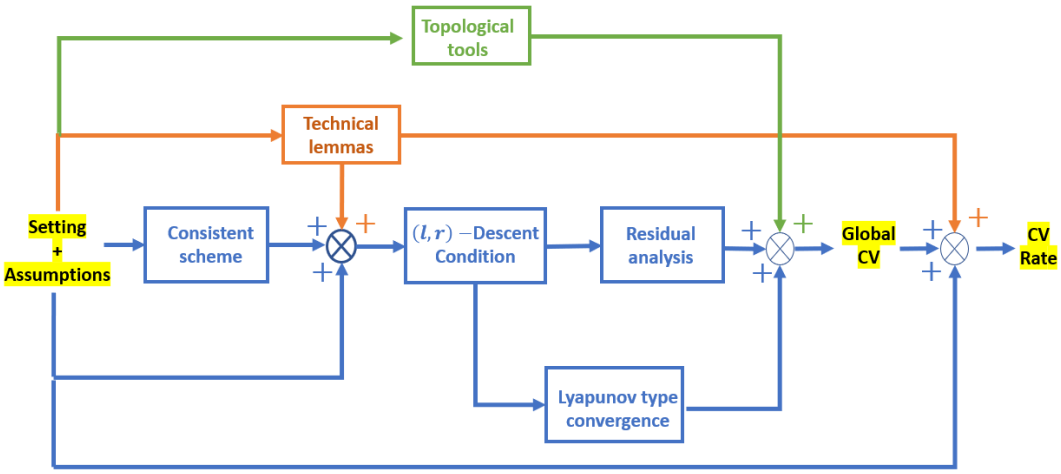


Figure 2.5: Summary diagram for the convergence study of an optimization algorithm in a differentiable framework

We notably use the strategy illustrated in figure 2.5 in Chapters 4 and 5 to investigate on the asymptotical behaviors of the algorithms developped in each of them. Our methodology is thus organized in various successive stages, the difficulties of each one depending on the mathematical tools available and, further upstream, on the assumptions we make about the cost function f.

2.4 Curvature properties of the cost function

In this section, we introduce the most frequent analytical properties (i.e. those on the cost function f) encountered in the differentiable optimization framework. Our aim is to highlight their usefulness, in particular by explaining the key role they play in the strategies developed in section 2.3.

2.4.1 Coercivity

We start this section with a reminder of the notion of coercivity that is often essential to ensure the existence of solutions to optimization problems in a continuous landscape.

Definition 2.3. *(Coercivity) Function $f : \mathcal{H} \longrightarrow \mathbb{R}$ is said to be coercive if it satisfies*

$$f(\boldsymbol{x}) \underset{\|\boldsymbol{x}\| \to +\infty}{\longrightarrow} +\infty. \tag{2.36}$$

In the field of optimization, although such a definition does not require any curvature knowledge, it is generally coupled with continuity through the fundamental following theorem:

Theorem 2.3. *(Existence of minimizers) If f is a continuous and coercive function, then the latter admits at least one minimizer.*

Proof. Several proofs of this result exist [19, 14] using classical topology arguments. We propose here a version adapted to our framework.

Following definition (2.36), there exists $M > 0$ such that $f(\boldsymbol{x}) \geq f(0)$ for every $\boldsymbol{x} \in \mathcal{H}$ satisfying $\|\boldsymbol{x}\| > M$. In finite dimension, $\overline{B}(0, M)$, the closed ball centered at 0 with radius M, is a compact set and thus the Weierstrass extreme value theorem [208] (f is continuous) guarantees that $\inf_{\boldsymbol{x} \in \overline{B}(0,M)} f(\boldsymbol{x})$ lies in $\mathbb{R}$ as well as the existence of $\boldsymbol{x}_0 \in \overline{B}(0, M)$ for which $f(\boldsymbol{x}_0) = \inf_{\boldsymbol{x} \in \overline{B}(0,M)} f(\boldsymbol{x})$. Since $0 \in \overline{B}(0, M)$, we thus have, for all $\boldsymbol{x} \in \mathcal{H}$, $\quad f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0)$ and $\boldsymbol{x}_0$ is a minimizer of f. □

Coercivity of f is thus a sufficient condition to the feasibility of the minimization problem (2.1). In the context of unconstrained optimization (or even more generally when the domain of study is unbounded), there are no real alternative result of existence. Proving that (2.1) is feasible for a non-coercive function generally requires a case-by-case approach to the function under study [11].

The notion of coercity is all the more important in convergence analysis of optimization schemes as it helps to bridge the gap between (l, r)-descent condition (see section 2.3.2) and boundedness.

Proposition 2.3. *Suppose that f is coercive and consider $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ a sequence of $\mathcal{H}^{\mathbb{N}}$. If the evaluations $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converge to a finite limit, $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is bounded.*

Proof. A proof of this result is proposed by [14] in the case where the limits of $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ corresponds to the minimal value of f. We here give another one in a slightly more general framework but with similar arguments.

Let us assume the contrary, i.e. $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is unbounded; we can extract $(\boldsymbol{x}_{\psi(k)})_{k\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$, a subsequence, verifying

$$\|\boldsymbol{x}_{\psi(k)}\| \underset{k\to\infty}{\longrightarrow} +\infty. \tag{2.37}$$

Moreover, denoting f_∞, the limit of sequence $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$, the coercivity of f ensures the existence of $M > 0$ for which any $\boldsymbol{x} \in \mathcal{H}$ s.t. $\|\boldsymbol{x}\| \geq M$ satisfies $f(\boldsymbol{x}) \geq f_\infty + 1$. Especially from (2.37), there exists $K \in \mathbb{N}$ s.t for every $k \geq K$ we have $\|\boldsymbol{x}_{\psi(k)}\| \geq M$ and thus also $f(\boldsymbol{x}_{\psi(k)}) \geq f_\infty + 1$. From this, sequence $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ finally cannot converge to f_∞ and here lies the contradiction. □

Beyond guaranteeing the existence of a solution for the global optimization problem (2.1), the notion of coercivity generally improves the knowledge provided by the descent inequality (2.18). Typically, keeping the same notation as in section 2.3.2 and under a simple descent condition , if f turns out to be continuous and coercive, it admits a minimizer, $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ thus converge to a finite limit (by monotonicity) and sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is finally bounded.

2.4.2 Lipschitz continuity of the gradient

As mentioned in subsection 2.3.2, the satisfaction of a (l, r)-descent condition for an optimization algorithm is the first step to conduct a study of convergence. The property we present in this subsection remains very classical in a differentiable framework and naturally promotes the existence of such a condition for any well-built scheme.

Definition 2.4. *(L-Lipschitz continuity of the gradient) Let $L > 0$. Function $f : \mathcal{H} \longrightarrow \mathbb{R}$ is said to be L-Lipschitz continuous gradient (or L-smooth) if*

$$(\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}) \quad \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{2.38}$$

More generally, f is said to be Lipschitz continuous gradient if there exists $L > 0$ for which f L-Lipschitz continuous gradient.

In addition to being a sufficient condition for gradient continuity, property (2.38) leads to the following fundamental result below, also named the descent lemma:

Proposition 2.4. *(Descent lemma) If $f : \mathcal{H} \longrightarrow \mathbb{R}$ is L-Lipschitz continuous gradient ($L > 0$), then:*

$$(\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}) \quad f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{2.39}$$

The proof directly comes from the integral version of Taylor's formula and can typically be found in [19, Appendix A.24]. Proposition 2.4 applied to f, thus guarantees the existence of a simple descent condition (see Definition 2.2) considering any sequence $(\boldsymbol{x}_k)_{k \in \mathbb{N}}$ for which the residual quantity $\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle < -(L/2)\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2$ from a certain rank.

In such a context, the class of descent methods (introduced in section 2.2) is particularly appropriate to the extent that the computed direction at every step already verifies (2.4). The Lipschitz continuity of the gradient for f especially conducts to a simple descent condition as long as sequence $(\boldsymbol{x}_k)_{k \in \mathbb{N}}$ is generated through scheme (2.3) and $(\boldsymbol{d}_k)_{k \in \mathbb{N}}, (\alpha_k)_{k \in \mathbb{N}}$ are chosen to verify the Wolfe conditions (2.15)+(2.16) [178, Theorem 3.2]. In particular, the associated residual sequence, initially due to Zoutendijk [250] can be defined as:

$$(\forall k \in \mathbb{N}) \quad r_k^Z := \cos^2 \theta_k \|\nabla f(\boldsymbol{x}_k)\|^2 \quad \text{with} \quad \cos \theta_k := \frac{\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle}{\|\nabla f(\boldsymbol{x}_k)\| \ \|\boldsymbol{d}_k\|}, \tag{2.40}$$

and sequence $(\theta_k)_{k \in \mathbb{N}}$ thus corresponds to those of angles between the gradient and the current direction.

2.4.3 Convexity

The notion of convexity remains one of the most essential in differentiable optimization to the extent it allows to make a strong link between minimizers and stationary points, i.e. between solutions to (2.1) and those of the Euler equation (2.2).

Definition 2.5. *(Convexity) Function $f : \mathcal{H} \longrightarrow \mathbb{R}$ is said to be convex if:*

$$(\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{H})(\forall t \in [0,1]) \quad f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq t f(\boldsymbol{x}) + (1-t) f(\boldsymbol{y}). \tag{2.41}$$

The reason why the knowledge of a convexity property for f is crucial in solving problem (2.1) is condensed in the next theorem.

Theorem 2.4. *Assuming that f is differentiable and convex, then its set of minimizers coincides with $\operatorname{zer} \nabla f$, the set of its stationary points.*

Theorem 2.4 therefore provides a partial reciprocal of Theorem 2.1. The latter actually guarantees that $\operatorname{zer} \nabla f$ contains the minimizers of f while adding the convexity assumption enables to prove the inverse inclusion. In other terms, if the convexity of f is satisfied, solving problem (2.1) becomes

equivalent to find its stationary points. It finally follows that any sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ converging to a point $\boldsymbol{x}^*$ such that $\nabla f(\boldsymbol{x}^*) = 0$ automatically converges to a minimizer of f. More generally and beyond our unconstrained differentiable framework, convexity is at the root of many subdomains of optimization in general [204, 90, 172, 38].

We end this subsection by reminding the reader two characterization of convexity relative to f in the differentiable setting:

Proposition 2.5. *Assuming f differentiable, the three following statements are equivalent*

(i) *f is convex,*

(ii) $(\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}) \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle,$

(iii) $(\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}) \quad \langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \geq 0.$

Statement (ii) therefore means that any point of the graph of f is above its associated tangent while (iii) indicates that the gradient is a monotone operator [14].

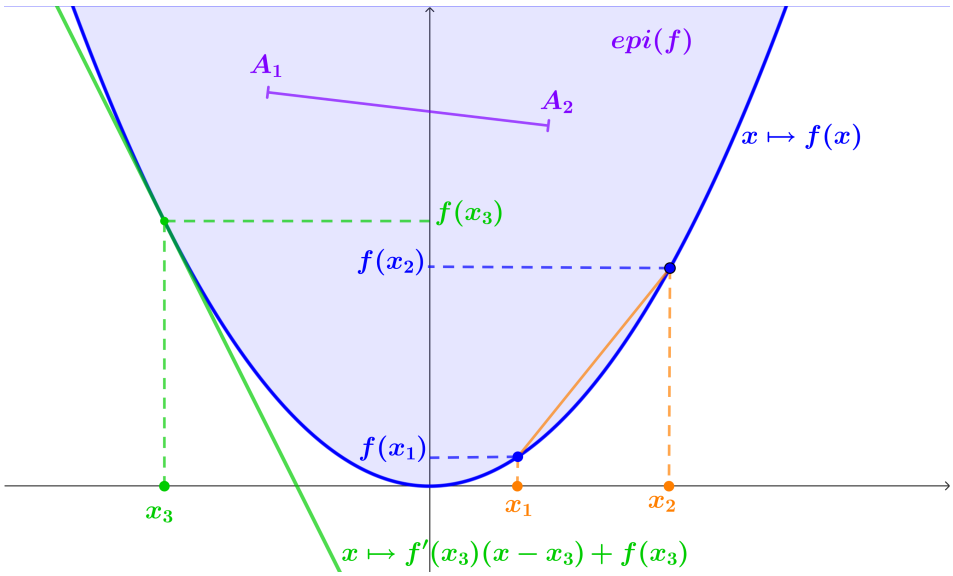


Figure 2.6: Three geometrical interpretations of the notion of convexity. The first one states that the string connecting two points $(\boldsymbol{x}_1, f(\boldsymbol{x}_1)), (\boldsymbol{x}_2, f(\boldsymbol{x}_2))$ of graph(f) always lies above the latter (Definition 2.5) The second one simply indicates that epi$(f) = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{H} \times \mathbb{R} \mid y \geq f(\boldsymbol{x})\}$ (the blue domain) is a convex set; the segment associated to any $\boldsymbol{A}_1, \boldsymbol{A}_2 \in$ epi(f) is still contained in epi(f). The last one illustrates characterization (ii) from Proposition 2.5 for which the tangent at any point $\boldsymbol{x}_3$ minorates the graph of f.

2.4.4 Strict/Strong Convexity

We end this section by introducing two additional properties which can be interpreted as specific versions of the convexity assumption.

Definition 2.6. *(Strict and strong convexity) Function $f : \mathcal{H} \to \mathbb{R}$ is said to be:*

(i) *Strictly convex if:*

$$(\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{H})(\boldsymbol{x} \neq \boldsymbol{y})(\forall t \in (0,1)) \quad f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) < tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}). \tag{2.42}$$

(ii) *μ-strongly convex (μ>0) if $f - \frac{\mu}{2}\|.\|^2$ is convex or equivalently:*

$$(\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{H})(\forall t \in (0,1)) \quad f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) - \frac{\mu t(1-t)}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{2.43}$$

In a more generic way, f is said to be strongly-convex if there exists $\mu > 0$ for which f is μ-strongly convex.

Strong convexity basically implies strict convexity and similarly strict convexity basically implies convexity. The following fundamental proposition is particularly noteworthy and its proof can be found in many optimization books as in [19, Proposition B.10]

Proposition 2.6. *If f is strictly convex, f admits at most one global minimizer.*

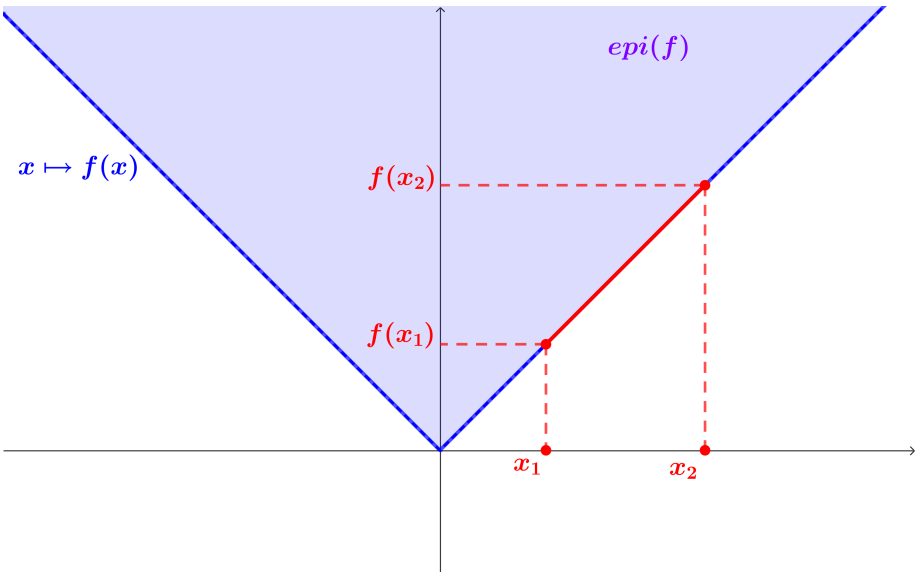


Figure 2.7: $f : x \in \mathbb{R} \to |x|$ is an example of a convex but not strictly-convex function in dimension 1. The segment connecting $(x_1, f(x_1)), (x_2, f(x_2))$ at the border of convex set epi(f) is not included in the interior of the latter.

Proposition 2.6 is therefore not a guarantee for f to admit a minimizer but ensures that, if so, the latter is necessarily unique. With f differentiable and admitting $\boldsymbol{x}_s$ as a minimizer, the combination of this property with Theorem 2.4 notably entails that $\operatorname{zer} \nabla f = \{\boldsymbol{x}_s\}$ every time we place ourselves in a strictly-convex setting. Such a result is particularly interesting from a topological point of view when it comes investigating the behavior of an optimization scheme. Typically, if a bounded sequence

$(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ and has its attached gradient sequence $(\nabla f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converging to zero, then χ^∞ (i.e. its set of accumulation points) is non-empty and is included in $\operatorname{zer}\nabla f$ (Proposition 2.2 (ii)). Strong convexity directly leading to $\chi^\infty = \{\boldsymbol{x}_s\}$, the global convergence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ to the (unique) minimizer of f is finally acted.

By opposition with the Lipschitz continuous gradient property providing a second order upper-bound for f (via Proposition 2.4), strong-convexity ensures the existence of a second order lower-bound and even more generally, we have the following classical characterizations

Proposition 2.7. *With f differentiable, the three following statements are equivalent*

(i) *f is μ-strongly convex,*

(ii) $(\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}) \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2,$

(iii) $(\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}) \quad \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \geq \mu\|\boldsymbol{y} - \boldsymbol{x}\|^2.$

While strict convexity allows to make relevant topological shortcuts and to easily conclude on the global convergence of an optimization algorithm, strong convexity allows to rule a more precise convergence analysis thanks to the attached μ parameter [32, 178]. Strong convexity also has the advantage of ensuring the coercivity of f as a direct corollary of Proposition 2.7 (ii).

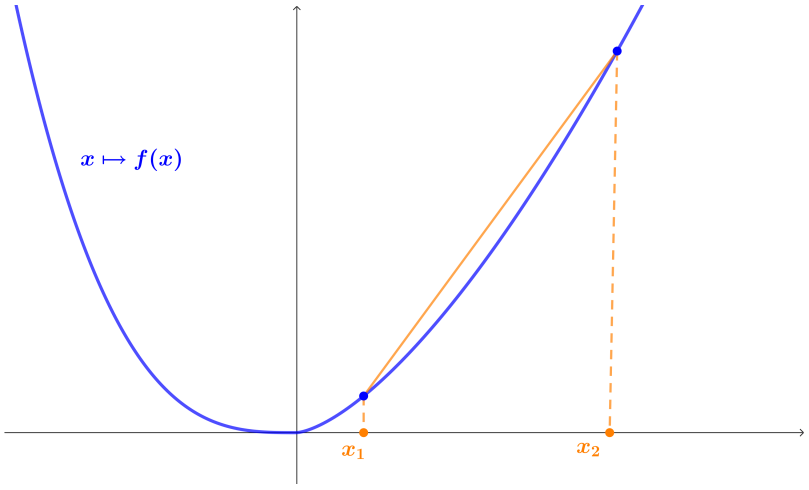


Figure 2.8: A dimension 1 example of a coercive and strictly but non-strongly-convex function $f : x \in \mathbb{R} \mapsto -x^3$ if $x < 0$, $x\sqrt{x}$ otherwise . The non-strongly convexity basically comes from the fact that $f(x)/x^2 \underset{x\to+\infty}{\longrightarrow} 0$.

Geometrically, strongly convex functions are ideal as their curvatures force their unique minimizer to be an attractive point for any decreasing process. We can reason by analogy by considering a minimization algorithm such as a ball moving on a gutter (i.e. the graph of the function) with the form of a pit. Regardless of the starting point of the ball, the latter will always be attracted to the equilibrium point, i.e. the bottom of the gutter or, mathematically speaking, the minimizer.

In view of the potential theoretical guarantees that may result, deciding on the strict/strong convexity or convexity in general, of the problem to be studied, is an essential step before any recourse to an optimization algorithm. However, the convexity property does not concern the vast majority of the cost functions encountered. In such a case, deeper theoretical investigations, require the use of specific tools, still under development at the present time and which are the subject of the last section of this chapter.

2.5 Dealing with the non-convex world in differentiable optimization

Our goal is here to introduce to the reader the fundamental theory we use in Chapter 4, 5 and 8 to deal with optimization problems in a non-convex setting. Historically, the notion of quasi-convexity can be considered as a natural weakening of those of convexity but still remain restrictive, especially regarding the number and the nature of the stationary points it imposes for the cost function f. Nowadays, the most successful theory, at the heart of this section, which is still being improved, is based on the works successively developed by S. Łojasiewicz and K. Kurdyka.

2.5.1 Convex setting limitations

2.5.1.1 *On the set of stationary points in general*

We previoulsy enhanced the interest of the notion of convexity in differentiable optimization for which solving problem (2.1) amounts to finding the stationary points. In the case where such assumption is not verified, it is therefore legitimate to wander what will happen to the relationship between these two problems, i.e. between minimizers and stationary points.

To do so, we first need to remind the notion of local extremum.

Definition 2.7. *(Local extremum) A point $\boldsymbol{x}^* \in \mathcal{H}$ is local extremum of f if there exists a neighborhood V of $\boldsymbol{x}^*$ for which one of the two conditions is satisfied*

(i) $(\forall \boldsymbol{x} \in V) \quad f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$,

(ii) $(\forall \boldsymbol{x} \in V) \quad f(\boldsymbol{x}^*) \geq f(\boldsymbol{x})$.

More specifically if $\boldsymbol{x}^*$ verifies (i) (resp. (ii)), the latter is said to be a local minimizer of f (resp. a local maximizer of f).

In particular, any minimizer of f is a local minimizer considering any neighborhood. The principle obstacle due to the lack of convexity, relies on the following theorem.

Theorem 2.5. *(First order optimality condition 2) Assuming only that f is differentiable, then its local extrema are also its stationary points, i.e. they verify* (2.2).

In a non-convex setting, the characterization of the minimizers of f via the stationary points is therefore lost. On an algorithmic aspect, and in a very pessimistic situation, a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ converging to a stationary point of f is potentially prone to converge to a local maximum. However, such a situation generally cannot be met for a well-built algorithm, as stated in the following proposition.

Proposition 2.8. *Let $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ be a sequence of points from $\mathcal{H}$ for which the evaluations $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ are strictly decreasing. If $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ converges to a limit, the latter is not a local maximizer.*

Proof. Again, we reason by contradiction assuming that $\boldsymbol{x}^*$, the limit of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$, is a local maximizer; there exits $r > 0$ s.t. $f(\boldsymbol{x}^*) \geq f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{H}$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq r$. On the one hand, following the definition of the limit, there exists $K \in \mathbb{N}$ for which $\|\boldsymbol{x}_K - \boldsymbol{x}^*\| \leq r$. On the other hand, the continuity of f ensures that $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to $f(\boldsymbol{x}^*)$ and the strict decay of the latter sequence finally leads to $f(\boldsymbol{x}^*) < f(\boldsymbol{x}_K)$. With $\|\boldsymbol{x}_K - \boldsymbol{x}^*\| \leq r$, this finally contradicts the initial statement. □

Proposition 2.8 thus ensures that any algorithm satisfying a simple descent condition (with a positive residual sequence $(r_k)_{k\in\mathbb{N}}$) cannot converge to a local maximizer. The most challenging situation finally remains the one where the limit point, cancelling the gradient, is a *saddle point*, i.e. a stationary point that is not a local extrema (Theorem 2.5 only ensuring that the set of local extrema is included in those of stationary points).

2.5.1.2 *Quasi-convexity*

As we have already seen, any local minimizer is contained in the set of stationary point (Theorem 2.5). Then, in a convex framework, for which the latter matches with those of minimizers, it thus follows that any local minimizer is actually a minimizer. The concept of quasi-convexity we briefly introduce in this subsection allows to preserve such a property in a less strict framework. Several possible definitions can be found in the literature [190, 114], we here give one of a geometrical nature.

Definition 2.8. *(Quasi-convexity) f is said to be quasi-convex if, for all $c \in \mathbb{R}$, the sub-level set $\{\boldsymbol{x} \in \mathcal{H} \mid f(\boldsymbol{x}) \leq c\}$ is convex.*

Especially, it remains easy to verify that the convexity of f implies its quasi-convexity. The advantage of such a notion relies on the fact that it does not break the connection between local minimizers and minimizers.

Proposition 2.9 ([190])**.** *If f is quasi-convex, each of its local minimizers is a minimizer whenever f is non-constant on all its associated neighborhoods.*

This characterization of quasi-convexity is particularly useful when the graph of f has neither a plateau nor a saddle point. Then, any well-built algorithm (in the sense of Proposition 2.8) which globally converges has its limit point as a solution to the initial problem (2.1). However, quasi-convexity also reduces the nature of the points that $\operatorname{zer}\nabla$ may contain in the following way:

Proposition 2.10. *If f is quasi-convex, its set of stationary points $\operatorname{zer}\nabla f$ does not contain any strict local maximizer, i.e. a point $\boldsymbol{x}^* \in \mathcal{H}$ for which inequality (ii) from Definition 2.7 is strict as soon as $\boldsymbol{x} \in V - \{\boldsymbol{x}^*\}$.*

Proof. We reason by contradiction another time by assuming the existence of $(\boldsymbol{x}^*, r) \in \mathcal{H}$ as well as $r > 0$ for which every $\boldsymbol{x} \in \overline{B}(\boldsymbol{x}^*, r) - \{\boldsymbol{x}^*\}$ verifies $f(\boldsymbol{x}^*) > f(\boldsymbol{x})$. Let us consider the continuous function $\varphi : t \in [-1, 1] \mapsto f(\boldsymbol{x}^* + tr)$ with $\varphi(-1) \leq \varphi(1)$ to set the ideas, the two cases being symmetric. Since $\boldsymbol{x}^* + tr \in \overline{B}(\boldsymbol{x}^*, r)$ for all $t \in [-1, 1]$, φ admits a unique maximizer in $t = 0$ and the intermediate value theorem then states the existence of $t_0 \in [-1, 0]$ s.t. $\varphi(t_0) = \varphi(1)$ (also because $\varphi(1) \geq \varphi(-1)$). Vectors $\boldsymbol{x}^* + t_0 r$ and $\boldsymbol{x}^* + r$ thus belong to $\{\boldsymbol{x} \in \mathcal{H} \mid f(\boldsymbol{x}) \leq \varphi(1)\}$ and, due to the quasi-convexity of f, so it is for $\boldsymbol{x}^* = \lambda(\boldsymbol{x}^* + t_0 r) + (1 - \lambda)(\boldsymbol{x}^* + r)$ with $\lambda = (1 - t_0)^{-1} \in [0, 1]$. It follows that $\varphi(0)(= f(\boldsymbol{x}^*)) \leq \varphi(1)$ which contradicts the fact that 0 is the unique maximizer of φ. □

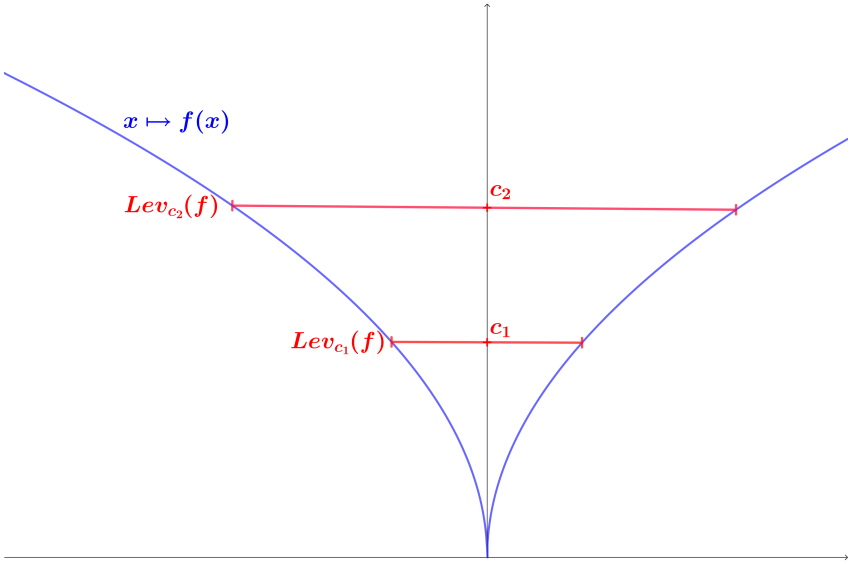


Figure 2.9: A classical example of a quasi-convex but non-convex function in dimension 1; the level sets of $f : x \to \sqrt{|x|^2}$ are all convex while its epigraph (i.e. the domain above graph(f) is not a convex set.

2.5.1.3 *The challenge imposed by the very generic non-convex framework*

Mild assumptions on f such as coercivity and Lipschitz continuity gradient, promote the existence of a sequence $(\boldsymbol{x}_k)_{k \in \boldsymbol{N}}$ able to ensure a (l, r)-descent condition (see Definition 2.2) and then to be bounded (see Proposition 2.3). However, linking the consequences of the descent condition and global convergence remains a considerable challenge, when considering a generic framework. Although the analysis of the residual $(\boldsymbol{r}_k)_{k \in \mathbb{N}}$ tends to promote preliminary interesting links between accumulations points of $(\boldsymbol{x}_k)_{k \in \mathbb{N}}$ and stationary points of f (see proposition 2.2), immediate convergence guarantees are generally only accessible considering a convex or, at least, a quasi-convex setting for which the structure of $\operatorname{zer} \nabla f$ possesses strong properties (absence of bumps, reduction to a singleton in the strictly convex case etc). In general, χ^∞ possesses a very complex structure and the only easily accessible guarantee, from a simple descent condition, is based on the next proposition.

Proposition 2.11. *Assume that $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is such that the evaluations $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to a finite limit, then f is constant on χ^∞.*

Proof. The arguments invoked are closed to those used in the proof of Proposition 2.2 (iii). The result is basically true by convention if χ^∞ is empty. Otherwise, for any $\boldsymbol{x}_\infty \in \chi^\infty$, the convergence of $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ and f continuity simply ensure that $f(\boldsymbol{x}_\infty) = f_\infty$, denoting f_∞ the limit of $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$. □

2.5.2 Kurdyka-Łojasiewicz theory

2.5.2.1 *Initial approach*

An alternative strategy to obtain the global convergence of a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ consists in proving that the latter is Cauchy, using the finiteness of its length i.e.

$$\sum_{k=0}^{+\infty} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| < +\infty. \tag{2.44}$$

Of course, the limiting point, at the origin of the theory we introduce here, lies in the fact that $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is not ensured to satisfy (2.44) even though it turns out to be bounded and to satisfy a (l, r)-descent condition. Instead of considering a theoretical counter-example, we prefer introducing to the reader a geometrical situation for which such a behavior cannot be verified. To do so, let us start from the simple example of steepest descent scheme we presented in section 2.2.1.1. One possible interpretation that we have not discussed so far is related to the proximity of our field with dynamical systems theory. Every sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ generated using the steepest gradient scheme is actually none other than the discrete analogue of a trajectory, $(\boldsymbol{x}(t))_{t\geq 0}$, solution of the *gradient flow* equation

$$\begin{aligned} &\boldsymbol{x}(0) \in \mathcal{H}, \\ &(\forall t > 0) \quad \dot{\boldsymbol{x}}(t) = -\nabla f(\boldsymbol{x}(t)), \end{aligned}$$

and thus remaining orthogonal to any level set of f encountered. In particular, there exist non-trivial categories of differentiable functions f for which $(f(\boldsymbol{x}(t)))_{t\geq 0}$ is decreasing and $(\boldsymbol{x}(t))_{t\geq 0}$ bounded but without converging. The most famous one, represented in figure 2.10, is the "Mexican hat" proposed by [132, 3]. Level sets of f are such that $(\boldsymbol{x}(t))_{t\geq 0}$ describes a spiral shape starting from $\boldsymbol{x}(0)$ and revolving the unit circle without reaching it.

The goal of this subsection is thus to introduce some analytical tools, historically developed in [159, 141], able to promote the global convergence of a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ under mild assumptions as those of (l, r)-descent condition or boundedness.

2.5.2.2 *Łojasiewicz inequality*

Keeping the analogy with the study of gradient flow problem, proving (2.44) for a generic $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ generated through steepest descent, amounts to showing, in the continuous case, that the arc associated to $(\boldsymbol{x}(t))_{t\geq 0}$ is finite, i.e.

$$\int_0^{+\infty} \|\dot{\boldsymbol{x}}(t)\| \, \mathrm{d}t < +\infty.$$

In such a context, a fundamental tool, especially satisfied, in the real case, for any analytical function, relies on an identity initially proposed by Łojasiewicz in the early 60s, [159]. It can be stated as follows:

Definition 2.9. *(Łojasiewicz inequality) Let $\boldsymbol{x}^*$ be a stationary point of f. f verifies the Łojasiewicz (L)-inequality (at $\boldsymbol{x}^*$) if there exist $\theta \in (0, 1/2], \kappa > 0$ and a neighborhood V of $\boldsymbol{x}^*$ s.t. identity $\kappa\|\nabla f(\boldsymbol{x})\| \geq |f(\boldsymbol{x}) - f(\boldsymbol{x}^*)|^{1-\theta}$ is satisfied for any $\boldsymbol{x} \in V$.*

Regarding our discussion in section 2.3.3, this latter property is particularly well-suited to our topological framework. Using a few operations to relate f to l, the Lyapunov function, a (l, r)-descent condition is generally enough to prove the convergence of $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ to a finite limit and also, in many scenarios, the existence of a converging subsequence $\left(\boldsymbol{x}_{\psi(k)}\right)_{k\in\mathbb{N}}$ for which the gradient tends to zero (e.g. through Proposition 2.2 (iii)). As a consequence, if (L)-inequality proves be to verified on the set $\operatorname{zer}\nabla f$, those of stationary points of f, it follows that the identity attached to Definition 2.9 will be satisfied with an infinite number of iterates of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$. In such a situation, (L)-inequality thus indicates that the decay of the gradient subsequence $\left(\nabla f(\boldsymbol{x}_{\psi(k)})\right)_{k\in\mathbb{N}}$ will remain "moderate" with those of $\left(f(\boldsymbol{x}_{\psi(k)})\right)_{k\in\mathbb{N}}$. Constants (θ, κ) acting as two control parameters.

Following the startegy adopted in [129, Theorem 1.1], the usefulness of (L)-inequality in proving the finite length of $(\boldsymbol{x}(t))_{t\geq 0}$ lies in the fact that it allows to upper-bound its associated arc by those of $\left(f^\theta(\boldsymbol{x}(t))\right)_{t\geq 0}$ every time the same $(\boldsymbol{x}(t))_{t\geq 0}$ is a decreasing and convergent process. The example of the gradient flow was notably covered in [142, 3]. It is therefore advisable to adapt such an approach to a discrete framework with the aim of obtaining (2.44) and finally a global convergence result.

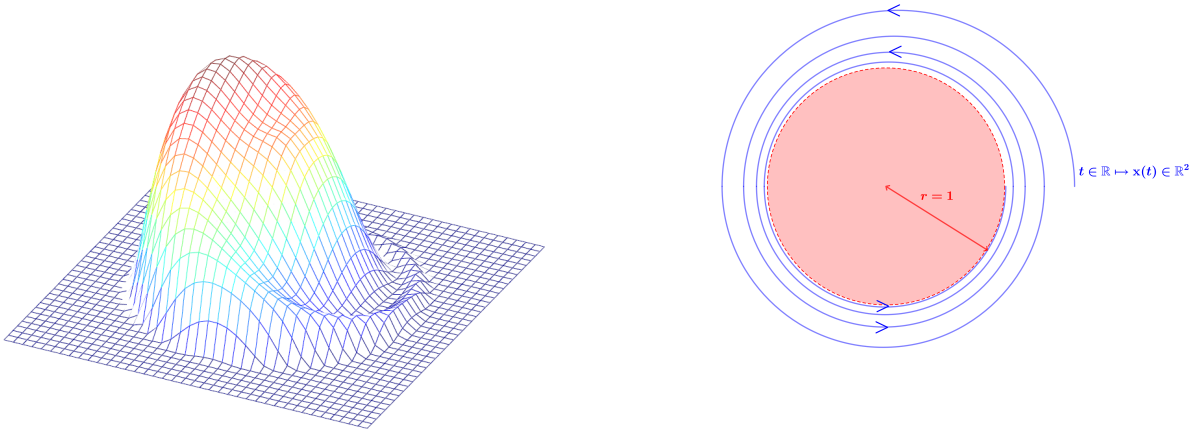


Figure 2.10: An example of function f, the Mexican hat, taken from [3] (left figure), whose associated gradient flow $\boldsymbol{x}(t)_{t\geq 0}$ satisfies $\nabla f(\boldsymbol{x}(t)) \underset{t\to+\infty}{\longrightarrow} 0$ but with a non-finite curvature length. Every orbit of $\boldsymbol{x}(t)_{t\geq 0}$ spinning infinitely around the unit circle (right figure).

2.5.2.3 *The Kurdyka extension*

Although, (L)-inequality does not possess a complex structure, it remains quite restrictive on the smoothness of f as it generally assumes that the latter is analytic. In response to this obstacle, Kurdyka

in [141] was able to propose an alternative identity satisfied by a much larger class of functions.

Definition 2.10. *(Kurdyka-Łojasiewicz property) f verifies the Kurdyka-Łojasiewicz KL-property at $\hat{\boldsymbol{x}} \in \mathcal{H}$ if there exist $\zeta \in (0, +\infty]$, V a neighborhood of $\hat{\boldsymbol{x}}$ and $\varphi : [0, \zeta) \to [0, +\infty)$ s.t.*

(i) $\varphi(0) = 0$,

(ii) φ *is continuously differentiable on* $(0, \zeta)$ *and continuous at* 0,

(iii) φ *is a concave function on* $[0, \zeta)$ *for which* $\varphi'(u) > 0$ *for all* $u \in (0, \zeta)$,

(iv) $\|\nabla f(\boldsymbol{x})\| \varphi'\left(f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}})\right) \geq 1$ *for any* $\boldsymbol{x} \in V$ *satisfying* $f(\hat{\boldsymbol{x}}) < f(\boldsymbol{x}) < f(\hat{\boldsymbol{x}}) + \zeta$.

More specifically, considering $E \subset \mathcal{H}$, f is said to satisfy the KL-property on E, if f satisfies the KL-property at every point of E.

KL-property is of local nature but tends to generalize the initial version of Łojasiewicz. Identity of Definition 2.9 can actually be recovered considering the situation where $E = \{\boldsymbol{x}^*\}$ for $\boldsymbol{x}^* \in \operatorname{zer} \nabla f$, $\zeta = +\infty$ and $\varphi : u \in [0, +\infty) \mapsto \kappa \theta^{-1} u^{\theta}$. In his very generic framework [141], Kurdyka managed to show that KL-property is satisfied for any function f definable through an o-minimal structure [229] including, in the real case (i.e. $\mathcal{H} = \mathbb{R}^N$ $(N \geq 1)$), those which are semi-algebraic. The definition we propose is not the original of [73] but rather a characterization due to [21].

Definition 2.11. *(Real semi-algebraicity) Considering $\mathcal{H} = \mathbb{R}^N$ $(N \geq 1)$, function $f : \mathbb{R}^N \to \mathbb{R}$ is said to be (real) semi-algebraic if its graph can be expressed as a finite combination of unions/intersections formed by polynomial domains, i.e. there exist $I, J \geq 1$ and a family $\{p_{i,j},\ q_{i,j} \mid 1 \leq i \leq I, 1 \leq j \leq J\}$ of $2 \times I \times J$ polynomials defined on $\mathbb{R}^{N+1}$ for which*

$$\operatorname{graph}(f) = \left\{(\boldsymbol{x}, f(\boldsymbol{x})) \mid \boldsymbol{x} \in \mathbb{R}^N\right\} = \bigcup_{i=1}^{I} \bigcap_{j=1}^{J} \left\{\boldsymbol{x} \in \mathbb{R}^{N+1} \mid p_{i,j}(\boldsymbol{x}) = 0,\ q_{i,j}(\boldsymbol{x}) > 0\right\}. \tag{2.45}$$

In particular, any set of this form is said to be semi-algebraic.

Since the class of semi-algebraic sets is stable considering the most usual operations [21] (finite intersection or union, complement, Cartesian product and orthogonal projection), semi-algebraic functions logically possess similar advantages. More specifically, they form a ring (regarding operation $+, \times$) endowed with the composition stability property [72]. In addition, the attached function in KL-property can be taken under the form of $\varphi : u \in (0, \zeta) \mapsto c u^{1-\theta}$ where θ lies in $(0, 1)$ and $c > 0$. Nevertheless, the estimation of exponent θ remains a difficult challenge which hardly depends on the involved function [149, 241].

2.5.2.4 *Back to global convergence*

By even adapting the Kurdyka-Łojasiewicz theory to a non-differentiable framework, H. Attouch and J. Bolte were the first to prove the global convergence of some optimization schemes in non-convex settings using finite length arguments [8, 9]. As such, the exploitation of KL-property as introduced in Definition 2.10 is not easy to handle, due to its extreme local nature. Parameter ζ and function

φ actually depend on the point we decide to stand and typically, in a sequential context for which $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges, we expect to have a different KL-inequality for each accumulation point of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$. One of the major contribution of [8, 27] in such a context relies on the construction of a so-call uniformization theorem derived from Definition 2.10, allowing a choice of ζ and φ which remains uniform for a specific category of $\mathcal{H}$ subsets. We shall retain the general theorem from [27]:

Theorem 2.6. *(Uniformized KL property) Let C be a non-empty compact set of $\mathcal{H}$. If f is constant on C and satisfies the KL property on this same subset, then there exists $(\zeta,\varepsilon)\in(0,+\infty)^2$ and $\varphi:[0,\zeta)\to\mathbb{R}_+$ satisfying i), ii) and iii) of Definition 2.10 s.t. $\|\nabla f(\boldsymbol{x})\|\varphi'\left(f(\boldsymbol{x})-f(\overline{\boldsymbol{x}})\right)\geq 1$ for any $(\boldsymbol{x},\overline{\boldsymbol{x}})\in\mathcal{H}\times C$ satisfying $\operatorname{dist}(\boldsymbol{x},C)<\varepsilon$ and $f(\overline{\boldsymbol{x}})<f(\boldsymbol{x})<f(\overline{\boldsymbol{x}})+\zeta$.*

Theorem 2.6 thus promotes the existence of uniform KL parameters over a given subset C as soon as the latter is compact and with a constant image by f. In the context of building global convergent algorithms, the set of accumulation point χ^∞ of a bounded sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$, for which $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to a finite limit, easily verifies such conditions (see Proposition 2.2). From this, a strategy to obtain the global convergence of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ can be summarized as follows: using the concavity of φ and assuming that the descent inequality of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is permissive enough, Theorem 2.6 allows to majorize the gradient sequence $(\|\nabla f(\boldsymbol{x}_k)\|)_{k\in\mathbb{N}}$ or even ideally $(\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|)_{k\in\mathbb{N}}$ by those of the difference $(l_{k+1}-l_k)_{k\in\mathbb{N}}$ keeping the notations of Definition 2.2. We here propose a short application example to clarify our point and give to the reader an taste of the KL-based resolution strategy.

Example 2.2. *Let us consider a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}\in\mathcal{H}^{\mathbb{N}}$ verifying a simple descent condition starting from a rank $k_0\in\mathbb{N}$ and whose residual can be chosen of the form of (2.25) with $n=2$, i.e. there exists a positive sequence (α_k) s.t.*

$$(\forall k\geq k_0)\quad f(\boldsymbol{x}_{k+1})\leq f(\boldsymbol{x}_k)-\alpha_k\|\nabla f(\boldsymbol{x}_k)\|^2. \tag{2.46}$$

Assuming f is coercive, $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to a finite limit $f_\infty\in\mathbb{R}$ while $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is basically bounded (Proposition 2.3) and as seen in Proposition 2.11, f remains constant on χ^∞. Moreover, the boundedness of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ also ensures χ^∞ is a non-empty compact set for which sequence $(\operatorname{dist}(\boldsymbol{x}_k,C))_{k\in\mathbb{N}}$ converges zero (Proposition 2.2). Let us then consider two cases.

- *If $f(\boldsymbol{x}_k)=f_\infty$ for a certain $K\geq k_0$, descent inequality forces $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ to be equal to f_∞ starting from this same rank and positivity of $(\alpha_k)_{k\in\mathbb{N}}$ ensures that $\nabla f(\boldsymbol{x}_k)=0$ for any $k\geq K$ which roughly leads to $\sum_{k=0}^{+\infty}\alpha_k\|\nabla f(\boldsymbol{x}_k)\|<+\infty$.*
- *In the opposite case, i.e. $f(\boldsymbol{x}_k)>f_\infty$ for any $k\geq k_0$, Theorem 2.6 can be applied considering $C=\chi^\infty$ and*

$$K_C:=\min\left\{l\geq k_0\mid\forall k\geq l,\ f_\infty<f(\boldsymbol{x}_k)<f_\infty+\zeta,\ d(\boldsymbol{x}_k,C)<\varepsilon\right\}, \tag{2.47}$$

is finite denoting $\zeta,\varepsilon,\varphi$ the attached KL parameters. We therefore have

$$(\forall k\geq K_C)\quad\|\nabla f(\boldsymbol{x}_k)\|\varphi'\left(f(\boldsymbol{x}_k)-f_\infty\right)\geq 1. \tag{2.48}$$

The next keypoint of the proof is then based on φ concavity; as a differentiable function on $(0,\zeta)$, Proposition 2.5 ensures that φ verifies:

$$(\forall k\geq K_C)\quad\varphi(f(\boldsymbol{x}_k)-f_\infty)-\varphi(f(\boldsymbol{x}_{k+1})-f_\infty)\geq\varphi'\left(f(\boldsymbol{x}_k)-f_\infty\right)\left(f(\boldsymbol{x}_k)-f(\boldsymbol{x}_{k+1})\right) \tag{2.49}$$

and due to descent inequality (2.46)*, we also deduce that*

$$(\forall k \geq K_C) \quad \varphi(f(\boldsymbol{x}_k) - f_\infty) - \varphi(f(\boldsymbol{x}_{k+1}) - f_\infty) \geq \alpha_k \varphi'\left(f(\boldsymbol{x}_k) - f_\infty\right) \|\nabla f(\boldsymbol{x}_k)\|^2. \tag{2.50}$$

The use the uniform KL inequality (2.48) *finally leads to*

$$(\forall k \geq K_C) \quad \varphi(f(\boldsymbol{x}_{k+1}) - f_\infty) - \varphi(f(\boldsymbol{x}_k) - f_\infty) \geq \alpha_k \|\nabla f(\boldsymbol{x}_k)\|. \tag{2.51}$$

The left term of (2.51) *being telescopic with* $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ *converging to* f_∞*, its attached sum (starting from* K_C*) is finite and equal to* $\varphi(f(\boldsymbol{x}_{K_C})-f_\infty)-\varphi(0)(=\varphi(f(\boldsymbol{x}_{K_C})-f_\infty)$*due to* $\varphi(0) = 0$*). Since every term involved in* (2.51) *is positive, we even obtain the existence of quantity* $\sum_{k=K_C}^{+\infty} \alpha_k \|\nabla f(\boldsymbol{x}_k)\|$ *with* $\sum_{k=K_C}^{+\infty} \alpha_k \|\nabla f(\boldsymbol{x}_k)\| \leq \varphi(f(\boldsymbol{x}_0) - f_\infty)$ *and so again we deduce that* $\sum_{k=0}^{+\infty} \alpha_k \|\nabla f(\boldsymbol{x}_k)\| < +\infty$*.*

In both cases, we finally concluded that sequence $(\alpha_k \|\nabla f(\boldsymbol{x}_k)\|)_{k\in\mathbb{N}}$ *is summable. Obtaining more specifically* (2.44) *depends on the initial structure of the adopted scheme. Considering the steepest descent methods (see 2.2.1.1),* (2.44) *is basically satisfied by* $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ *since, independently of* $(\alpha_k)_{k\in\mathbb{N}}$ *(in that case this corresponds to the stepsize), we simply have* $(\alpha_k \|\nabla f(\boldsymbol{x}_k)\|)_{k\in\mathbb{N}} = (\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|)_{k\in\mathbb{N}}$*.*

2.6 Conclusion

Throughout this chapter, we have provided an introduction to the field of unconstrained differentiable optimization that we hope will be as didactic as possible. Although sections 2.2 and 2.4 introduce very classical notions, those of 2.3 and 2.5 are generally less encountered and a good understanding of the different tools and strategies exposed is necessary to fully grasp the interest of the different works we present in the rest of this manuscript. All the theoretical reasoning in deterministic terms is actually first based on the results of section 2.3 and then refined by applying those of section 2.5. The notions recalled in sections 2.2 and 2.3 nevertheless remain essential to set the scene for Chapter 3, dedicated to Quadratic Majorization-Minimization methods.

Chapter 3

Quadratic Majorization-Minimization algorithms

Contents

3.1 Outlines

The goal of this chapter is to introduce an important class of differentiable methods, generated from the Quadratic Majorization-Minimization (QMM) schemes and at the root of this thesis. Although the latter is related to gradient descent approaches, our presentation differs from the one of Chapter 2 as we first focus, in section 3.2, on the so-called Majorization-Minimization (MM) principle. Section 3.3 then introduces the QMM class of methods whose updates rules are based on the existence of specific surrogates of $f : \mathcal{H} \to \mathbb{R}$, the cost function, we call quadratic-tangent majorization approximations. The usual construction strategies of such objects are detailed in section 3.4. The update rules of basic QMM schemes involving too costly operations in high dimension, it is generally appropriate to incorporate additional steps into the resulting algorithms for the purposes of complexity reduction. In such a context and in section 3.5, we speak more specifically about subspaces steps (or techniques), the most widespread in our field nowadays. Section 3.6 aims to highlight the robustness of QMM schemes, with subspaces incorporations or not, by showing their ability to verify a simple descent condition. On the one hand, it allows us to make the link with the theoretical notions discussed in Chapter 2, which we use throughout this manuscript. On the other hand, it provides a clear mathematical background for our further theoretical analysis in Chapter 4,5 and 7. Section 3.7 finally gives few words of conclusion.

3.2 Motivations

As already discussed, working in a differentiable setting does not necessary imply that the minimization of f is easy. When f possesses a complex structure, a reasonable strategy consists in studying an approximation of it, constructed in regards with the available information. In such a context, Majoration-Minimization (MM) principle assumes, at each point $\boldsymbol{x} \in \mathcal{H}$, the existence of a surrogate application having a graph always above that of f, and merged with this latter at $\boldsymbol{x}$. This is mathematically referred to as tangent majorization approximation of f at point $\boldsymbol{x}$.

Definition 3.1. *(Tangent majorization approximation) A bi-component function $h : \mathcal{H}^2 \to \mathcal{H}$ is said to be a tangent majorization approximation (or surrogate) (of f on $\mathcal{H}$) if the two following conditions are verified:*

$$(\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{H}^2) \quad \begin{cases} f(\boldsymbol{y}) \leq h(\boldsymbol{y}, \boldsymbol{x}), \\ f(\boldsymbol{x}) = h(\boldsymbol{x}, \boldsymbol{x}). \end{cases} \tag{3.1}$$

The next step consists in building a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ whose updates simply relies on successive minimizations of h at current iterates:

$$\begin{aligned} &\boldsymbol{x}_0 \in \mathcal{H}, \\ &(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} \in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{H}} h(\boldsymbol{x}, \boldsymbol{x}_k). \end{aligned} \tag{3.2}$$

Implicitly, the tangent majorization approximation h should remain relatively easy to handle. Otherwise, the use of an approximation as structurally complex as the original function would be of low

interest. The construction of an algorithm based on the minimization of successive majorization surrogates was historically proposed in [181]. However, the article that popularised such an approach is probably [81]. The latter proposed a new statistical method, known as Expectation-Maximization (EM), to deal with robust parameters estimation, when considering incomplete observation data. Explicit denomination (3.2) was formulated in the early 2000s in [125, 144].

As such, update rule (3.2) necessary leads to the decay of the evaluations of the iterates, the tangent majorization behavior of h, (Definition 3.1) ensuring that

$$(\forall k \in \mathbb{N}) \quad f(\boldsymbol{x}_{k+1}) \leq h(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) \leq h(\boldsymbol{x}_k, \boldsymbol{x}_k) = f(\boldsymbol{x}_k).$$

The resulting scheme has therefore the advantage of naturally possessing a certain stability when minimizing the differentiable function f. Indeed, $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ follows a simple descent condition in every situation (at worst case with a zero residual sequence). Convergence and stability of scheme (3.2) have notably been investigated in [127], considering generic tangent majorization approximations.

As mentioned, to build a relevant update rule, it becomes necessary to assemble tangent majorization approximations that are also relatively easy to minimize. More specifically, we will see in the next section that the class of quadratic functions is a good candidate to meet such a requirement.

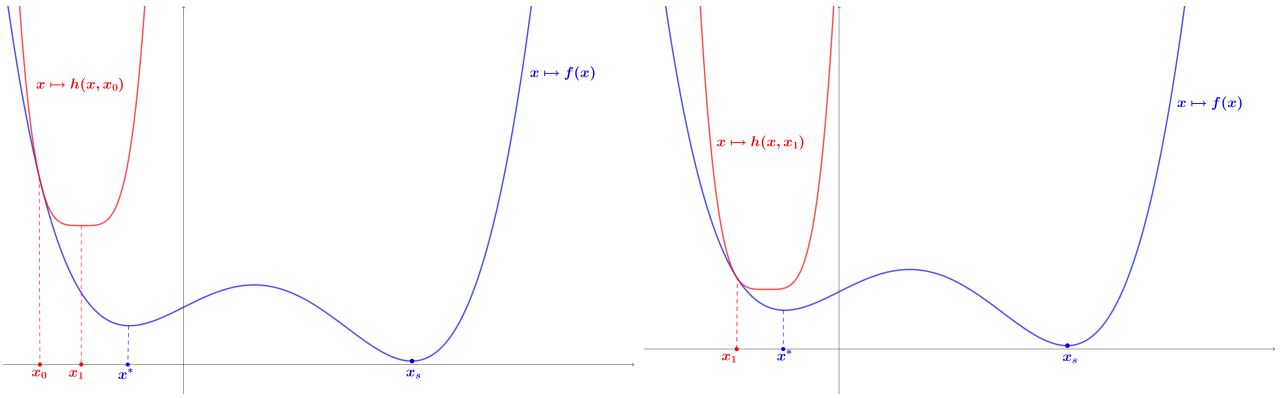


Figure 3.1: A simple graphical illustration of MM principle. Here we consider a non-convex fonction possessing a local minimizer $\boldsymbol{x}^*$. For a well-built tangent majorization approximation of f, the MM sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is expected to converge to $\boldsymbol{x}^*$.

3.3 Quadratic Majorization-Minimization approach

Similarly with second degree polynomials in real analysis, quadratic functions are fundamental tools in differentiable optimization. They appear implicitly in a large number of minimization methods. The use of quadratic functions to build tangent majorization approximations gave birth to a specific version, known as quadratic, of the MM algorithm. This approach is used in a majority of situations due to its simplicity and efficiency.

3.3.1 Quadratic optimization reminder

Before going further, we need to introduce few useful notations we use throughout this chapter.

We denote by $\mathcal{S}(\mathcal{H})$, the set of (bounded) self-adjoint linear operators of $\mathcal{H}$ and $\mathcal{S}(\mathcal{H})^* = \mathcal{S}(\mathcal{H}) - \{\mathbf{0}\}$. $\mathcal{S}^{++}(\mathcal{H})$ then corresponds to the set of elements of $\mathcal{S}(\mathcal{H})$ that are definite positive. In such a case, we remind that for all $\boldsymbol{M} \in \mathcal{S}^{++}(\mathcal{H})$, $\lambda_{min}(\boldsymbol{M})$ and $\lambda_{max}(\boldsymbol{M})$, the smallest and biggest eigenvalues of $\boldsymbol{M}$, verify $\lambda_{min}(\boldsymbol{M}) > 0$ and $\lambda_{min}(\boldsymbol{M})\|\boldsymbol{x}\|^2 \leq \langle \boldsymbol{M}\boldsymbol{x}, \boldsymbol{x}\rangle \leq \lambda_{max}(\boldsymbol{M})\|\boldsymbol{x}\|^2$ for any $\boldsymbol{x} \in \mathcal{H}$. More specifically, a mapping $\boldsymbol{A} : \mathcal{H} \to \mathcal{S}^{++}(\mathcal{H})$ is said to be uniformly-bounded if there exist $\underline{\mu}, \overline{\mu} \in (0, +\infty)$ for which inequalities $\underline{\mu}\|\boldsymbol{y}\|^2 \leq \langle \boldsymbol{A}(\boldsymbol{x})\boldsymbol{y}, \boldsymbol{y}\rangle \leq \overline{\mu}\|\boldsymbol{y}\|^2$ are satisfied for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$.

Definition 3.2. *($(\boldsymbol{M}, \boldsymbol{b}, c)$-Quadratic function) Let $\boldsymbol{M} \in \mathcal{S}^*(\mathcal{H})$, $\boldsymbol{b} \in \mathcal{H}$ and $c \in \mathbb{R}$. A function $q : \boldsymbol{x} \in \mathcal{H} \to \mathbb{R}$ is said to be $(\boldsymbol{M}, \boldsymbol{b}, c)$-quadratic (on $\mathcal{H}$) if it can be written as:*

$$(\forall \boldsymbol{x} \in \mathcal{H}) \quad q(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{M}\boldsymbol{x}, \boldsymbol{x}\rangle - \langle \boldsymbol{b}, \boldsymbol{x}\rangle + c. \tag{3.3}$$

More generally, q is said to be quadratic if there exists $(\boldsymbol{M}, \boldsymbol{b}, c) \in \mathcal{S}^*(\mathcal{H}) \times \mathcal{H} \times \mathbb{R}$ for which q is $(\boldsymbol{M}, \boldsymbol{b}, c)$-quadratic.

The advantage of such functions lies in simple form for which the associated operator $\boldsymbol{M}$ concentrates the curvature information:

Proposition 3.1. *Let $(\boldsymbol{M}, \boldsymbol{b}, c) \in \mathcal{S}^*(\mathcal{H}) \times \mathcal{H} \times \mathbb{R}$ and $q : \mathcal{H} \to \mathbb{R}$ be a $(\boldsymbol{M}, \boldsymbol{b}, c)$-quadratic function (on $\mathcal{H}$):*

(i) *q is convex (resp. strictly convex) if and only if $\boldsymbol{M}$ is positive (resp. definite positive).*

(ii) *If $\boldsymbol{M}$ is definite positive, q is coercive, $\boldsymbol{M}$ is invertible and q admits $\boldsymbol{M}^{-1}\boldsymbol{b}$ as unique minimizer.*

Proof. These results are classical and their respective proofs can easily be found in a majority of differentiable optimization courses. Here is a reminder.

(i) q is differentiable as a polynomial function with gradient $\nabla q : \boldsymbol{x} \in \mathcal{H} \mapsto \boldsymbol{M}\boldsymbol{x} - \boldsymbol{b}$. Proposition 2.5 (iii) then means that q is convex if and only if $\langle \boldsymbol{M}(\boldsymbol{x} - \boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \geq 0$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$ which is equivalent for $\boldsymbol{M}$ to be positive. The same reasoning applies for the strictly convex case.

(ii) If $\boldsymbol{M}$ is definite positive, denoting $\lambda_{min}(\boldsymbol{M})$ its smallest eigenvalue, we have $\lambda_{min}(\boldsymbol{M}) > 0$ and then Cauchy-Schwarz inequality leads to

$$(\forall \boldsymbol{x} \in \mathcal{H}) \quad q(\boldsymbol{x}) \geq \lambda_{min}(\boldsymbol{M})\|\boldsymbol{x}\|^2 - \|\boldsymbol{b}\|\|\boldsymbol{x}\| + c.$$

$\lambda_{min}(\boldsymbol{M})$ positivity directly ensures that $q(\boldsymbol{x}) \underset{\|\boldsymbol{x}\| \to +\infty}{\longrightarrow} +\infty$ and hence q coercivity. According to (i), function is also a strictly convex and therefore admits a unique minimizer $\boldsymbol{x}^*$ which is also its unique stationary point i.e. $\boldsymbol{M}\boldsymbol{x}^* - b = 0$ or, equivalently, $\boldsymbol{x}^* = \boldsymbol{M}^{-1}\boldsymbol{b}$ using $\boldsymbol{M}$ invertibility. □

3.3.2 Quadratic MM (QMM) scheme

The great strength of quadratic functions lies in their omnipresence in the interpretation of gradient-descent approaches. As Example 3.1 below highlights, any well-conditioned scheme of the form of (2.3) can be interpreted as the minimization of successive quadratic functions whose order 0 and 1 terms coincide with the first order Taylor expansion of f around the current iterate. More generally, the approach based on quadratic approximations tends to promote monotonicity of the resulting process [24] and may easily lead to simple descent conditions (typically as those of Zoutendijk mentioned in section 2.4.2).

Example 3.1. *Let us consider $(\boldsymbol{H}_k)_{k\in\mathbb{N}}$ a family of $\mathcal{S}^{++}(\mathcal{H})$. As a consequence of Proposition 3.1, the update of any sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ generated from a descent scheme (2.3) (see section 2.2), whose descent directions are $(\boldsymbol{d}_k)_{k\in\mathbb{N}} := (-\boldsymbol{H}_k\nabla f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$, can be rewritten, using Proposition 3.1 for all $k\in\mathbb{N}$, as*

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\alpha_k \boldsymbol{H}_k \nabla f(\boldsymbol{x}_k) = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{H}} q_k(\boldsymbol{x} - \boldsymbol{x}_k),$$

where q_k is $((\alpha_k\boldsymbol{H}_k)^{-1}, \nabla f(\boldsymbol{x}_k), f(\boldsymbol{x}_k))$-quadratic, i.e.

$$(\forall \boldsymbol{x}\in\mathcal{H}) \quad q_k(\boldsymbol{x}) = f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}\rangle + \frac{1}{2}\langle \alpha_k^{-1}\boldsymbol{H}_k^{-1}, \boldsymbol{x}\rangle.$$

In a similar way, for what we are interested in next, one can refine the notion of tangent majorization approximation (see section 3.2) within a quadratic setting.

Definition 3.3. *($\boldsymbol{A}$-quadratic tangent majorization approximation). Let $\boldsymbol{A} : \mathcal{H} \to \mathcal{S}^{++}(\mathcal{H})$ be a uniformly-bounded mapping. A tangent majorization approximation $h_q : \mathcal{H}^2 \to \mathcal{H}$ (of f on $\mathcal{H}$) is said to be an $\boldsymbol{A}$-quadratic tangent majorization approximation (or surrogate) (of f on $\mathcal{H}$) if for all $\boldsymbol{x}\in\mathcal{H}$, function $\boldsymbol{y}\in\mathcal{H} \mapsto h_q(\boldsymbol{x}+\boldsymbol{y}, \boldsymbol{x})$ is $(\boldsymbol{A}(\boldsymbol{x}), \nabla f(\boldsymbol{x}), f(\boldsymbol{x}))$-quadratic, i.e.*

$$(\forall \boldsymbol{x},\boldsymbol{y}\in\mathcal{H}) \quad f(\boldsymbol{y}) \le h_q(\boldsymbol{y},\boldsymbol{x}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle + \frac{1}{2}\langle \boldsymbol{A}(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle. \tag{3.4}$$

- *More generally, f is said to be quadratic tangent majorizable if there exists a mapping $\boldsymbol{A} : \mathcal{H} \to \mathcal{S}^{++}(\mathcal{H})$ for which f admits an $\boldsymbol{A}$-quadratic tangent majorization approximation.*
- *A uniformly-bounded mapping $\boldsymbol{A} : \mathcal{H} \to \mathcal{S}^{++}(\mathcal{H})$ for which f possesses an $\boldsymbol{A}$-quadratic tangent majorization approximation is said to be a majorization mapping (for f).*

This new class of quadratic surrogates thus induces a specific MM scheme that we name Quadratic MM (QMM), relying on the existence of a majorization mapping $\boldsymbol{A} : \mathcal{H} \to \mathcal{S}^{++}(\mathcal{H})$ for f. According to Proposition 3.1 (ii), the general update formula (3.2), in that case, can be rewritten under a closed-form expression as

$$(\forall k\in\mathbb{N}) \quad \boldsymbol{x}_{k+1} = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{H}} h_q(\boldsymbol{x}, \boldsymbol{x}_k) = \boldsymbol{x}_k - \boldsymbol{A}_k^{-1}\nabla f(\boldsymbol{x}_k), \tag{3.5}$$
$$\text{where } \boldsymbol{A}_k := \boldsymbol{A}(\boldsymbol{x}_k).$$

Update rule (3.5) can be intrepreted as an alternative descent scheme for which directions $(\boldsymbol{d}_k)_{k\in\mathbb{N}}$ are taken as $\boldsymbol{d}_k = -\boldsymbol{A}_k^{-1}\nabla f(\boldsymbol{x}_k)$ for all $k\in\mathbb{N}$, and stepsize is fixed to one. Contrary to the class

of Quasi-Newton type updates, it is not required for sequence $(\boldsymbol{A}_k)_{k\in\mathbb{N}}$ to accurately account for the second-order information of f in the neighborhood of the current iterate. As a majorization mapping, the primary role of $\boldsymbol{A}$ always consists in preserving the majorization inequality (3.4) and thus guaranting a theoretical stability through $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ decreasing (see subsection 3.2). As we will see in the next section, a large choice of majorization mappings is generally possible and even desired to take $\boldsymbol{A}$ as accurate as possible so that the curvature to the associated surrogate at stake $h_q(.,\boldsymbol{x}_k)$ $(k\in\mathbb{N})$ remains closed to that of f.

3.4 Majorization mappings construction strategies

In this section, we present a range of different analytical strategies for the explicit construction of quadratic tangent majorization approximations.

3.4.1 Existence results

As such, the possibility or not of constructing quadratic tangent majorization approximations for a given f, remains the first factor limiting the use of QMM approaches. Nevertheless, the existence of such approximations is ensured under mild assumptions, typically the Lipschitz continuous gradient property, on the cost function.

Proposition 3.2. *If f is L-Lipschitz continuous gradient, $\boldsymbol{A}:\boldsymbol{x}\in\mathcal{H}\mapsto L\boldsymbol{I}\in\mathcal{S}^{++}(\mathcal{H})$ is a majorization mapping for f.*

Proof. This result is a straightforward consequence of descent lemma (see Proposition 2.4). The fact that f is L-Lipschitz gradient continuous leads to

$$(\forall \boldsymbol{x},\boldsymbol{y}\in\mathcal{H})\quad f(\boldsymbol{y})\leq f(\boldsymbol{x})+\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle+\frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2. \tag{3.6}$$

i.e. f admits a $(L\boldsymbol{I},\nabla f(\boldsymbol{x}),f(\boldsymbol{x}))$-quadratic tangent majorization approximation at every $\boldsymbol{x}\in\mathcal{H}$, which directly ends the proof. □

More generally, there exists an analytic characterization of functions possessing a majorization mapping.

Proposition 3.3. *The two following items are equivalent:*

(i) *There exists $\beta>0$ s.t. $\langle\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle\leq\beta\|\mathbf{y}-\boldsymbol{x}\|^2$ for all $\boldsymbol{x},\boldsymbol{y}\in\mathcal{H}$,*

(ii) *f is quadratic tangent majorizable.*

Proof. Implication $(i)\Longrightarrow(ii)$ is a direct consequence of [127, Lemma 3.3 (a)]. This simply implies that the constant mapping $\boldsymbol{x}\in\mathcal{H}\mapsto\beta\boldsymbol{I}$ is a majorization one for f. If (ii) holds then there exists a majorization mapping of f, $\boldsymbol{A}$ and $\overline{\mu}>0$ such that $\langle\boldsymbol{A}(\boldsymbol{x})\boldsymbol{y},\boldsymbol{y}\rangle\leq\overline{\mu}\|\boldsymbol{y}\|^2$ for any $\boldsymbol{x},\boldsymbol{y}\in\mathcal{H}$. Let us now fix $\boldsymbol{x},\boldsymbol{y}$ two vectors of $\mathcal{H}$. Using the quadratic majorization property (3.4), we then deduce that

$$f(\boldsymbol{y})\leq f(\boldsymbol{x})+\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle+\frac{1}{2}\langle\boldsymbol{A}(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle\leq f(\boldsymbol{x})+\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle+\frac{\overline{\mu}}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \tag{3.7}$$

and symmetrically

$$f(\boldsymbol{x}) \leq f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\overline{\mu}}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{3.8}$$

Summing (3.7) with (3.8) directly leads to

$$\langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq \overline{\mu} \|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{3.9}$$

Relation (3.9) being verified for generics $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$, (i) is thus obtained simply taking $\beta = \overline{\mu}$. □

As condition (i) is satisfied by any Lipschitz continuous gradient function, the latter characterization can be seen as an extension of the existence result that Proposition 3.2 guarantees. Beyond ensuring the existence of majorization mappings for a certain class of differentiable functions. Proposition 3.3 conversely can be used to identify those which do not possess any.

Example 3.2. *Let us consider $f : \boldsymbol{x} \in \mathcal{H} \mapsto \|\boldsymbol{x}\|^{2r}$ where $r > 1$ and whose gradient is $\nabla f : \boldsymbol{x} \in \mathcal{H} \mapsto 2r\|\boldsymbol{x}\|^{2r-2}\boldsymbol{x}$. f does not satisfy (i) of Proposition 3.3 and thus is not quadratic tangent majorizable.*

Otherwise, there exists $\beta > 0$ for which for any $\boldsymbol{x} \in \mathcal{H}$, $\boldsymbol{x} \neq \boldsymbol{0}$, taking $\boldsymbol{y} = 0$, $\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} \rangle \leq \beta \|\boldsymbol{x}\|^2$ i.e. $2r\|\boldsymbol{x}\|^{2r} \leq \beta\|\boldsymbol{x}\|^2$. This conducts to $\|\boldsymbol{x}\|^{2r-2} \leq (2r)^{-1}\beta$ for all $\boldsymbol{x} \in \mathcal{H}$, and thus $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|^{2r-2}$ is a bounded function, hence the contradiction with $r > 1$.

3.4.2 A key construction lemma

Taking $\boldsymbol{A} = L\boldsymbol{I}$ as the majorization mapping for f (L the Lipschitz constant), the associated scheme (3.5) becomes equivalent to the steepest descent considering L^{-1} as constant stepsize. Although such a choice always guarantees a certain stability, it remains quite pessimistic and tends to generate poor approximations of f taking only very little account of its curvature properties. Wherever possible, a most appropriate choice would consist in taking $\boldsymbol{A} : \boldsymbol{x} \mapsto \nabla f^2(\boldsymbol{x})$. The resulting quadratic tangent majorization surrogates involved in the algorithm then would exactly match with the second order Taylor expansion of f at the current iterates. Such a framework remains quite restrictive and the latter point especially generally implies for f to be strongly convex. However, there exists intermediary approaches which typically promotes the existence of non-constant majorization mappings without making use of the Lipschitz constant, that we describe below.

Lemma 3.1. *Let $u : t \in \mathbb{R} \mapsto \mathbb{R}$ be a differentiable function satisfying the following conditions,*

(i) *u is even,*

(ii) *u is increasing on $[0, +\infty)$,*

(iii) *$u(\sqrt{\cdot})$ is concave on $[0, +\infty)$,*

(iv) *Function $w : t \in (0, +\infty) \mapsto \frac{u'(t)}{t}$ is positive and admits a positive continuous extension to 0.*

Then, the following inequality holds

$$(\forall (s,t) \in \mathbb{R}^2) \quad u(t) \leq u(s) + u'(s)(t-s) + \frac{1}{2} w(|s|)(t-s)^2 \tag{3.10}$$

and $s \in \mathbb{R} \mapsto w(|s|) \in \mathbb{R}$ is a majorization mapping of u if w is uniformly bounded on $[0, +\infty)$.

Proof. Set $s \in \mathbb{R}^*$ and $t \in \mathbb{R}$. Function $u(\sqrt{.})$ is derivable on $(0, +\infty)$ with $u'(\sqrt{.})/(2\sqrt{.})$ as derivative. We then apply (iii) considering point t^2 and s^2. The characterization of concavity in the differentiable case (Proposition 2.5(ii)), directly leads to:

$$u\left(\sqrt{t^2}\right) \leq u\left(\sqrt{s^2}\right) + \frac{u'\left(\sqrt{s^2}\right)}{2\sqrt{s^2}}(t^2 - s^2). \tag{3.11}$$

Since u is even, we have $u\left(\sqrt{t^2}\right) = u(|t|) = u(t)$, $u\left(\sqrt{s^2}\right) = u(|s|) = u(s)$. Then, introducing function w, relation (3.11) can be rewritten as

$$u\left(t\right) \leq u\left(s\right) + \frac{w(|s|)}{2}(t^2 - s^2), \tag{3.12}$$

or equivalently

$$u\left(t\right) \leq u\left(s\right) + u'(s)(t-s) + \frac{w(|s|)}{2}(t-s)^2 + \left(\frac{w(|s|)}{2}(t^2 - s^2) - u'(s)(t-s) - \frac{w(|s|)}{2}(t-s)^2\right). \tag{3.13}$$

We then need to simplify the term of (3.13) between parenthesis. Since u is even, its derivative u' is odd and we can deduce that $u'(|s|) = u'(s)\frac{|s|}{s}$ $(s \neq 0)$, i.e. $u'(s) = w(|s|)s$. This gives:

$$\frac{w(|s|)}{2}(t^2 - s^2) - u'(s)(t-s) - \frac{w(|s|)}{2}(t-s)^2 = \frac{w(|s|)}{2}\left((t^2 - s^2) - 2(ts - s^2) - (t-s)^2\right) = 0. \tag{3.14}$$

(3.10) is thus verified for all $(s,t) \in \mathbb{R}^* \times \mathbb{R}$. As w admits a continuous extension in 0, the latter remains true for $s = 0$ passing to the limit. Finally, since u is increasing, u' is positive and so it is for w. Mapping $w(|.|)$ is thus positive and straightforwardly uniformly bounded, every time this is the case for w. □

In our context, Lemma 3.1, initially derived from [124, Lemma 8.3], is a key result which enables to build majorization mappings of f when the latter possesses a half-quadratic structure [108, 109, 5]. We illustrate its use in the academical Example 3.3 below. Such a technique is used extensively in Chapters 4,5 and 7 for the construction of quadratic tangent majorization approximations for our test functions.

Example 3.3. *Let $\mathcal{H} = \mathbb{R}^N$ $(N \geq 1)$ and consider $f : \mathbb{R}^N \to \mathbb{R}$ of the half-quadratic form of [5] for which we aim to build a quadratic tangent majorization approximation i.e.*

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) \text{ with } f_1(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{H}\boldsymbol{x} - \boldsymbol{z}\|^2 \text{ and } f_2(\boldsymbol{x}) = \sum_{c=1}^{C} u([\boldsymbol{V}\boldsymbol{x}]_p), \tag{3.15}$$

considering $\boldsymbol{z} \in \mathbb{R}^M$ $(M \geq 1)$, $\boldsymbol{H} \in \mathbb{R}^{M \times N}, \boldsymbol{V} \in \mathbb{R}^{C \times N}$ $(C \geq 1)$ and $u : \mathbb{R} \to \mathbb{R}$ a differentiable function.

First of all, using academic algebra, f_1 can be rewritten so as to be $(\boldsymbol{H}^\top\boldsymbol{H}, \boldsymbol{H}\boldsymbol{z}, \|\boldsymbol{z}\|^2)$-quadratic. Basically, it is a $\boldsymbol{H}^\top\boldsymbol{H}$-quadratic tangent majorization approximation of itself. Quadratic tangent majorization property being preserved by summation, it remains to build a majorization mapping for f_2. Of course, such a construction depends on the available information on u. Let us display two important cases.

- *([108]): If u is β-Lipschitz continuous gradient($\beta > 0$), the descent lemma (Proposition 2.4) ensures that f_2 admits $\beta^{-1}\boldsymbol{V}^\top\boldsymbol{V}$ as a constant majorization mapping. The resulting majorization mapping for f is then Geman-Yang one*

$$\boldsymbol{x} \in \mathbb{R}^N \mapsto \boldsymbol{A}_{GY}(\boldsymbol{x}) := \boldsymbol{H}^\top\boldsymbol{H} + \beta^{-1}\boldsymbol{V}^\top\boldsymbol{V}. \tag{3.16}$$

- *([109]): If u verifies all conditions of Lemma 3.1, the latter allows to choose non constant majorization mapping $\boldsymbol{x} \in \mathbb{R}^N \mapsto \boldsymbol{V}^\top \operatorname{Diag}_{1\leq c\leq C}\left(\frac{u'(|[\boldsymbol{V}\boldsymbol{x}]_c|)}{|[\boldsymbol{V}\boldsymbol{x}]_c|}\right)\boldsymbol{V}$ for f_2. The resulting majorization mapping for f is then the Geman-Reynolds one*

$$\boldsymbol{x} \in \mathbb{R}^N \mapsto \boldsymbol{A}_{GR}(\boldsymbol{x}) := \boldsymbol{H}^\top\boldsymbol{H} + \boldsymbol{V}^\top \operatorname*{Diag}_{1\leq c\leq C}\left(\frac{u'(|[\boldsymbol{V}\boldsymbol{x}]_c|)}{|[\boldsymbol{V}\boldsymbol{x}]_c|}\right)\boldsymbol{V}. \tag{3.17}$$

Given the diversity of differentiable optimization problems, many other construction techniques have been developed recently. Typically, [102] proposed a generalization of Geman-Yang/Geman-Reynolds constructions in the complex framework notably in the case where f_1 (keeping notations of Example 3.3) is non-necessary quadratic. At the same time, [64] promoted quadratic majorization techniques for functions f of the form of f_2.

3.5 Subspace strategies

Once the methods for constructing quadratic tangent majorization approximations have been presented, the implementation of sufficiently accurate QMM algorithms, i.e. with a non-necessary trivial majorization mapping, becomes possible for a large class of functions. However, in a large-scale optimization context, such an implementation is usually easily compromised by the dimension of the problem under consideration. One way improvement, almost systematically used today in the MM field for this kind of setting, relies on a subspace technique.

3.5.1 Limitations of the QMM algorithm

Considering $\boldsymbol{A}$ as a definite positive majorization mapping of f, the associated scheme (3.5) possesses a relatively simple structure but requires to be able to compute the inverse of operator $\boldsymbol{A}_k := \boldsymbol{A}(\boldsymbol{x}_k)$ for every iteration $k \in \mathbb{N}$. The degree of difficulty of such an operation being strongly dependents on $\mathcal{H}$ dimension, the latter shall not be recommendable and is even impossible in a large-scale framework. Contrary to Quasi-Newton method (see 2.2.1.2), one cannot generally express sequence $\left(\boldsymbol{A}_k^{-1}\right)_{k\in\mathbb{N}}$ under a recursive process (as it is typically the case for SR1 or BFGS methods [178]) to the extent that every $\boldsymbol{A}_k$ ($k \in \mathbb{N}$) here strictly results of the evaluation of mapping $\boldsymbol{A}$ at a certain point of $\mathcal{H}$. Moreover, since the construction of a quadratic tangent majorization surrogates may also require the use of complex analytical tools, there is nothing to suggest that the structure of the $\boldsymbol{A}_k$ operator is any simpler and that it will therefore be relatively easy to invert.

3.5.2 Subspace Quadratic MM (SQMM) scheme

The necessity of computing the inverse of operator $\boldsymbol{A}_k$ at any iteration $k \in \mathbb{N}$ comes from the fact we aim to minimize, on the whole space $\mathcal{H}$, the corresponding quadratic tangent majorization approximation $h_q(., \boldsymbol{x}_k)$. Logically, the larger the dimension of $\mathcal{H}$, the greater the number of directions to investigate.

One way to overcome such an issue consists in adopting a so-called subspace acceleration strategy; the new iterate $\boldsymbol{x}_{k+1}$ ($k \in \mathbb{N}$) is chosen so as to minimize $h_q(., \boldsymbol{x}_k)$ no longer on the whole space but along M_k, a fixed number of directions $\boldsymbol{d}_k^1, \ldots, \boldsymbol{d}_k^{M_k} \in \mathcal{H}$. Starting from $\boldsymbol{x}_k$, the next iterate $\boldsymbol{x}_{k+1}$ is thus of the form of $\boldsymbol{x}_k + \boldsymbol{d}_k$ where $\boldsymbol{d}_k$ lies in subspace $\mathcal{V}_k := \mathrm{Vect}(\boldsymbol{d}_k^1, \ldots, \boldsymbol{d}_k^{M_k})$. With $\dim(\mathcal{V}_k) \leq M_k = \dim(\boldsymbol{R}^{M_k})$, we consider a linear transformation $\boldsymbol{D}_k : \mathbb{R}^{M_k} \to \mathcal{H}$ for which $\mathrm{im}(\boldsymbol{D}_k) = \mathcal{V}_k$ and the search for $\boldsymbol{d}_k$ amounts to finding a vector $\boldsymbol{u}_k \in \mathbb{R}^{M_k}$ verifying $\boldsymbol{d}_k = \boldsymbol{D}_k \boldsymbol{u}_k$. The incorporation of a subspace research into the QMM update (3.5) thus leads to the Subspace QMM (SQMM) scheme:

$$
\begin{aligned}
(\forall k \in \mathbb{N}) \quad & \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{u}_k, \qquad (3.18)\\
& \text{with } \boldsymbol{u}_k \in \underset{\boldsymbol{u} \in \mathbb{R}^{M_k}}{\arg\min}\, h_q(\boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{u}, \boldsymbol{x}_k) := f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{D}_k \boldsymbol{u} \rangle + \frac{1}{2} \langle \boldsymbol{A}_k \boldsymbol{D}_k \boldsymbol{u}, \boldsymbol{D}_k \boldsymbol{u} \rangle.
\end{aligned}
$$

The following proposition ensures the latter scheme is well-defined.

Proposition 3.4. *Let $\boldsymbol{A} : \mathcal{H} \to \mathcal{S}^{++}(\mathcal{H})$ be a uniformly bounded mapping and assume that f admits $h_q : \mathcal{H}^2 \to \mathcal{H}$, as a $\boldsymbol{A}$-quadratic tangent majorization approximation. Let also $M \geq 1$ and consider $\boldsymbol{D} : \mathbb{R}^M \to \mathcal{H}$ a linear operator. Then, for any $\boldsymbol{x} \in \mathcal{H}$, the set $\underset{\boldsymbol{u} \in \mathbb{R}^M}{\arg\min}\, h_q(\boldsymbol{x} + \boldsymbol{D}\boldsymbol{u}, \boldsymbol{x})$ is not empty.*

Proof. Let $\boldsymbol{x}$ be a vector of $\mathcal{H}$. We denote $E_{\boldsymbol{x}} := \left\{ h_q(\boldsymbol{x} + \boldsymbol{D}\boldsymbol{u}, \boldsymbol{x}) \mid \boldsymbol{u} \in \mathbb{R}^M \right\}$ and $\mu_{\boldsymbol{x}} := \inf(E_{\boldsymbol{x}}) \in [-\infty, +\infty)$ ($E_{\boldsymbol{x}}$ is obviously non-empty).

Following Definition 3.3, $\boldsymbol{y} \in \mathcal{H} \mapsto h_q(\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{x})$ is $(\boldsymbol{A}(\boldsymbol{x}), \nabla f(\boldsymbol{x}), f(\boldsymbol{x}))$-quadratic with $\boldsymbol{A}(\boldsymbol{x}) \in \mathcal{S}^{++}(\mathcal{H})$ and Proposition 3.1 (ii) ensures the coercivity of such a function. Set $\{h_q(\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{x}) \mid \boldsymbol{y} \in \mathcal{H}\}$ thus admits a minimum (Theorem 2.3). To the extent that $E_{\boldsymbol{x}} \subset \{h_q(\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{x}) \mid \boldsymbol{y} \in \mathcal{H}\}$, $E_{\boldsymbol{x}}$ is bounded below and thus $\mu_{\boldsymbol{x}} (:= \inf(E_{\boldsymbol{x}})) \in \mathbb{R}$. In addition, by inf characterization, there exists a sequence $(\boldsymbol{v}_k)_{k \in \mathbb{N}}$ of vector lying in $\mathbb{R}^M$ s.t.

$$
h_q(\boldsymbol{x} + \boldsymbol{D}\boldsymbol{v}_k, \boldsymbol{x}) \underset{k \to +\infty}{\longrightarrow} \mu_{\boldsymbol{x}}. \qquad (3.19)
$$

Decomposition $\mathcal{H} = \ker(\boldsymbol{D}) \oplus \ker(\boldsymbol{D})^{\perp}$ being always valid in finite dimension, there also exist two sequences $(\boldsymbol{v}_k^1)_{k \in \mathbb{N}}$, lying in $\ker(\boldsymbol{D})$, and $(\boldsymbol{v}_k^2)_{k \in \mathbb{N}}$, lying in $\ker(\boldsymbol{D})^{\perp}$, s.t. $\boldsymbol{v}_k = \boldsymbol{v}_k^1 + \boldsymbol{v}_k^2$ for all $k \in \mathbb{N}$. Considering this, behaviour (3.19) can be rewritten as:

$$
h_q(\boldsymbol{x} + \boldsymbol{D}\boldsymbol{v}_k^2, \boldsymbol{x}) \underset{k \to +\infty}{\longrightarrow} \mu_{\boldsymbol{x}}. \qquad (3.20)
$$

Denoting $\boldsymbol{D}$ the adjoint operator of $\boldsymbol{D}$, with $\boldsymbol{v}_k^2 \in \ker(\boldsymbol{D})^{\perp} = \ker(\boldsymbol{D}^* \boldsymbol{D})^{\perp}$ for any $k \in \mathbb{N}$, the Courant-Fischer minoration leads to

$$
(\forall k \in \mathbb{N}) \quad \|\boldsymbol{D}\boldsymbol{v}_k^2\|^2 = \langle \boldsymbol{D}^* \boldsymbol{D} \boldsymbol{v}_k^2, \boldsymbol{v}_k^2 \rangle \geq \lambda_{min}^+(\boldsymbol{D}^* \boldsymbol{D}) \|\boldsymbol{v}_k^2\|^2, \qquad (3.21)
$$

where $\lambda^+_{min}(\boldsymbol{D}^*\boldsymbol{D})$ denotes the smallest positive eigenvalue of the (positive) bounded self-adjoint operator $\boldsymbol{D}^*\boldsymbol{D}$. Moreover, coercivity of $\boldsymbol{y} \mapsto h_q(\boldsymbol{x}+\boldsymbol{y}, \boldsymbol{x})$ guarantees that $(\|\boldsymbol{D}\boldsymbol{v}^2_k\|)_{k\in\mathbb{N}}$ does not converge to $+\infty$ (the contrary would contradict (3.20) since $\mu_{\boldsymbol{x}} \in \mathbb{R}$). As a consequence, there exist $B > 0$ as well as a subsequence $\left(\boldsymbol{v}^2_{\psi_1(k)}\right)_{k\in\mathbb{N}}$ for which $\left(\left\|\boldsymbol{D}\boldsymbol{v}^2_{\psi_1(k)}\right\|\right)_{k\in\mathbb{N}}$ is uniformly bounded by B. Inequality (3.21) then leads to

$$(\forall k \in \mathbb{N}) \quad \left\|\boldsymbol{v}^2_{\psi_1(k)}\right\|^2 \leq \left(\lambda^+_{min}(\boldsymbol{D}^*\boldsymbol{D})\right)^{-1} \|\boldsymbol{D}\boldsymbol{v}^2_{\psi_1(k)}\|^2 \leq \left(\lambda^+_{min}(\boldsymbol{D}^*\boldsymbol{D})\right)^{-1} B^2, \tag{3.22}$$

and thus guarantees the boundedness of $\left(\boldsymbol{v}^2_{\psi_1(k)}\right)_{k\in\mathbb{N}}$. $\mathbb{R}^M$ being of finite dimension, we can finally extract a convergent subsequence $\left(\boldsymbol{v}^2_{(\psi_1\circ\psi_2)(k)}\right)_{k\in\mathbb{N}}$. Denoting $\boldsymbol{v}^*$ its attached limit, (3.20) and continuities of $\boldsymbol{y} \mapsto h_q(\boldsymbol{x}+\boldsymbol{D}\boldsymbol{y}, \boldsymbol{x})$ finally conduct to $h_q(\boldsymbol{x}+\boldsymbol{D}\boldsymbol{v}^*, \boldsymbol{x}) = \mu_{\boldsymbol{x}}$ and so $\underset{\boldsymbol{u}\in\mathbb{R}^M}{\arg\min}\, h_q(\boldsymbol{x}+\boldsymbol{D}\boldsymbol{u}, \boldsymbol{x})$ contains $\boldsymbol{v}^*$. □

In particular, considering $\boldsymbol{D}^*_k$, the adjoint of $\boldsymbol{D}_k$ ($k \in \mathbb{N}$), elementary algebraic manipulation are sufficient to prove that function $\boldsymbol{u} :\in \mathbb{R}^{M_k} \mapsto h_q(\boldsymbol{x}_k+\boldsymbol{D}_k\boldsymbol{u}, \boldsymbol{x}_k)$ is basically $(\boldsymbol{D}^*_k\boldsymbol{A}_k\boldsymbol{D}_k, \boldsymbol{D}^*_k\nabla f(\boldsymbol{x}_k), f(\boldsymbol{x}_k))$-quadratic and more specifically $\boldsymbol{D}^*_k\boldsymbol{A}_k\boldsymbol{D}_k$ has the advantage of always being positive. Considering this, let us make the distinction between two cases.

- If operator $\boldsymbol{D}^*_k\boldsymbol{A}_k\boldsymbol{D}_k$ is invertible, then the latter is definite positive and, in virtue of Proposition 3.1 (ii), $\boldsymbol{u} :\in \mathbb{R}^{M_k} \mapsto h_q(\boldsymbol{x}_k + \boldsymbol{D}_k\boldsymbol{u}, \boldsymbol{x}_k)$ thus admits a unique minimizer whose closed-form is

$$\boldsymbol{u}_k = (\boldsymbol{D}^*_k\boldsymbol{A}_k\boldsymbol{D}_k)^{-1}\, \boldsymbol{D}^*_k\nabla f(\boldsymbol{x}_k). \tag{3.23}$$

  Note that, it is necessary for directions $\boldsymbol{d}^1_k, \ldots, \boldsymbol{d}^{M_k}_k$ to be linearly independant for such a situation to occur.

- If $\boldsymbol{D}^*_k\boldsymbol{A}_k\boldsymbol{D}_k$ is not invertible, choice

$$\boldsymbol{u}_k = (\boldsymbol{D}^*_k\boldsymbol{A}_k\boldsymbol{D}_k)^{\dagger}\, \boldsymbol{D}^*_k\nabla f(\boldsymbol{x}_k), \tag{3.24}$$

  is, by default, retained. Notation † corresponds here to the Moore-Penrose inversion.

In both cases, the inversion step of $\boldsymbol{A}_k : \mathcal{H} \to \mathcal{H}$, of expected complexity close to $\dim(\mathcal{H})^3$ (by analogy with the real space), is replaced by the inversion (or the pseudo-inverse computation) of operator $\boldsymbol{D}^*_k\boldsymbol{A}_k\boldsymbol{D}_k : \mathbb{R}^{M_k} \to \mathbb{R}^{M_k}$ of complexity of order M^3_k and so much less costly.

3.5.3 Choice of subspace directions

As seen, the great advantage of the subspace method lies in its ability to highly reduce the complexity of the update step when M_k is chosen to be very small compared to $\dim(\mathcal{H})$. Remark that taking $M_k = \dim(\mathcal{H})$ and directions $\boldsymbol{d}_k, \ldots, \boldsymbol{d}^{M_k}_k$ so as to form a basis of $\mathcal{H}$ makes (3.18) equivalent to (3.5).

Historically and strictly speaking, [164] is the first article mentioning subspace strategies for gradient descent schemes. In the latter, the authors built subspace $\mathcal{V}_k$ ($k \in \mathbb{N}$) by only retaining two components (i.e $M_k = 2$); $\boldsymbol{d}^1_k = -\nabla f(\boldsymbol{x}_k)$ to fully capture the first order information and $\boldsymbol{d}^2_k = \boldsymbol{x}_k - \boldsymbol{x}_{k-1}$

in a way to preserve those brought by the past iterate $\boldsymbol{x}_{k-1}$ (with convention $\boldsymbol{x}_{-1} = 0$ if k is negative). In the real case ($\mathcal{H} = \mathbb{R}^N$ $(N \geq 1)$) for which $\boldsymbol{D}_k$ can basically be assimilated to its representative matrix in the canonical basis, $\boldsymbol{D}_k = [-\nabla f(\boldsymbol{x}), \boldsymbol{x}_k - \boldsymbol{x}_{k-1}]$ is usually find under the name of the memory gradient matrix [58]. Let us also underline the fact that the subspace technique using these two directions appears naturally for some classical algorithms as the non-linear conjugate gradient one [55].

Other choices of search subspaces have also been proposed. This typically includes some generalized versions of memory acceleration as those of [74], collecting more differences of past iterates in the hope of preserving even more information. Most of the time, the directions are chosen by reasoning by analogy with already existing schemes in the literature or by interpreting them as naturally accelerated [242]. As an example, the Fletcher-Reeves conjugate gradient algorithm [101] (see 2.2.1.3) is simply equivalent to a memory gradient strategy incorporated in the steepest descent procedure (see 2.2.1.1) in the case where f is quadratic [46]. The incorporation of subspace techniques into the QMM algorithm was initially proposed by Chouzenoux *et al* in [58].

3.6 Existing asymptotical results

As mentioned, the theoretical study of the generic scheme (3.2) was conducted in [127]. Without any assumption on the structure of the quadratic tangent majorization surrogates, the results highlighted in [127] are mainly of a topological nature and similar as those of section 2.3.3. In this section, we mainly exhibit an overview of the existing assumptions and theorems promoting the existence of simple descent inequalities (see Definition 2.2) for QMM schemes. The structure of the resulting quadratic tangent majorization approximations allowing to obtain recurrence relations on $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ quite easily.

3.6.1 Descent inequality for QMM scheme with stepsize

In the literature, the QMM (3.5) procedure is generally studied under a slightly more general form with the advantage of incorporating an additional stepsize sequence $(\alpha_k)_{k\in\mathbb{N}}$:

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{A}_k^{-1} \nabla f(\boldsymbol{x}_k). \tag{3.25}$$

The following theorem stipulates the existence of a simple descent condition for scheme (3.25) every time the stepsize is wisely chosen:

Theorem 3.1. *(Quadratic MM descent condition) Let $\boldsymbol{A} : \mathcal{H} \to \mathcal{S}^{++}(\mathcal{H})$ be a majorization mapping of f and denote by $\overline{\nu}$ the constant attached to its uniform upper-bound (see notations at the beginning of subsection 3.3.1). Also assume that there exist $\underline{\alpha}, \overline{\alpha} \in (0, 2)$ for which $\underline{\alpha} \leq \alpha_k \leq 2 - \overline{\alpha}$ for every $k \in \mathbb{N}$. Then, sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$, satisfies the following simple descent condition,*

$$(\forall k \in \mathbb{N}) \quad f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \underline{\alpha}\overline{\nu}^{-1} \left(1 - \frac{\overline{\alpha}}{2}\right) \|\nabla f(\boldsymbol{x}_k)\|^2. \tag{3.26}$$

Moreover, $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ admits a finite limit and $(\nabla f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to zero if f is bounded-below.

Proof. Let us set $k \in \mathbb{N}$, we first use the fact that f admits an $\boldsymbol{A}$-quadratic tangent majorization approximation:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{1}{2} \langle \boldsymbol{A}_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle. \tag{3.27}$$

With $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\alpha_k \boldsymbol{A}_k^{-1} \nabla F(\boldsymbol{x}_k)$ and $\boldsymbol{A}_k$ self-adjoint, (3.27) can be turned to:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_{k+1}) - \alpha_k \left(1 - \frac{\alpha_k}{2}\right) \langle \boldsymbol{A}_k^{-1} \nabla f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k) \rangle. \tag{3.28}$$

Since $\boldsymbol{A}_k$ is definite positive and $\underline{\nu}$-bounded, passing to the inverse ensures that $\boldsymbol{A}_k^{-1}$ is definite positive and bounded-below by $\underline{\nu}^{-1}$ and so $\left\langle \boldsymbol{A}_k^{-1} \nabla f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k) \right\rangle \geq \overline{\nu}^{-1} \|\nabla f(\boldsymbol{x}_k)\|^2$. As $\alpha_k \in [\underline{\alpha}, \overline{\alpha}] \subset (0, 2)$, we deduce that the residual term of (3.28) is positive and

$$\alpha_k \left(1 - \frac{\alpha_k}{2}\right) \langle \boldsymbol{A}_k^{-1} \nabla F(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k) \rangle \geq \underline{\alpha} \overline{\nu}^{-1} \left(1 - \frac{\overline{\alpha}}{2}\right) \|\nabla f(\boldsymbol{x}_k)\|^2, \tag{3.29}$$

which directly gives (3.26). Adding the fact that f is lower-bounded classically ensures that $(f(\boldsymbol{x}_k))_{k \in \mathbb{N}}$ converges to a finite limite f_∞ (see 2.3.3). With a residual of the form (2.25) (taking $n = 2$ and $(\gamma_k)_{k \in \mathbb{N}}$ positively constant), $\sum_{k=0}^{+\infty} \|\nabla f(\boldsymbol{x}_k)\|^2 < +\infty$ follows and finally ensures the convergence of $(\nabla f(\boldsymbol{x}_k))_{k \in \mathbb{N}}$ to zero. □

The arguments invoked in the previous proof can notably be found in the convergence analysis of [5]. The reader may also consult [55, Chapter 4] to have an extension of Theorem 3.1 with a vectorized version of the stepsize.

3.6.2 Descent inequality for SQMM scheme

Let us start this section by a technical definition which will be useful to conduct our convergence proof.

Definition 3.4. *(Uniformed gradient related directions) Let $(\boldsymbol{x}_k)_{k \in \mathbb{N}}$ be a sequence of $\mathcal{H}$*

(i) *A sequence of directions $(\boldsymbol{d}_k)_{k \in \mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ is said to be uniformly gradient related to $(\boldsymbol{x}_k)_{k \in \mathbb{N}}$ (regarding f) if there exist $c_1, c_2 > 0$ s.t.*

$$(\forall k \in \mathbb{N}) \quad \begin{cases} \boldsymbol{d}_k = 0 \ \text{ if and only if } \nabla f(\boldsymbol{x}_k) = 0, \\ \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle \leq -c_1 \|\nabla f(\boldsymbol{x}_k)\|^2, \\ \|\boldsymbol{d}_k\| \leq c_2 \|\nabla f(\boldsymbol{x}_k)\|. \end{cases} \tag{3.30}$$

(ii) *A sequence of linear operator $(\boldsymbol{D}_k)_{k \in \mathbb{N}}$ is said to be uniformly gradient related to $(\boldsymbol{x}_k)_{k \in \mathbb{N}}$ (regarding f) if there exists a sequence $(\boldsymbol{d}_k)_{k \in \mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ uniformly gradient related to $(\boldsymbol{x}_k)_{k \in \mathbb{N}}$ and s.t. $\boldsymbol{d}_k \in \mathrm{im}(\boldsymbol{D}_k)$ for any $k \in \mathbb{N}$.*

Theorem 3.2. *(SQMM descent condition) Let $\boldsymbol{A} : \mathcal{H} \to \mathcal{S}^{++}(\mathcal{H})$ be a majorization mapping of f and denote once again by $\overline{\nu}$, the constant attached to its uniform upper-bound. Consider also $(\boldsymbol{x}_k)_{k \in \mathbb{N}}$ a process following the SQMM update rule* (3.18) *(associated to $\boldsymbol{A}$) and for which subspace sequence*

$(\boldsymbol{D}_k)_{k\in\mathbb{N}}$ is uniformly gradient related to $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ given constants $c_1, c_2 > 0$. Then, $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ verifies the simple descent condition:

$$(\forall k \in \mathbb{N}) \quad f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{1}{2\overline{\nu}} \left(\frac{c_1}{c_2}\right)^2 \|\nabla f(\boldsymbol{x}_k)\|^2. \tag{3.31}$$

Moreover, $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ admits a finite limit and $(\nabla f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to zero, if is bounded-below.

Proof. Let us set $k \in \mathbb{N}$. According to Definition 3.4, there exist $(\boldsymbol{d}_k, \hat{\boldsymbol{u}}_k) \in \mathcal{H}^* \times \mathbb{R}^{M_k}$ for which $\boldsymbol{D}_k\hat{\boldsymbol{u}}_k = \boldsymbol{d}_k$ and (3.30) is satisfied. Moreover, as $\boldsymbol{u}_k$ minimize $\boldsymbol{u} \in \mathbb{R}^{M_k} \mapsto h_q(\boldsymbol{x} + \boldsymbol{D}_k\boldsymbol{u}, \boldsymbol{x}_k)$, it follows that, for any $t \in \mathbb{R}$,

$$f(\boldsymbol{x}_{k+1}) = h_q(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) = h_q(\boldsymbol{x}_k + \boldsymbol{D}_k\boldsymbol{u}_k, \boldsymbol{x}_k) \leq h_q(\boldsymbol{x}_k + \boldsymbol{D}_k\ t\ \hat{\boldsymbol{u}}_k, \boldsymbol{x}_k) = h_q(\boldsymbol{x}_k + t\boldsymbol{d}_k, \boldsymbol{x}_k). \tag{3.32}$$

If $\boldsymbol{d}_k = 0$ then $h_q(\boldsymbol{x}_k + t\boldsymbol{d}_k, \boldsymbol{x}_k) = f(\boldsymbol{x}_k)$ and inequality (3.31) holds.

For the rest of the proof, let us assume that $\boldsymbol{d}_k \neq 0$, (and so $\nabla f(\boldsymbol{x}_k) \neq 0$). The fact that $\boldsymbol{A}_k$ is definite positive with $\boldsymbol{d}_k \neq 0$ ensures that function $t \in \mathbb{R} \mapsto h_q(\boldsymbol{x}_k + t\boldsymbol{d}_k, \boldsymbol{x}_k) := f(\boldsymbol{x}_k) + t\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle + \frac{t^2}{2}\langle \boldsymbol{A}_k\boldsymbol{d}_k, \boldsymbol{d}_k\rangle$ is second order polynomial whose dominant term $\langle \boldsymbol{A}_k\boldsymbol{d}_k, \boldsymbol{d}_k\rangle$ is positive. It thus admits a unique minimizer t_{min} which can be expressed under the closed form

$$t_{min} = -\frac{\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle}{\langle \boldsymbol{A}_k\boldsymbol{d}_k, \boldsymbol{d}_k\rangle}. \tag{3.33}$$

Since (3.32) is verified for any $t \in \mathbb{R}$, it is satisfied at $t = t_{min}$, and

$$f(\boldsymbol{x}_{k+1}) \leq h_q(\boldsymbol{x}_k + t_{min}\boldsymbol{d}_k, \boldsymbol{x}_k) = f(\boldsymbol{x}_k) - \frac{\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle^2}{2\langle \boldsymbol{A}_k\boldsymbol{d}_k, \boldsymbol{d}_k\rangle}. \tag{3.34}$$

With $\boldsymbol{A}_k$ is spectrum-bounded by $\overline{\nu}$, $\langle \boldsymbol{A}_k\boldsymbol{d}_k, \boldsymbol{d}_k\rangle \leq \overline{\nu}\|\boldsymbol{d}_k\|^2$ and we deduce that

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle^2}{2\overline{\nu}\|\boldsymbol{d}_k\|^2}. \tag{3.35}$$

Finally, the use of (3.30) leads to:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{c_1^2\|\nabla f(\boldsymbol{x}_k)\|^4}{2\overline{\nu}c_2^2\|\nabla f(\boldsymbol{x}_k)\|^2} = f(\boldsymbol{x}_k) - \frac{1}{2\overline{\nu}} \left(\frac{c_1}{c_2}\right)^2 \|\nabla f(\boldsymbol{x}_k)\|^2, \tag{3.36}$$

and so to inequality (3.31). The end of the proof is similar to those of Theorem 3.2. □

The tools we used to conduct our proof was initially proposed in [58]. In the latter and in the case where $\mathcal{H} = \mathbb{R}^N$ ($N \geq 1$), the authors considered the specific case where the gradient related direction is the first column of the the subspace matrix $\boldsymbol{D}_k$ ($k \in \mathbb{N}$). Moreover, their framework differs somewhat from ours to the extent they allow the incorporation of a stepsize research in their update rule [55].

3.6.3 *Bridging the gap with global convergence*

The strategies of proof we applied for Theorem 3.1-3.2 remain in fact relatively close to those invoked to study usual descent gradient schemes [178]. The main difference relies on the fact that we use the quadratic tangent majorization property (3.4) as a substitute to classical descent lemma 2.4. Of course, the access to extra curvatures properties on f would logically allow finer results. For instance, the coercivity of f typically ensures boundedness of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ and thus allows the use of Proposition 2.2 for better investigations.

In a non-convex framework, the construction of proof strategies based on KL theory (see Chapter 2 section 2.5.2) have been considered over the past few years. The use of such tools in a way to establish the global convergence of a QMM type scheme was made for the first time in [59]. In the latter, the authors conducted a similar line of reasoning as those we exhibited in Example 2.2. Their convergence investigation remains rather adapted to the situation where f is semi-algebraic insofar as the KL function is specifically taken as a power of $\theta \in (0, 1)$. Another interest of the KL theory in this context may lie in its ability to construct convergence rates. Although their setting is somewhat different from the QMM one, [66] actually highlighted the fact that θ exponent may play a crucial role in the estimation convergence speed of the iterates. Up to our knowledge, only [61] was able to exhibit a convergence rate for QMM schemes.

3.7 Conclusion

The presented chapter can thus be considered as a specific literature review, on the central subject of this thesis, i.e. the QMM methods. Our main goal being here the setting up of a sufficiently clear theoretical framework which will constitute, in all the continuation, an essential working base to really determine the interest of the new methods which we present in Chapters 4, 5 and finally in Chapter 7 once our stochastic framework is introduced.

Chapter 4

Convergence analysis of block majorize-minimize subspace approach

In this chapter, we consider the minimization of a differentiable Lipschitz gradient but non necessarily convex, function F defined on $\mathbb{R}^N$. We propose an accelerated gradient descent approach which combines three strategies, namely (i) a variable metric derived from the majorization-minimization principle ; (ii) a subspace strategy incorporating information from the past iterates ; (iii) a block alternating update. Under the assumption that F satisfies the Kurdyka-Łojasiewicz property, we give conditions under which the sequence generated by the resulting block majorize-minimize subspace algorithm converges to a critical point of the objective function, and we exhibit convergence rates for its iterates.

This work is based on our article: E. Chouzenoux and J-B. Fest, *Convergence analysis of block majorize-minimize subspace approach*, that we submitted to Optimization Letters in 2022 (now under minor revisions).

Contents

4.1 Introduction

Our work focuses on the resolution of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \quad F(\boldsymbol{x}), \tag{4.1}$$

with $F : \mathbb{R}^N \to \mathbb{R}$ is a differentiable Lipschitz gradient function which is not assumed to be convex. Instead, we address the case when F satisfies the Kurdyka-Łojasiewicz (KL) inequality [141, 25].

In the case of large scale optimization problems, one major concern is to find an optimization algorithm able to deliver reliable numerical solutions in a reasonable time. Numerous works have been devoted to accelerate the first order gradient descent technique. These methods aim to increase the convergence rate while preserving theoretical guarantees and limited computational cost/memory burden per iteration. Three main families of acceleration strategies can be distinguished in the literature. The first approach, adopted for example in the well-known L-BFGS [156] and non-linear conjugate gradient [101] methods, relies on subspace acceleration [243, 231]. The convergence rate is improved by using information from past iterates for the construction of new estimates. Another efficient way to accelerate the convergence of a minimization algorithm is based on a variable metric (i.e., preconditioning) strategy [28, 103]. The underlying metric is modified at each iteration thanks to a preconditioning matrix, which may incorporate structural second-order information about the function to minimize. The third technique to limit the dependence of an optimization algorithm on the dimension of the problem, is to adopt a block alternating scheme where, at each iteration, only a subset of the variables are updated [27].

Among various choices for preconditioning first-order methods, an important class of techniques rely on the principle of Majorization-Minimization (MM) [223, 246]. At each iteration, a quadratic convex surrogate function majorizing F is constructed. The inverse of its curvature (i.e., Hessian) matrix then serves to define a weighted Euclidean metric used for updating the next iterate. This idea is at the core of the half-quadratic algorithm [203, 5] for image restoration. It has also been exploited in [64] to build an accelerated proximal gradient method for non smooth optimization, with guaranteed convergence of the iterates to a stationary point, in the non-convex case. The latter result has been then extended in [66], where block alternating updates are introduced. Block alternating MM approaches have also been explored in [122, 215, 128], although without established convergence guarantees on their iterates in the non-convex setting. MM metrics are also well suited to the construction of efficient subspace optimization methods [58, 59, 61, 223]. In [58], subspace acceleration is employed to reduce the complexity of an MM algorithm in large scale image processing problems, and in [59], convergence guarantees are obtained on the iterates under the KL assumption. This algorithm has recently been extended in [60] to the resolution of convex constrained optimization problems.

In this chapter, we propose to bridge the gap between the theoretical analysis from [64] and [58]. We introduce the Block MM Subspace (B2MS) algorithm to solve (4.1), that incorporates the three aforementioned catalyzing effects, namely (i) MM-based preconditioning, (ii) subspace acceleration, (iii) block alternating update. We show the convergence of its iterates to a critical point of F. We furthermore perform its convergence rate analysis, relying on the KL exponent properties from [8].

The chapter is organized as follows. Section 4.2 introduces the notation, the proposed B2MS algorithm, and the considered assumptions. Section 4.3 provides technical descent lemmas essential to

our analysis. Section 4.4 presents our main contribution, namely the proof of convergence for B2MS, and a study of its convergence rate. Section 4.5 discusses the advantages of B2MS acceleration features through an ablation study on a practical problem of sparse signal blind deconvolution. Section 4.6 concludes the work.

4.2 Block MM subspace algorithm

4.2.1 Notation

We consider the Euclidean space $\mathbb{R}^N$, endowed with the scalar product $\langle\cdot|\cdot\rangle$ and norm $\|\cdot\|$. $\boldsymbol{I}_N$ states for the identity matrix of $\mathbb{R}^N$. For any $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ symmetric definite positive (SDP), we also introduce the weighted norm $\|\cdot\|_{\boldsymbol{A}} = \sqrt{\langle\cdot|\boldsymbol{A}\cdot\rangle}$. Let $\mathcal{S} \subset \{1,\dots,N\} \triangleq [\![1,N]\!]$ with cardinal $|\mathcal{S}|$ and complementary set $\overline{\mathcal{S}} \triangleq [\![1,N]\!]/\mathcal{S}$. For all $\boldsymbol{x} = (x^n)_{n\in[\![1,N]\!]} \in \mathbb{R}^N$, we denote $\boldsymbol{x}^{(\mathcal{S})} \triangleq \left(\boldsymbol{x}^i\right)_{i\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$. Similarly, the restriction to block $\mathcal{S}$ of the gradient of F, at some $\boldsymbol{x} \in \mathbb{R}^N$ reads $\nabla F^{(\mathcal{S})}(\boldsymbol{x}) \in \mathbb{R}^{|\mathcal{S}|}$. For any $\boldsymbol{x} \in \mathbb{R}^N$, we finally introduce function $F^{(\mathcal{S})}(.\,,\boldsymbol{x}) : \boldsymbol{v} \in \mathbb{R}^{|S|} \mapsto F(\boldsymbol{u})$ where $\boldsymbol{u}^{(\mathcal{S})} = \boldsymbol{v}$ and $\boldsymbol{u}^{(\overline{\mathcal{S}})} = \boldsymbol{x}^{(\overline{\mathcal{S}})}$.

4.2.2 B2MS scheme

The proposed B2MS algorithm solves (4.1) through a block alternating minimization approach. Let $\mathbb{S}$ a family of $C \geq 1$ nonempty subsets of $[\![1,N]\!]$ (not necessarily disjoints). Let $\boldsymbol{x}_0 \in \mathbb{R}^N$. At every iteration $k \in \mathbb{N}$, the entries of the current iterate $\boldsymbol{x}_k$ within a selected block $\mathcal{S}_k \in \mathbb{S}$ are updated using one iteration of the MM subspace algorithm [58] on the restriction of F to the k-th block $F^{(\mathcal{S}_k)}(\cdot,\boldsymbol{x}_k)$. The entries of $\boldsymbol{x}_k$ within the complementary set $\overline{S}_k$ remain constant.

To implement the MM subspace update, we first build the following *quadratic majorization approximation* [223] of $F^{(\mathcal{S}_k)}(\cdot,\boldsymbol{x}_k)$ at $\boldsymbol{x}_k$,

$$(\forall \boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}_k|}) \quad Q^{(\mathcal{S}_k)}(\boldsymbol{v},\boldsymbol{x}_k) \triangleq F(\boldsymbol{x}_k) + \langle \nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k), \boldsymbol{v} - \boldsymbol{x}_k^{(\mathcal{S}_k)}\rangle + \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{x}_k^{(\mathcal{S}_k)}\|^2_{\boldsymbol{A}^{(\mathcal{S}_k)}(\boldsymbol{x}_k)}, \tag{4.2}$$

where $\boldsymbol{A}^{(S_k)}(\boldsymbol{x}_k) \in \mathbb{R}^{|\mathcal{S}_k|\times|\mathcal{S}_k|}$ is an SDP matrix such that:

$$(\forall \boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}_k|}) \quad F^{(\mathcal{S}_k)}(\boldsymbol{v},\boldsymbol{x}_k) \leq Q^{(S_k)}(\boldsymbol{v},\boldsymbol{x}_k). \tag{4.3}$$

Second, we choose a *subspace acceleration* matrix $\boldsymbol{D}_k \in \mathbb{R}^{|\mathcal{S}_k|\times M_k}$ [243]. The block update $\boldsymbol{x}_{k+1}^{(\mathcal{S}_k)}$ is then defined as a minimizer of $Q^{(\mathcal{S}_k)}(\cdot,\boldsymbol{x}_k)$ within the vectorial subspace spanned by the columns of $\boldsymbol{D}_k$. Iterating the above procedure in a block alternating fashion yields Algorithm (4.4):

$$\begin{array}{l}
\text{Initialize} \quad \boldsymbol{x}_0 \in \mathbb{R}^N. \\
\text{For} \quad k = 0,1,2,\dots \\
\left\lfloor \begin{array}{l}
\text{Choose } \mathcal{S}_k \in \mathbb{S} \text{ and } \boldsymbol{D}_k \in \mathbb{R}^{|\mathcal{S}_k|\times M_k} \\
\boldsymbol{u}_k \in \underset{\boldsymbol{u}\in\mathbb{R}^{M_k}}{\arg\min}\, Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_k^{(\mathcal{S}_k)} + \boldsymbol{D}_k\boldsymbol{u}, \boldsymbol{x}_k\right) \\
\boldsymbol{x}_{k+1}^{(\mathcal{S}_k)} = \boldsymbol{x}_k^{(\mathcal{S}_k)} + \boldsymbol{D}_k\boldsymbol{u}_k \\
\boldsymbol{x}_{k+1}^{(\overline{\mathcal{S}_k})} = \boldsymbol{x}_k^{(\overline{\mathcal{S}_k})}
\end{array} \right.
\end{array} \tag{B2MS}$$

If $\mathbb{S} = \{[\![1, N]\!]\}$ (so $C = 1$), we retrieve the MM subspace algorithm from [58, 59]. When $\boldsymbol{D}_k = \boldsymbol{I}_{|\mathcal{S}_k|}$, the above approach can be viewed as a particular case of the BSUM scheme [122, 195] using quadratic surrogates, or of the approach from [66] using a null proximal term.

4.2.3 Assumptions

Assumption 4.1.
Family $\mathbb{S}$ verifies $\bigcup_{\mathcal{S}\in\mathbb{S}} \mathcal{S} = [\![1, N]\!]$. Moreover, there exists $K \in \mathbb{N}^*$ such that, for all $k \in \mathbb{N}$, every $\mathcal{S} \in \mathbb{S}$ belongs to $\{\mathcal{S}_k, \ldots, \mathcal{S}_{k+K-1}\}$.

Assumption 4.2.
F is $\mathcal{C}^1$ and coercive on $\mathbb{R}^N$.

Assumption 4.3.
(i) For all $k \in \mathbb{N}$, $\boldsymbol{D}_k$ has full column rank.
(ii) There exists $(\gamma_0, \gamma_1) > 0$ such that, for all $k \in \mathbb{N}$,

$$(\boldsymbol{d}_k)^\top \nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k) \leq -\gamma_0 \|\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k)\|^2, \tag{4.5}$$

$$\|\boldsymbol{d}_k\| \leq \gamma_1 \|\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k)\|, \tag{4.6}$$

with $\boldsymbol{d}_k \in \mathbb{R}^{|\mathcal{S}_k|}$ the first column of $\boldsymbol{D}_k$.

Assumption 4.4.
(i) For every $k \in \mathbb{N}$, there exists an SDP matrix $\boldsymbol{A}^{(\mathcal{S}_k)}(\boldsymbol{x}_k)$ such that inequality (4.3) holds.
(ii) There exists $(\eta, \nu) > 0$ such that

$$(\forall k \in \mathbb{N}) \quad \eta \boldsymbol{I}_{|\mathcal{S}_k|} \preceq \boldsymbol{A}^{(\mathcal{S}_k)}(\boldsymbol{x}_k) \preceq \nu \boldsymbol{I}_{|\mathcal{S}_k|}. \tag{4.7}$$

Assumption 4.5.

F is Lipschitz differentiable on every bounded subset of $\mathbb{R}^N$. In other words, for each bounded $E \subset \mathbb{R}^N$, there exists $\beta(E) > 0$ such that

$$(\forall (\boldsymbol{x}, \boldsymbol{y}) \in E^2) \quad \|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \leq \beta(E) \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{4.8}$$

Assumption 4.6. *For $\xi \in \mathbb{R}$ and any bounded $E \subset \mathbb{R}^N$, there exists $(\kappa, \zeta, \theta) \in \mathbb{R}_+^* \times \mathbb{R}_+^* \times]0, 1[$ such that, for all $\boldsymbol{x} \in E$ with $|F(\boldsymbol{x}) - \xi| \leq \zeta$,*

$$\|\nabla F(\boldsymbol{x})\| \geq \kappa \, |F(\boldsymbol{x}) - \xi|^\theta . \tag{4.9}$$

Assumption 4.1, also adopted in [64], implies that every set of $\mathbb{S}$ is updated at least once during any K-length cycle. It is also known as quasi-cyclic or acyclic rule [128], the cyclic rule being a special case of it. Assumption 4.2 ensures the existence of a minimizer for F. Assumption 4.3(ii) is equivalent to imposing a gradient-related condition [19] on the first column of $\boldsymbol{D}_k$. This is satisfied by a large number of typical subspace acceleration matrices [58][Tab. I]. In particular, setting $\boldsymbol{D}_k =$

$[-\nabla F(\boldsymbol{x}_k)|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}] \in \mathbb{R}^{N\times 2}$ for $k \in \mathbb{N}^*$ yields the memory gradient subspace [164, 102] which has strong connections with the non-linear conjugate gradient method [46]. Assumption 4.4 is rather mild, and inherent to the well-posedness and stability of quadratic MM schemes [58, 64]. Assumption 4.5 is a standard smoothness assumption, also considered in [25]. Finally, Assumption 4.6 is usually refereed to as the KL inequality [141, 25], and arises from the literature of non-smooth analysis. It is satisfied by a large variety of functions, non necessarily convex, such as semi-algebraic or analytical functions, to name a few. Its use has become popular in the last decade, as it provides a key tool for establishing convergence of iterates for descent methods in the non convex setting [8, 66, 103].

4.3 Technical lemmas

This section presents technical lemmas that turn out to be essential to our convergence analysis.

Lemma 4.1. *Under Assumptions 4.1-4.2-4.3-4.4(i), $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to a finite limit and sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is bounded. Moreover, under Assumptions 4.1-4.2-4.3-4.4, the B2MS sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ satisfies:*

$$(\forall k \in \mathbb{N}) \quad F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k+1}) \geq \frac{\gamma_0^2}{2\gamma_1^2\nu}\|\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k)\|^2, \tag{4.10}$$

$$\sum_{k=0}^{+\infty} \|\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k)\|^2 < +\infty, \tag{4.11}$$

$$(\forall k \in \mathbb{N}) \quad \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \leq \frac{1}{\eta^2}\|\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k)\|^2, \tag{4.12}$$

$$(\forall k \in \mathbb{N}) \quad \|\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k)\|^2 \leq \frac{\gamma_1^2\nu^2}{\gamma_0^2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2. \tag{4.13}$$

Proof. Consider $k \in \mathbb{N}$. Since $\boldsymbol{x}_{k+1}$ is the concatenation of $\boldsymbol{x}_{k+1}^{(\mathcal{S}_k)}$ and $\boldsymbol{x}_k^{(\overline{\mathcal{S}_k})}$, the use of the majorizing inequality (4.3) (Assumption 4.4(i)) gives,

$$F(\boldsymbol{x}_{k+1}) = F^{(\mathcal{S}_k)}(\boldsymbol{x}_{k+1}^{(\mathcal{S}_k)}, \boldsymbol{x}_k) \leq Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_{k+1}^{(\mathcal{S}_k)}, \boldsymbol{x}_k\right) = Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_k^{(\mathcal{S}_k)} + \boldsymbol{D}_k\boldsymbol{u}_k, \boldsymbol{x}_k\right). \tag{4.14}$$

Let $\boldsymbol{e} = (1, 0, \ldots, 0)^\top \in \mathbb{R}^{M_k}$. Since $\boldsymbol{u}_k$ minimizes $Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_k^{(\mathcal{S}_k)} + \boldsymbol{D}_k\ .\ , \boldsymbol{x}_k\right)$, any $t \in \mathbb{R}$ satisfies

$$Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_k^{(\mathcal{S}_k)} + \boldsymbol{D}_k\boldsymbol{u}_k, \boldsymbol{x}_k\right) \leq Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_k^{(\mathcal{S}_k)} + \boldsymbol{D}_k t\boldsymbol{e}, \boldsymbol{x}_k\right) = Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_k^{(\mathcal{S}_k)} + t\ \boldsymbol{d}_k, \boldsymbol{x}_k\right).$$

Moreover, $t \mapsto Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_k^{(\mathcal{S}_k)} + t\ \boldsymbol{d}_k, \boldsymbol{x}_k\right)$ is scalar quadratic with $F(\boldsymbol{x}_k) - \frac{\left\langle \nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k), \boldsymbol{d}_k\right\rangle^2}{2\|\boldsymbol{d}_k\|^2_{\boldsymbol{A}^{(\mathcal{S}_k)}(\boldsymbol{x}_k)}}$ as minimal value. Hence, Assumptions 4.3 leads to

$$Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_k^{(\mathcal{S}_k)} + \boldsymbol{D}_k\boldsymbol{u}_k, \boldsymbol{x}_k\right) \leq F(\boldsymbol{x}_k) - \frac{\left\langle \nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k), \boldsymbol{d}_k\right\rangle^2}{2\|\boldsymbol{d}_k\|^2_{\boldsymbol{A}^{(\mathcal{S}_k)}(\boldsymbol{x}_k)}} \leq F(\boldsymbol{x}_k) - \frac{\gamma_0^2\|\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k)\|^2}{2\|\boldsymbol{d}_k\|^2_{\boldsymbol{A}^{(\mathcal{S}_k)}(\boldsymbol{x}_k)}}. \tag{4.15}$$

Combining (4.15) and (4.14) ensures that $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ is a decreasing sequence. The coercivity of F in Assumption 4.2 both guarantees that $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to a finite limit F^∞ and that $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is bounded, which concludes the first part of the proof. Using now Assumption 4.4(ii) and (4.15) leads to

$$Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_k^{(\mathcal{S}_k)} + \boldsymbol{D}_k\boldsymbol{u}_k, \boldsymbol{x}_k\right) \leq F(\boldsymbol{x}_k) - \frac{\gamma_0^2}{2\gamma_1^2\nu}\|\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k)\|^2. \tag{4.16}$$

Combination of (4.16) with (4.14) directly gives (4.10) in Lemma 4.1. Moreover,

$$\sum_{k=0}^{+\infty}\|\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k)\|^2 \leq \frac{2\gamma_1^2\nu}{\gamma_0^2}\left(F^\infty - F(\boldsymbol{x}_0)\right), \tag{4.17}$$

so that (4.11) in Lemma 4.1 is obtained. Since function $Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_k^{(\mathcal{S}_k)} + \boldsymbol{D}_k\ .\ , \boldsymbol{x}_k\right)$ is quadratic, we also deduce that its minimizer $\boldsymbol{u}_k$ satisfies:

$$\left(\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k)\right)^\top \boldsymbol{D}_k\boldsymbol{u}_k = -\left\|\boldsymbol{D}_k\boldsymbol{u}_k\right\|^2_{\boldsymbol{A}^{(\mathcal{S}_k)}(\boldsymbol{x}_k)}. \tag{4.18}$$

Equality $\boldsymbol{D}_k\boldsymbol{u}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$, and Assumption 4.4(ii), lead to

$$Q^{(\mathcal{S}_k)}\left(\boldsymbol{x}_k^{(\mathcal{S}_k)} + \boldsymbol{D}_k\boldsymbol{u}_k, \boldsymbol{x}_k\right) = F(\boldsymbol{x}_k) - \frac{1}{2}\|\boldsymbol{D}_k\boldsymbol{u}_k\|^2_{\boldsymbol{A}^{(\mathcal{S}_k)}(\boldsymbol{x}_k)} \geq F(\boldsymbol{x}_k) - \frac{\nu}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2. \tag{4.19}$$

Equation (4.13) of Lemma 4.1 then comes by plugging (4.19) into (4.16). Using again the expression of $\boldsymbol{u}_k$ as a minimizer of a quadratic form, we can rewrite one iteration of B2MS scheme as

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\boldsymbol{B}_k\ \nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k), \tag{4.20}$$

with $\boldsymbol{B}_k \triangleq \boldsymbol{D}_k\left(\boldsymbol{D}_k^\top\boldsymbol{A}^{(\mathcal{S}_k)}(\boldsymbol{x}_k)\boldsymbol{D}_k\right)^{-1}\boldsymbol{D}_k^\top$. Remark that Assumption 4.3(i) ensures that $\boldsymbol{B}_k$ is a well-defined symmetric definite positive matrix. Then, by Assumption 4.4(ii),

$$\boldsymbol{B}_k \preceq \frac{1}{\eta}\boldsymbol{D}_k\left(\boldsymbol{D}_k^\top\boldsymbol{D}_k\right)^{-1}\boldsymbol{D}_k^\top \preceq \frac{1}{\eta}\boldsymbol{I}_{|\mathcal{S}_k|}. \tag{4.21}$$

Plugging (4.21) into (4.20) and taking the squared norm of the quantities gives (4.12).

□

Lemma 4.2. *Let $(u_k)_{k\in\mathbb{N}}, (v_k)_{k\in\mathbb{N}}$ two sequences of positive real. If there exists $k^* \geq K$ such that*

$$(\forall k \geq k^*) \quad u_k \leq \rho\sum_{i=k-K}^{k-1} u_i + v_{k-1}, \tag{4.22}$$

if $\rho < \frac{1}{K}$ and $\sum_{k=0}^{+\infty} v_k < +\infty$, then $\sum_{k=0}^{+\infty} u_k < +\infty$.

Proof. Summing (4.22) from k^* to $n \geq k^*$ leads to

$$\sum_{k=k^*}^{n} u_k \leq \rho\sum_{k=k^*}^{n}\sum_{i=k-K}^{k-1} u_i + \sum_{k=k^*}^{n} v_{k-1}, \tag{4.23}$$

with

$$\sum_{k=k^*}^{n}\sum_{i=k-K}^{k-1} u_i = \sum_{k=k^*}^{n}\sum_{i=1}^{K} u_{k-i} = \sum_{i=1}^{K}\sum_{k=k^*-i}^{n-i} u_k \leq \sum_{i=1}^{K}\sum_{k=0}^{n} u_k. \tag{4.24}$$

Plugging (4.24) into (4.23), yields

$$\sum_{k=k^*}^{n} u_k \leq \rho K \sum_{k=k*}^{n} u_k + \left(\rho K \sum_{k=0}^{k^*-1} u_k + \sum_{k=k^*}^{n} v_{k-1}\right) \leq \rho K \sum_{k=k*}^{n} u_k + \left(\rho K \sum_{k=0}^{k^*-1} u_k + \sum_{k=0}^{+\infty} v_k\right), \tag{4.25}$$

that is $(1-\rho K)\sum_{k=k^*}^{n} u_k \leq \rho K \sum_{k=0}^{k^*-1} u_k + \sum_{k=0}^{+\infty} v_k$. With $0 < 1-\rho K < 1$, we deduce the summability of $(u_k)_{k\in\mathbb{N}}$. □

Lemma 4.1 gathers the different inequalities and descent properties which result from B2MS scheme (4.4). Equations (4.10)-(4.11) can be interpreted as a generalized block version of [58, Theorem 1]. Lemma 4.2 is an alternative of [34, Lemma 3], [76, Lemma 5.1].

4.4 Asymptotical behaviour

For the sake of clarity, our presentation for the convergence analysis of scheme (4.4) is divided into three parts. First, we establish the convergence of the gradient of the B2MS iterates to zero, under Assumptions 4.1-4.5. Second, under the additional Assumption 4.6 (i.e., KL inequality), we show the convergence of the iterates of B2MS to a stationary point. Third, we establish a convergence rate result involving the KL exponent θ of function F.

4.4.1 Global convergence

Theorem 4.1. *Let $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ the B2MS sequence. Under Assumptions 4.1-4.5, sequence $(\|\nabla F(\boldsymbol{x}_k)\|)_{k\in\mathbb{N}}$ converges to 0. Moreover, there exists $\boldsymbol{x}^*$, a stationary point of F, such that $F(\boldsymbol{x}_k) \underset{k\to+\infty}{\longrightarrow} F(\boldsymbol{x}^*)$.*

Proof. Let $k \geq K$. For all $\mathcal{S} \in \mathbb{S}$, Assumption 4.1 ensures that $\{t \in [\![k-K, k-1]\!] \ / \ \mathcal{S}_t = \mathcal{S}\}$ is a non-empty set. We can thus rewrite every set of $\mathbb{S}$ as $\mathcal{S} = \mathcal{S}_{T_k^{\mathcal{S}}}$ with

$$T_k^{\mathcal{S}} \triangleq \max\{t \in [\![k-K, k-1]\!] \ / \ \mathcal{S}_t = \mathcal{S}\}. \tag{4.26}$$

The application of $k - T_k^{\mathcal{S}}$ Jensen's inequalities on $\|\nabla F^{\mathcal{S}}(\boldsymbol{x}_k)\|^2$ leads to

$$\|\nabla F^{(\mathcal{S})}(\boldsymbol{x}_k)\|^2 \leq \sum_{i=T_k^{\mathcal{S}}}^{k-1} 2^{k-i}\|\nabla F^{(\mathcal{S})}(\boldsymbol{x}_{i+1}) - \nabla F^{(\mathcal{S})}F(\boldsymbol{x}_i)\|^2 + 2^{k-T_k^{\mathcal{S}}}\|\nabla F^{(\mathcal{S})}F(\boldsymbol{x}_{T_k^{\mathcal{S}}})\|^2. \tag{4.27}$$

We now majorize both parts of the right term of (4.27). According to Lemma 4.1, the iterates belong to a bounded set that we denote E. Using Assumption 4.5, we apply inequality (4.8), using the short

notation $\beta \triangleq \beta(E)$, and then apply (4.12). The latter part of (4.27) is handled by using $\mathcal{S} = \mathcal{S}_{T_k^{\mathcal{S}}}$ and noticing that $k - i \in]0, K[$ for $i \in [\![T_k^{\mathcal{S}}, k-1]\!]$. It yields

$$\begin{aligned}
\|\nabla F^{(\mathcal{S})}(\boldsymbol{x}_k)\|^2 &\leq \frac{2^K \beta^2}{\eta^2} \sum_{i=T_k^{\mathcal{S}}}^{k-1} \|\nabla F^{(\mathcal{S}_i)}(\boldsymbol{x}_i)\|^2 + 2^K \|\nabla F^{(\mathcal{S}_{T_k^{\mathcal{S}}})}(\boldsymbol{x}_{T_k^{\mathcal{S}}})\|^2, \\
&\leq 2^K \left(\frac{\beta^2}{\eta^2} + 1\right) \sum_{i=T_k^{\mathcal{S}}}^{k-1} \|\nabla F^{(\mathcal{S}_i)}(\boldsymbol{x}_i)\|^2, \\
&\leq 2^K \left(\frac{\beta^2}{\eta^2} + 1\right) \sum_{i=k-K}^{k-1} \|\nabla F^{(\mathcal{S}_i)}(\boldsymbol{x}_i)\|^2. \qquad (4.28)
\end{aligned}$$

Then from (4.11) we have $\|\nabla F^{(\mathcal{S})}(\boldsymbol{x}_k)\|^2 \underset{k\to+\infty}{\longrightarrow} 0$. Since $[\![1, N]\!] = \bigcup_{\mathcal{S}\in\mathbb{S}} \mathcal{S}$ (by Ass. 4.1) and $\mathbb{S}$ finite,

$$\|\nabla F(\boldsymbol{x}_k)\|^2 \leq \sum_{\mathcal{S}\in\mathbb{S}} \|\nabla F^{(\mathcal{S})}(\boldsymbol{x}_k)\|^2 \underset{k\to+\infty}{\longrightarrow} 0. \qquad (4.29)$$

According to Lemma 4.1, $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to a finite limit denoted F^∞. The boundedness of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ guarantees the existence of $\boldsymbol{x}^* \in \mathbb{R}^n$, an accumulation point of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$. As $\|\nabla F(\boldsymbol{x}_k)\| \underset{k\to+\infty}{\longrightarrow} 0$ and ∇F is continuous (by Ass. 4.2), $\nabla F(\boldsymbol{x}^*) = \mathbf{0}$ directly follows, so that $\boldsymbol{x}^*$ is a stationary point of F. Moreover, $F^\infty = F(\boldsymbol{x}^*)$ since F is continuous, which concludes our proof. □

Theorem 4.1 shows a classical behaviour for a descent method applied to a non-convex Lipschitz differentiable objective function [178].

4.4.2 Sequence convergence

This part is dedicated to refine the result of Theorem 4.1, when we additionally introduce Assumption 4.6. We first state a technical inequality giving a direct relation between the gradient at the current iterate and the differences of past iterates over a cycle (i.e a K-length period)

Proposition 4.1. *Let $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ the B2MS sequence. Under Assumptions 4.1-4.5, for all $k \geq K$,*

$$\|\nabla F(\boldsymbol{x}_k)\|^2 \leq 2^K C \left(\beta^2 + \frac{\gamma_1^2}{\gamma_0^2}\nu^2\right) \sum_{i=k-K}^{k-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|^2. \qquad (4.30)$$

Proof. Let $k \geq K$. The beginning of the proof is identical to that of Theorem 4.1 until Eq.(4.27). We then derive a new majoration for the quantity involved in the right term of (4.27). We first majorize

using Ass. 4.2 combined with $\mathcal{S} = \mathcal{S}_{T_k^{\mathcal{S}}}$. We then use (4.13) of Lemma 4.1. This yields:

$$\begin{aligned}
\|\nabla F^{(\mathcal{S})}(\boldsymbol{x}_k)\|^2 &\leq 2^K \beta^2 \sum_{i=T_k^{\mathcal{S}}}^{k-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|^2 + 2^K \|\nabla F^{(\mathcal{S}_{T_k^{\mathcal{S}}})}(\boldsymbol{x}_{T_k^{\mathcal{S}}})\|^2, \\
&\leq 2^K \beta^2 \sum_{i=T_k^{\mathcal{S}}}^{k-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|^2 + 2^K \frac{\nu^2 \gamma_1^2}{\gamma_0^2} \|\boldsymbol{x}_{T_k^{\mathcal{S}}+1} - \boldsymbol{x}_{T_k^{\mathcal{S}}}\|^2, \\
&\leq 2^K \left(\beta^2 + \frac{\gamma_1^2}{\gamma_0^2}\nu^2\right) \sum_{i=T_k^{\mathcal{S}}}^{k-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|^2, \\
&\leq 2^K \left(\beta^2 + \frac{\gamma_1^2}{\gamma_0^2}\nu^2\right) \sum_{i=k-K}^{k-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|^2. \qquad (4.31)
\end{aligned}$$

We then sum (4.31) over $\mathbb{S}$ (of cardinal C). The right term being independent from $\mathcal{S}$, it is simply multiplied by C. Then, noting that $\|\nabla F(\boldsymbol{x}_k)\|^2 \leq \sum_{\mathcal{S}\in\mathbb{S}} \|\nabla F^{(\mathcal{S})}(\boldsymbol{x}_k)\|^2$ concludes our proof. □

Remark that an alternative proof of Theorem 4.1 could be obtained from Prop. 4.1. However, inequality (4.28) is more direct to demonstrate than (4.30).

We finally state our main theoretical result, namely the convergence of the iterates of B2MS scheme.

Theorem 4.2. *Let $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ the B2MS sequence. Under Assumptions 4.1-4.6,*

$$\sum_{k=0}^{+\infty} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| < +\infty. \qquad (4.32)$$

Moreover, $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ converges to a stationary point of F.

Proof. Following the same notations as those of Theorem 4.1, let us apply (4.9) to set $E = \{\boldsymbol{x}_k \,/\, k \in \mathbb{N}\}$, which is a bounded set by Lemma 4.1, and $\xi = F^\infty$. Then, by KL inequality (Ass. 4.6), there exists $(\kappa, \zeta, \theta) \in \mathbb{R}_+^* \times \mathbb{R}_+^* \times]0,1[$ such that for every $k \in \mathbb{N}$ verifying $|F(\boldsymbol{x}_k) - F^\infty| \leq \zeta$,

$$\|\nabla F(\boldsymbol{x}_k)\| \geq \kappa \left|F(\boldsymbol{x}_k) - F^\infty\right|^\theta. \qquad (4.33)$$

Using (4.13) and (4.10) gives, for every $k \in \mathbb{N}$,

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \leq \frac{2\gamma_1^2\nu}{\gamma_0^2\eta^2}\left(F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k+1})\right) = \frac{2\gamma_1^2\nu}{\gamma_0^2\eta^2}\left[(F(\boldsymbol{x}_k) - F^\infty) - (F(\boldsymbol{x}_{k+1}) - F^\infty)\right]. \qquad (4.34)$$

Let us now invoke the convexity of $v \in \mathbb{R}_+ \mapsto v^{\frac{1}{1-\theta}}$. It follows that for all $v, w \in \mathbb{R}_+$ with $v \leq w$

$$w - v \leq (1-\theta)^{-1} w^\theta \left(w^{1-\theta} - v^{1-\theta}\right). \qquad (4.35)$$

Plugging (4.35) in (4.34) with $w = F(\boldsymbol{x}_k) - F^\infty$ and $v = F(\boldsymbol{x}_{k+1}) - F^\infty$ yields, for every $k \in \mathbb{N}$,

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \leq \frac{2\gamma_1^2\nu}{\gamma_0^2\eta^2(1-\theta)}[F(\boldsymbol{x}_k) - F^\infty]^\theta \Delta^k, \qquad (4.36)$$

with $\Delta_k \triangleq \left(F(\boldsymbol{x}^k) - F^\infty\right)^{1-\theta} - \left(F(\boldsymbol{x}_{k+1}) - F^\infty\right)^{1-\theta}$. Since $F(\boldsymbol{x}_k) \underset{k\to+\infty}{\longrightarrow} F^\infty$, there exists $k_0 \geq K$ such that

$$(\forall k \geq k_0) \quad |F(\boldsymbol{x}_k) - F^\infty| \leq \zeta. \tag{4.37}$$

Thus, using (4.33),

$$(\forall k \geq k_0) \quad \|\nabla F(\boldsymbol{x}_k)\| \geq \kappa \left|F(\boldsymbol{x}_k) - F^\infty\right|^\theta. \tag{4.38}$$

Combining (4.38) and (4.36) leads to

$$(\forall k \geq k_0) \quad \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \leq \frac{2\gamma_1^2\nu}{\gamma_0^2\eta^2\kappa(1-\theta)} \|\nabla F(\boldsymbol{x}_k)\| \Delta_k. \tag{4.39}$$

We now rely on the majoration of Proposition 4.1, to obtain

$$(\forall k \geq k_0) \quad \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \leq \Lambda \sqrt{\sum_{i=k-K}^{k-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|^2} \ \Delta_k \leq \Lambda \sum_{i=k-K}^{k-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\| \ \Delta_k, \tag{4.40}$$

with $\Lambda \triangleq \frac{2^{\frac{K}{2}+1}}{\kappa(1-\theta)} \frac{\gamma_1^2\nu}{\gamma_0^2\eta^2} \sqrt{C} \sqrt{\beta^2 + \frac{\gamma_1^2}{\gamma_0^2}\nu^2}$. We extract the square root of (4.40) and invoke the inequality $\sqrt{ab} \leq \frac{a}{c} + \frac{bc}{4}$ with $a = \sum_{i=k-K}^{k-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|$, $b = \Delta_k$ and some $c > 0$. Then,

$$(\forall k \geq k_0) \quad \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| \leq \frac{\sqrt{\Lambda}}{c} \sum_{i=k-K}^{k-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\| + \frac{c}{4}\sqrt{\Lambda}\Delta_k. \tag{4.41}$$

$(\Delta_k)_{k\in\mathbb{N}}$ is summable and $\sqrt{\Lambda}/c \in]0, 1/K[$ for $c > \sqrt{\Lambda}K$, we apply Lemma 4.2 with $k^* = k_0$. Finally, $(\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|)_{k\in\mathbb{N}}$ is summable and $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is a Cauchy sequence possessing $\boldsymbol{x}^*$ as an accumulation point. Sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ thus converges to $\boldsymbol{x}^*$. □

4.4.3 Convergence rate

As highlighted above, Theorem 4.2 guarantees the convergence of sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ to $\boldsymbol{x}^*$, a stationary point of function F. Our last contribution lies in characterizing the convergence rate for B2MS algorithm. Here again, KL inequality (Ass. 4.6) is an anchor point of our analysis.

Theorem 4.3. *Let $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ the B2MS sequence. Under Assumptions 4.1-4.6, the following holds.*

(i) if $\theta \in]1/2, 1[$, then $\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \underset{k\to+\infty}{=} \mathcal{O}\left(k^{-\frac{1-\theta}{2\theta-1}}\right)$;

(ii) If $\theta \in]0, 1/2]$, then there exists $\varepsilon \in]0, 1[$ such that $\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \underset{k\to+\infty}{=} \mathcal{O}\left(\varepsilon^k\right)$.

Proof. Keeping the same notations as previously, let $c > 0$ and $k \geq k_0$. We sum (4.41) from kK

$$\Gamma_k \leq \frac{\sqrt{\Lambda}}{c} \sum_{j=kK}^{+\infty} \sum_{i=j-K}^{j-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\| + \frac{c}{4}\sqrt{\Lambda} \sum_{j=kK}^{+\infty} \Delta_j, \tag{4.42}$$

with $\Gamma_k \triangleq \sum_{j=kK}^{+\infty} \|\boldsymbol{x}_{j+1} - \boldsymbol{x}_j\|$. On the one hand,

$$\begin{aligned}\sum_{j=kK}^{+\infty} \sum_{i=j-K}^{j-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\| &= \sum_{j=kK}^{+\infty} \sum_{i=0}^{K-1} \|\boldsymbol{x}_{i+j-K+1} - \boldsymbol{x}_{i+j-K}\|, \\ &= \sum_{i=0}^{K-1} \sum_{j=kK}^{+\infty} \|\boldsymbol{x}_{i+j-K+1} - \boldsymbol{x}_{i+j-K}\| \leq K\Gamma_{k-1}. \end{aligned} \tag{4.43}$$

On the other hand, using $F(\boldsymbol{x}) - F(\boldsymbol{x}^*) \underset{k\to+\infty}{\longrightarrow} 0$, (4.33) and Proposition 4.1, yields

$$\begin{aligned}[F(\boldsymbol{x}_{kK}) - F(\boldsymbol{x}^*)]^{1-\theta} \leq \kappa^{\frac{\theta-1}{\theta}} \|\nabla F(\boldsymbol{x}_{kK})\|^{\frac{1-\theta}{\theta}} &\leq \Lambda' \left(\sum_{j=(k-1)K}^{kK-1} \|\boldsymbol{x}_{j+1} - \boldsymbol{x}_j\|^2 \right)^{\frac{1-\theta}{2\theta}}, \\ &\leq \Lambda' \left(\sum_{j=(k-1)K}^{kK-1} \|\boldsymbol{x}_{j+1} - \boldsymbol{x}_j\| \right)^{\frac{1-\theta}{\theta}}, \\ &= \Lambda' \left(\Gamma_{k-1} - \Gamma_k\right)^{\frac{1-\theta}{\theta}}, \end{aligned} \tag{4.44}$$

with $\Lambda' \triangleq \kappa^{\frac{\theta-1}{\theta}} \left(2^K C \left(\beta^2 + \frac{\gamma_1^2}{\gamma_0^2}\nu^2\right)\right)^{\frac{1-\theta}{2\theta}}$. Plugging (4.43), (4.44) in (4.42), with $c = 2\sqrt{\Lambda}K$ and $\sum_{j=kK}^{+\infty} \Delta_j = [F(\boldsymbol{x}_{kK}) - F(\boldsymbol{x}^*)]^{1-\theta}$ gives

$$\Gamma_k \leq \frac{1}{2}\Gamma_{k-1} + \frac{1}{2}K\Lambda\Lambda' \left(\Gamma_{k-1} - \Gamma_k\right)^{\frac{1-\theta}{\theta}}. \tag{4.45}$$

By multiplying (4.45) by 2 and removing Γ_k from each side,

$$\Gamma_k \leq (\Gamma_{k-1} - \Gamma_k) + K\Lambda\Lambda' \left(\Gamma_{k-1} - \Gamma_k\right)^{\frac{1-\theta}{\theta}}. \tag{4.46}$$

Since $(\Gamma_k)_{k\in\mathbb{N}}$ is a positive decreasing sequence to zero, we can apply [8, Theorem 2].

- If $\theta \in]1/2, 1[$, there exists $\lambda > 0$ such that

$$(\forall k \geq k_0) \quad \Gamma_k \leq \lambda k^{-\frac{1-\theta}{2\theta-1}}. \tag{4.47}$$

Denoting $q(k), r(k)$ the quotient and remainder of the Euclidean division of k by K, leads to

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| = \|\boldsymbol{x}_{q(k)K+r(k)} - \boldsymbol{x}^*\| \leq \Gamma_{q(k)}. \tag{4.48}$$

Combining (4.48) with (4.47) gives, for k large enough,

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq \lambda q(k)^{-\frac{1-\theta}{2\theta-1}} = \lambda \left(\frac{k - r(k)}{K}\right)^{-\frac{1-\theta}{2\theta-1}} \leq \lambda \left(\frac{k}{K} - 1\right)^{-\frac{1-\theta}{2\theta-1}}, \tag{4.49}$$

with $\lambda \left(\frac{k}{K} - 1\right)^{-\frac{1-\theta}{2\theta-1}} \underset{k\to+\infty}{=} \mathcal{O}\left(k^{-\frac{1-\theta}{2\theta-1}}\right)$.

- If $\theta \in]0, 1/2]$, there exist $\mu > 0$ and $\delta \in]0, 1[$ such that

$$(\forall k \geq k_0) \quad \Gamma_k \leq \mu \ \delta^k. \tag{4.50}$$

Similarly with the previous case, for k large enough,

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq \mu \ \delta^{q(k)} = \mu \ \delta^{\frac{k-r(k)}{K}} \leq \mu \ \delta^{\frac{k}{K}}. \tag{4.51}$$

Conclusion is obtained by taking $\varepsilon = \delta^{\frac{1}{K}} \in]0, 1[$.

□

One can notice the dependency of the convergence rate with the KL exponent θ. This result is similar to the one in [66], though with a more direct proof following naturally from our global convergence Theorem 4.2. We also notice that the gradient Lipschitz property (see Assumption 4.5) entails that the convergence rate obtained for $(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|)_{k\in\mathbb{N}}$ also holds for the sequence $(\|\nabla F(\boldsymbol{x}_k)\|)_{k\in\mathbb{N}}$.

4.5 Numerical illustration

The numerical part of this chapter arises from previous works conducted by Emilie Chouzenoux and Jean-Christophe Pesquet, which has not been submitted for publication until now.

4.5.1 Problem formulation

Let us illustrate numerically the benefits of each acceleration features introduced in B2MS. To do so, we focus on the signal processing problem of sparse signal blind deconvolution [118]. Let $\boldsymbol{y} \in \mathbb{R}^P$, $P \geq 1$, a vector of observations, related to a sought signal $\overline{\boldsymbol{z}} \in \mathbb{R}^P$ through the model

$$\boldsymbol{y} = \overline{\boldsymbol{h}} * \overline{\boldsymbol{z}} + \boldsymbol{e}, \tag{4.52}$$

with $\overline{\boldsymbol{h}} \in \mathbb{R}^L$, $L \geq 1$, a blur kernel, $*$ the 1D discrete convolution operator with zero-padding assumption, and $\boldsymbol{e} \in \mathbb{R}^P$ a realization of an i.i.d. zero-mean Gaussian distribution. Blind deconvolution amounts to retrieving estimates $(\widetilde{\boldsymbol{z}}, \widetilde{\boldsymbol{h}})$ of $(\overline{\boldsymbol{z}}, \overline{\boldsymbol{h}})$ from $\boldsymbol{y}$ under some structural assumptions. Namely here, we consider a sparse signal $\overline{\boldsymbol{z}}$, with few non-zero entries within the range $[z_{\min}, z_{\max}]$, $-\infty < z_{\min} < z_{\max} < +\infty$, and a blur kernel with bounded energy and entries within the range $[h_{\min}, h_{\max}]$, $-\infty < h_{\min} < h_{\max} < +\infty$. Under these specifications, we propose to solve (4.1) where

$$(\forall \boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{h}) \in \mathbb{R}^{P+L})$$
$$F(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{h} * \boldsymbol{z} - \boldsymbol{y}\|^2 + \lambda \, \mathrm{SOOT}(\boldsymbol{z}) + \frac{\rho}{2} d^2_{[z_{\min}, z_{\max}]^P}(\boldsymbol{z}) + \frac{\xi}{2} d^2_{[h_{\min}, h_{\max}]^L}(\boldsymbol{h}) + \frac{\zeta}{2}\|\boldsymbol{h}\|^2, \tag{4.53}$$

with $(\lambda, \rho, \xi, \zeta) > 0$ some penalization weights. Hereabove, SOOT denotes the smoothed ℓ_1-over-ℓ_2 sparsity promoting penalty introduced in [196], defined as

$$(\forall \boldsymbol{z} \in \mathbb{R}^P) \quad \mathrm{SOOT}(\boldsymbol{z}) = \log\left(\frac{\ell_{1,\alpha}(\boldsymbol{z}) + \beta}{\ell_{2,\eta}(\boldsymbol{z})}\right) \tag{4.54}$$

where

$$(\forall \boldsymbol{z} = (z^p)_{p\in[\![1,P]\!]} \in \mathbb{R}^P) \quad \ell_{1,\alpha}(\boldsymbol{z}) = \sum_{p=1}^{P} \left(\sqrt{(z^p)^2 + \alpha^2} - \alpha \right), \quad \ell_{2,\eta}(\boldsymbol{z}) = \sqrt{\sum_{p=1}^{P} (z^p)^2 + \eta^2}. \tag{4.55}$$

Function (4.54) is non-convex and Lipschitz-differentiable on $\mathbb{R}^P$ (see [53, Prop.2] with $p = 1$ and $q = 2$). The scalars $(\alpha, \beta, \eta) > 0$ act as smoothing hyper-parameters so that the penalty term (4.54) can be viewed as a smoothed non-convex proxy of the norm ratio ℓ_1 over ℓ_2. SOOT penalty was shown in [196] to suitably enhance the restoration of sparse signals in the context of blind deconvolution, when compared to standard ℓ_1-based formulation. It was later on generalized in [53, 247] to tackle signal processing tasks arising in chemistry, and hereagain showed superior results when compared to various state-of-the-art sparsity priors. Function d_C denotes the Euclidean distance to a set C. If C is non-empty and convex, function $\frac{1}{2}d_C^2$ is convex and 1-Lipschitz differentiable [14]. The two distance terms in (4.53) act as smoothed exterior penalty terms, weighted by (ρ, ξ) favoring the fulfillment of the considered range constraints. Finally, the quadratic penalty term weighted by ζ controls the boundedness of the estimated kernel energy.

4.5.2 B2MS implementation

Let us discuss the practical application of the proposed B2MS scheme to the minimization of function (4.53). The latter is $\mathcal{C}^1$ and coercive on $\mathbb{R}^{P+L}$ so Assumption 4.2 holds. Moreover, it satisfies the Lipschitz condition from Assumption 4.5. A similar analysis than in [196] leads to the fulfillment of Assumption 4.6. We define the family

$$\mathbb{S} = \{[\![1, P]\!], [\![P+1, P+L]\!]\}, \tag{4.56}$$

with the aim to build a B2MS scheme alternating between updates on the signal variable $\boldsymbol{z}$ and the kernel variable $\boldsymbol{h}$. We adopt a quasi-cyclic strategy

$$(\forall k \in \mathbb{N}) \quad \mathcal{S}_k = \begin{cases} [\![P+1, P+L]\!] \text{ if } (k \equiv K) = 0, \\ [\![1, P]\!] \text{ otherwise,} \end{cases} \tag{4.57}$$

with $\equiv$ the modulo operation, and $K \geq 1$ a predefined value of the number of updates on variable $\boldsymbol{z}$ before processing again (and a single time) the variable $\boldsymbol{h}$. Definitions (4.56)-(4.57) satisfy by construction Assumption 4.1. Let $\boldsymbol{H} \in \mathbb{R}^{P\times P}$ the Toeplitz operator such that $\boldsymbol{h} * \boldsymbol{z} = \boldsymbol{H}\boldsymbol{z}$, and $\boldsymbol{Z} \in \mathbb{R}^{L\times L}$ the correlation operator such that $\boldsymbol{h} * \boldsymbol{z} = \boldsymbol{Z}\boldsymbol{h}$. We define

$$(\forall \boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{h}) \in \mathbb{R}^{P+L})$$
$$\boldsymbol{A}^{([\![1,P]\!])}(\boldsymbol{x}) = \boldsymbol{H}^\top \boldsymbol{H} + \frac{9\lambda}{8\eta^2}\boldsymbol{I}_P + \frac{\lambda}{\ell_{1,\alpha}(\boldsymbol{z}) + \beta} \mathrm{Diag}\left((((z^p)^2 + \alpha^2)^{-1/2})_{p\in[\![1,P]\!]} \right) + \rho \boldsymbol{I}_P, \tag{4.58}$$

and

$$(\forall \boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{h}) \in \mathbb{R}^{P+L}) \quad \boldsymbol{A}^{([\![P+1,P+L]\!])}(\boldsymbol{x}) = \boldsymbol{Z}^\top \boldsymbol{Z} + (\xi + \zeta)\boldsymbol{I}_L. \tag{4.59}$$

Using [196, Prop.2], the descent lemma [14] and the block definitions (4.56)-(4.57) allows to show that the majorizing condition (4.3) holds for every $k \in \mathbb{N}$. Hence, Assumption 4.4(i) holds. Regarding the

subspace choice, we adopt unless specified otherwise the memory gradient subspace from [164, 59], which reads in the block alternating case as ([43]):

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{D}_k = [-\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k) \mid \boldsymbol{x}_k^{(\mathcal{S}_k)} - \boldsymbol{x}_{\iota_k}^{(\mathcal{S}_k)}] \in \mathbb{R}^{|\mathcal{S}_k| \times 2}, \tag{4.60}$$

where, for every $k \in \mathbb{N}$, $\iota_k \in [\![0, k-1]\!]$ is the largest index when the same block $\mathcal{S}_{\iota_k} = \mathcal{S}_k$ was updated (with default choice $\iota_k = 0$). This choice of subspace trivially meets Assumption 4.3. In a nutshell, Assumption 4.1 to Assumption 4.4(i) hold, so that, by Lemma 4.1, the B2MS iterates are bounded. This allows to deduce, through a straightforward analysis of (4.58)-(4.59), that Assumption 4.4(ii) holds. Furthermore, Assumptions 4.5 and 4.6 hold so that our convergence Theorems 4.1 and 4.2 apply.

4.5.3 Numerical results

We now present our numerical results. We rely on the same dataset and model implementation than in the SOOT Matlab toolbox[1]. The ground truth signal/kernel $(\overline{\boldsymbol{z}}, \overline{\boldsymbol{h}})$ have size $P = 784$ and $L = 41$. We set the range values $(z_{\min}, z_{\max}, h_{\min}, h_{\max})$ to the ground truth minimal and maximal values of the sought quantities. The SOOT parameters $(\lambda, \alpha, \beta, \eta)$ and the initialization $(\boldsymbol{z}_0, \boldsymbol{h}_0)$ are kept unchanged with respect to the toolbox implementation. A rough grid search is used to set the hyper-parameters $(\rho, \xi, \zeta) = (10^{-1}, 10^{-1}, 10^{-2})$ so as to reach an accurate restoration. Example of observed signal $\boldsymbol{y}$, ground truth signal/kernel vectors $(\overline{\boldsymbol{z}}, \overline{\boldsymbol{h}})$ and estimated ones $(\widetilde{\boldsymbol{z}}, \widetilde{\boldsymbol{h}})$ are displayed in Figure 4.1. Running times displayed in the forthcoming sections are for a Matlab 2021a implementation on a x64 Dell Destop with 11th Gen Intel(R) Core(TM) i7-1185G7 @ 3.00GHz with 32 Go RAM.

4.5.3.1 *Role of quasi-cyclic rule*

The rule (4.57) corresponds to the standard cyclic (i.e., Gauss Seidel) rule when $K = 1$. When $K > 1$, a quasi-cyclic rule is obtained, when B2MS iterates several times on the signal before updating (once) the kernel. Our theoretical analysis encompasses such choice. To illustrate the benefit for such versatility, we display on Figure 4.2 the computational time required to satisfy

$$\|\boldsymbol{z}_k - \widetilde{\boldsymbol{z}}\|_1 \leq 10^{-4}, \quad \text{and} \quad \|\boldsymbol{h}_k - \widetilde{\boldsymbol{h}}\|_1 \leq 10^{-4}, \tag{4.61}$$

as a function of K in (4.57), with $(\widetilde{\boldsymbol{z}}, \widetilde{\boldsymbol{h}})$ the B2MS iterate after a very large number of iterations (typically 10^4). One can observe that the minimal running time is not obtained for $K = 1$ (i.e., cyclic rule) but for a larger value of K, here around 100. This shows the advantage of adopting Assumption 4.1 instead of the standard cyclic update requirement. This phenomenon was already observed in the study in [196] but for another block alternating scheme. The setting $K = 100$ is retained for our next experiments.

[1] https://www.mathworks.com/matlabcentral/fileexchange/50481-soot-l1-l2-norm-ratio-sparse-blind-deconvolution

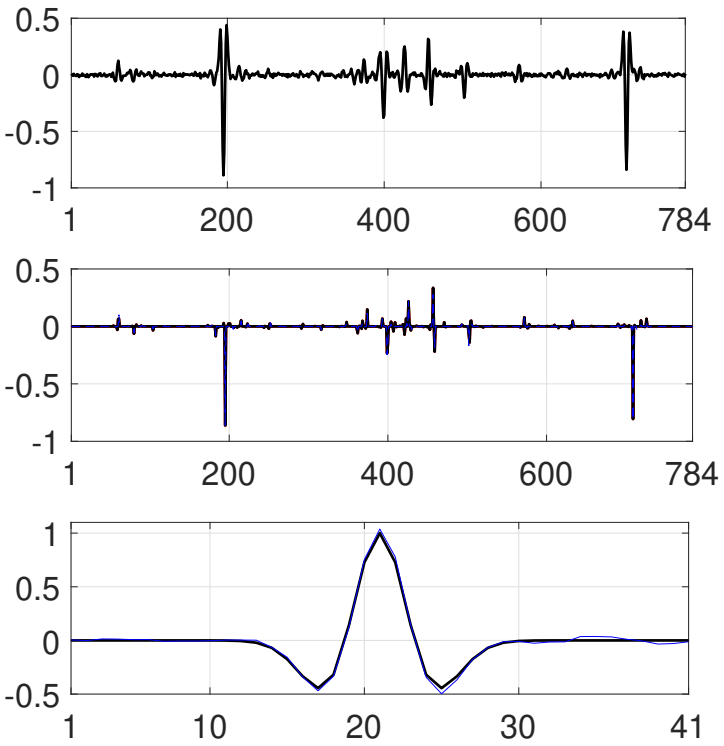


Figure 4.1: Observed signal $\boldsymbol{y}$ (top). Ground truth (black continuous line) and restored (blue dashed line) signal (middle) and kernel (bottom). Reconstruction errors $\|\overline{\boldsymbol{z}} - \widetilde{\boldsymbol{z}}\|_1 = 3.4 \times 10^{-3}$ and $\|\overline{\boldsymbol{h}} - \widetilde{\boldsymbol{h}}\|_1 = 1.7 \times 10^{-2}$.

4.5.3.2 *Ablation study on B2MS*

In order to assess the role of both subspace and MM acceleration in B2MS, we perform an ablation study. To that aim, we implement three ablated versions of B2MS, where we removed either one or both aforementioned features. Namely, we can remove subspace acceleration (B2MS-NoSub) by simply setting

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{D}_k = -\nabla F^{(\mathcal{S}_k)}(\boldsymbol{x}_k) \in \mathbb{R}^{|\mathcal{S}_k|}. \tag{4.62}$$

We can also remove the effect of the MM preconditioner (B2MS-NoPrec) by setting, for every $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{h}) \in \mathbb{R}^{P+L}$, $\boldsymbol{A}^{([\![1,P]\!])}(\boldsymbol{x}) = B(\boldsymbol{h})\boldsymbol{I}_P$ and $\boldsymbol{A}^{([\![P+1,P+L]\!])}(\boldsymbol{x}) = B(\boldsymbol{z})\boldsymbol{I}_L$, with $(B(\boldsymbol{h}), B(\boldsymbol{z})) \in]0, +\infty[^2$ some upper bounds on the spectra of (4.58)-(4.59). For every $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{h}) \in \mathbb{R}^{P+L}$, a straightforward analysis leads to

$$(\forall \boldsymbol{h} \in \mathbb{R}^L) \quad B(\boldsymbol{h}) = |||\boldsymbol{H}|||^2 + \frac{9\lambda}{8\eta^2} + \frac{\lambda}{\alpha\beta} + \rho, \tag{4.63}$$

$$(\forall \boldsymbol{z} \in \mathbb{R}^P) \quad B(\boldsymbol{z}) = |||\boldsymbol{Z}|||^2 + \xi + \zeta, \tag{4.64}$$

with $|||\cdot|||$ the spectral norm. We display in Figure 4.3 the evolution of the gradient norm and the estimation error along time for the four tested methods. A first observation is that the methods not using the MM preconditioner (i.e., B2MS-NoPrec, B2MS-NoPrec-NoSub) exhibit very slow convergence, in

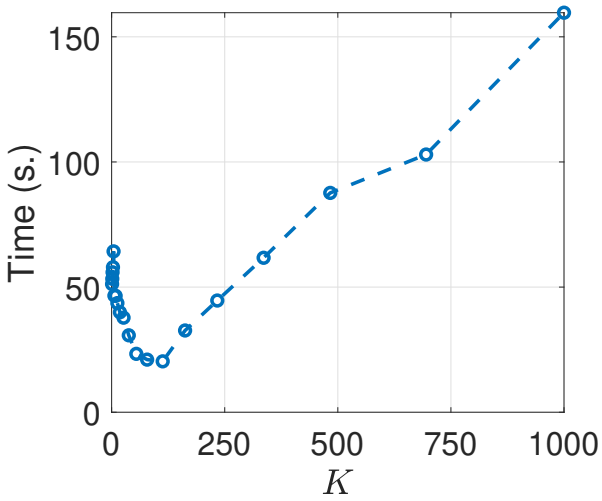


Figure 4.2: Computational time in seconds required for B2MS iterates to satisfy stopping criterion (4.61) as a function of K.

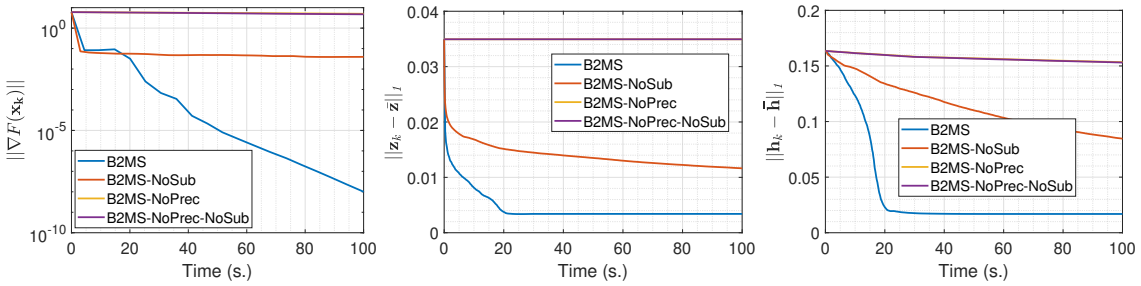


Figure 4.3: Evolution of the gradient norm of F (left), of the estimation error on the signal (middle) and of the estimation error on the kernel (right) along time in seconds for B2MS and its ablated versions. B2MS-NoPrec-NoSub and B2MS-NoPrec plots are almost superimposed.

comparison with the two others. This shows the crucial role of the MM acceleration technique, especially in this example when the bounds (4.64) reach very high values (typically of the order of 10^8). As an additional remark, note that (4.64) requires to recompute, at each iteration, spectral norms of large operators which is cumbersome. Second, using the memory gradient subspace instead of a basic gradient descent search improves the convergence speed of B2MS, as can be seen when comparing B2MS and B2MS-NoSub plots. In a nutshell, both MM preconditioning and subspace catalyzers are essential in this problem.

4.5.3.3 *Comparison with state-of-the-art*

We conclude our analysis by comparing B2MS with several non alternating minimization schemes from the state-of-the-art. Namely, we implemented the nonlinear conjugate gradient (NLCG) algorithm with

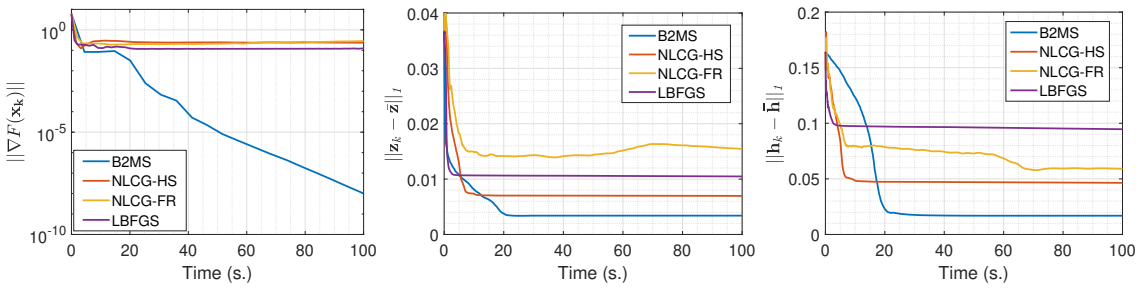


Figure 4.4: Evolution of the gradient norm of F (left), of the estimation error on the signal (middle) and of the estimation error on the kernel (right) along time in seconds for B2MS and its competitors.

either Hestenes-Stiefel (HS) or Fletcher-Reeves (FR) conjugacy formula [119, 101] and a linesearch satisfying strong Wolfe conditions [178]. We also tested the quasi-Newton LBFGS approach [156] combined with an Armijo backtracking linesearch [178]. The linesearch and memory parameters were manually tuned to reach the best performance of each competitor. Note that, up to our knowledge, neither NLCG nor LBFGS possess any theoretical guarantees on the convergence of their iterates, in this example, due to the non-convexity of F. No major numerical instabilities were observed in our experiments. Figure 4.4 shows the convergence plots in terms of gradient norm and restoration error. One can observe the fast convergence of B2MS. In contrast, the three competitors do not seem to reach a convergent point (see Fig. 4.4(left)). Moreover, they lead to solutions with higher estimation error (see Fig. 4.4(middle-right)). This shows again the advantage of the proposed scheme which benefits from sounded convergence and good practical behavior.

4.6 Conclusion

This chapter introduces a block alternating MM subspace algorithm and provides its convergence analysis in the non-convex case. Numerical experiments illustrate the performance of the proposed method on a practical problem of blind signal restoration. When combined with a memory gradient subspace (see, for e.g., [102]), the proposed method can be viewed as a convergent preconditioned block alternating non linear conjugate gradient algorithm for non-convex large scale optimization. Future work will be dedicated to building a distributed implementation for B2MS, for instance by adopting an asynchronous approach [215] where block updates are spread among core machines with possible update delay.

Chapter 5

Block delayed majorize-minimize subspace algorithm for large scale image restoration

In this chapter, we propose an asynchronous Majorization-Minimization (MM) algorithm for solving large scale differentiable non-convex optimization problems. The proposed algorithm runs efficient MM memory gradient updates on blocks of coordinates, in a parallel and possibly asynchronous manner. We establish the convergence of the resulting sequence of iterates under mild assumptions. The performance of the algorithm is illustrated on the restoration of 3D images degraded by depth-variant 3D blur, arising in multiphoton microscopy. Significant computational time reduction, scalability and robustness are observed on synthetic data, when compared to state-of-the-art methods. Experiments on the restoration of real acquisitions of a muscle structure illustrate the qualitative performance of our approach and its practical applicability.

This chapter arises from the article: M. Chalvidal, E. Chouzenoux, J-B. Fest and C. Lefort. *Block delayed Majorize-Minimize subspace algorithm for large scale image restoration*, published in Inverse Problems, volume 39, 2023.

Contents

5.1 Introduction

Large-scale optimization algorithms, benefiting from fast convergence, capable of utilizing modern computing infrastructures, and dealing with distributed datasets are becoming compulsory for solving inverse problems in modern imaging [68]. The ever-growing need for fast processing solutions that can operate on high-dimensional problems (i.e implying a huge number of variables) calls for the development of parallel methods harnessing the power of distributed computational architectures. In addition, the expansion of IoT systems and remote highly parallel computing induce new network issues with specific constraints. For instance, instabilities may occur whenever the volume of data dwarfs the memory capacity of a single agent or when the processing power is shared (potentially unevenly) between devices [139]. Several classes of so-called distributed optimization methods, have been investigated under various assumptions on the computing scenario and on the optimization problem itself, that we review hereafter (see also [248, 240]).

Distributed optimization approaches inherit from block alternating methods. In the latter, at each iteration, only a subset of the variables are updated, by minimizing the objective function with respect to only those variables, the others being fixed. The blocks are selected sequentially following a cyclic (or quasi-cyclic) order or a random rule. Exact minimization with respect to a given block of variables is rarely possible in a closed form. It is not even desirable as it may lead to convergence issues [226]. More efficient and stable block alternating schemes rely on a so-called Majorization-Minimization (MM) strategy [128]. It consists in building, at each iteration, a majorizing approximation for the objective function within the active block of variables, whose minimizer has a more tractable form. Many powerful algorithms fall within this framework, such as BSUM [121], PALM [27], multiplicative methods for NMF [146], to name a few. By exploiting the structure of the objective function, block alternating MM methods can reach fast convergence rates [196, 98, 174, 179] while offering theoretical guarantees in non-convex cases [27, 65, 29].

When the problem size increases, as in 3D microscopy imaging [147] and astronomy [180, 192], running block alternating methods gets inefficient. Parallel implementations have been devised, where the block updates are performed simultaneously, allowing to distribute computations on different nodes (or machines) [43, 222, 200]. Implementation on parallel architecture requires to pay attention to communication cost. The latter can be reduced by resorting to an asynchronous parallel implementation, yielding the so-called distributed optimization approach. Each computation node has now its own iteration loop, so local variables are updated without the need to wait for distant variables update. This however raises challenging questions, in terms of convergence analysis, as the communication delays may introduce instabilities. A plethora of recent works have focused on proposing distributed optimization algorithms with assessed convergence, based on stochastic proximal primal [115, 155, 165] or primal-dual [185, 117, 245, 54, 180, 1] techniques. Recent contributions in the field of federated learning are also highly related [123, 158, 232]. However, as the aforementioned works rely on specific fixed-point analysis tools involving Fenchel-Rockafellar duality [14], the proposed algorithms are limited to convex (sometimes even strongly convex) optimization and often require specific probabilistic assumptions on the block update rule difficult to meet in practice. In the context of MM algorithms, although the need for distributed implementation strategies is crucial (see the discussion in [121] and the specific examples in [228, 94]), theoretical results regarding convergence guarantees of MM tech-

nique in a distributed context are rather scarce. Let us first mention the work of [76, 77], that proposes an asynchronous version of PALM, with proven convergence of the iterates in non-convex case, and good practical behaviour [225]. The convergence of distributed MM methods was also explored in the recent works [45, 154]. However, the analysis of [154] is limited to the convex case. In [45], the analysis covers non-convex terms in the objective function, but it only shows the convergence of the sequence of objective function values, and not the convergence of the iterates themselves (thus, the results is weaker than the one of [76]).

In this chapter, we aim at solving a smooth optimization problem of the form

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \quad f(\boldsymbol{x}), \tag{5.1}$$

where $f : \mathbb{R}^N \mapsto \mathbb{R}$ is (Fréchet) differentiable but non necessarily convex. In the context of inverse problems in imaging, f typically reads as the sum of a data fidelity term (e.g., a least-squares term) measuring the discrepancy between an acquired, degraded (e.g., blurry, noisy) image, and its estimate (usually, through a linear observation operator), and a regularization term incorporating prior information on the sought solution [68, 17] (see also our Section 5.5). We introduce the block delayed MM memory gradient (BD3MG) algorithm for the resolution of Problem (5.1). BD3MG is a distributed MM algorithm designed for an efficient implementation on a multi-CPU computing architecture, such as a high performance calculation unit. Our contributions are:

- Introduction of the BD3MG algorithm, that implements an advanced distributed asynchronous update rule within the block alternating MM method we recently proposed in [56].
- Proof for the convergence of BD3MG iterates to a stationary point of f under mild assumptions (in particular, no convexity is assumed), using recent tools of Lyapunov analysis [235].
- Illustration of the performance of BD3MG by means of various experiments on a real inverse problem of 3D image restoration arising in the context of multiphotonic miscroscopy.

The chapter is organized as follows. Section 5.2 introduces our notations, recalls the principle of MM schemes and finally presents our proposed algorithm. Section 5.3 states our mathematical assumptions for the convergence analysis and presents preliminary technical propositions and lemmas. Section 5.4 presents our main theoretical contribution, dedicated to the convergence analysis of the proposed BD3MG scheme. Section 5.5 illustrates the qualitative and computational performance of BD3MG in the applicative context of 3D image deblurring in the presence of a depth-variant 3D blur. Section 5.6 concludes the chapter.

5.2 Proposed algorithm

5.2.1 Notations

Throughout this chapter, we consider the euclidean space $\mathbb{R}^N$ endowed with the usual scalar product $\langle\cdot|\cdot\rangle$ (or, equivalently, $\cdot^\top\cdot$) and the norm $\|\cdot\|$. $\mathbf{0}_N$ is the vector with null entries of $\mathbb{R}^N$. $\boldsymbol{I}_N$ is the identity matrix of $\mathbb{R}^N$. We use the short notation $[\![1, N]\!]$, to denote $\{1, 2, \ldots, N\}$, i.e. the set of integers from

1 to N. $\mathbb{S}^N$ denotes the set of symmetric matrices of $\mathbb{R}^{N\times N}$, and $\mathbb{S}^N_+$ (resp. $\mathbb{S}^N_{++}$) the set of positive (resp definite positive) symmetric matrices. Given some $\boldsymbol{M} \in \mathbb{S}^N_{++}$, we denote by $\|\cdot\|_{\boldsymbol{M}}$ the induced weighted euclidean norm, such that, for all $\boldsymbol{v} \in \mathbb{R}^N$, $\|\boldsymbol{v}\|^2_{\boldsymbol{M}} = \boldsymbol{v}^\top \boldsymbol{M}\boldsymbol{v}$. We use the Loewner orders symbols $\prec$ and $\preceq$, to compare real symmetric matrices $(\boldsymbol{A}, \boldsymbol{B}) \in (\mathbb{S}^N)^2$ i.e., $\boldsymbol{A} \preceq \boldsymbol{B}$ (resp. $\boldsymbol{A} \prec \boldsymbol{B}$) is verified when difference $\boldsymbol{B} - \boldsymbol{A}$ belongs to $\mathbb{S}^N_+$ (resp. $\mathbb{S}^N_{++}$).

Let us introduce extra notations, that will be useful to present block coordinate optimization strategy. Most notations hereafter are reminiscent from [56]. Let $\mathcal{S} \subset [\![1, N]\!]$.

- ▷ We denote by $\overline{\mathcal{S}}$ its complementary set $[\![1, N]\!]\backslash\mathcal{S}$, $|\mathcal{S}|$ its cardinal and $\left(\mathbb{R}^{|\mathcal{S}|}, \langle\cdot,\cdot\rangle\right)$ the resulting euclidean space (with a slight abuse of notation). Moreover, we also denote by $\mathbb{S}^{|\mathcal{S}|}, \mathbb{S}^{|\mathcal{S}|}_+, \mathbb{S}^{|\mathcal{S}|}_{++}$ respectively the set of symmetric, symmetric positive, and symmetric definite positive matrices of $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$.

- ▷ Let $\boldsymbol{x} = (x_n)_{n\in[\![1,N]\!]} \in \mathbb{R}^N$. We denote $\boldsymbol{x}_{(\mathcal{S})} = (x_i)_{i\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$ the vector gathering the entries of $\boldsymbol{x}$ with indexes within the set $\mathcal{S}$ of coordinates.

- ▷ Let $\boldsymbol{x} \in \mathbb{R}^N$. $\nabla f(\boldsymbol{x})$ is the gradient of f evaluated at $\boldsymbol{x}$. Moreover, $\nabla_{(\mathcal{S})} f(\boldsymbol{x}) = ([\nabla f(\boldsymbol{x})]_i)_{i\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$ denotes the partial gradient of f with respect to the coordinates with indexes in $\mathcal{S}$, evaluated at $\boldsymbol{x}$.

- ▷ Let $\boldsymbol{M} \in \mathbb{S}^N$. We denote the (symmetric) sub-matrix $\boldsymbol{M}_{(\mathcal{S})} = (M_{i,j})_{(i,j)\in\mathcal{S}^2} \in \mathbb{S}^{|\mathcal{S}|}$. If $\boldsymbol{M}_{(\mathcal{S})} \in \mathbb{S}^{|\mathcal{S}|}_+$, we define the induced weighted Euclidean norm as $\|\cdot\|_{\boldsymbol{M}_{(\mathcal{S})}}$.

- ▷ For any $\boldsymbol{x} \in \mathbb{R}^N$, we introduce the restriction of f to the block $\mathcal{S}$ and vector $\boldsymbol{x}$ as the function $f_{(\mathcal{S})}(.\ , \boldsymbol{x}) : \boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|} \mapsto f(\boldsymbol{u})$ where $\boldsymbol{u}$ is related to $(\boldsymbol{v}, \boldsymbol{x})$ through the relations $\boldsymbol{u}_{(\mathcal{S})} = \boldsymbol{v}$ and $\boldsymbol{u}_{(\overline{\mathcal{S}})} = \boldsymbol{x}_{(\overline{\mathcal{S}})}$.

5.2.2 Block MM principle

MM approach to the resolution of Problem (5.1) is a generic iterative procedure where each iteration amounts to minimizing (exactly or not) a surrogate for f satisfying a majorizing property [223, 246, 128, 126]. The theoretical and practical properties of an MM algorithm greatly depend on (i) the family of considered surrogates, (ii) the procedure to minimize it. In this chapter, we focus on quadratic MM techniques, where f is such that it can be upper bounded by quadratic functions (typically, f is Lipschitz differentiable). In such context, the inner step of an MM algorithm amounts to minimizing a quadratic function on $\mathbb{R}^N$ or, otherwise stating, to invert an $N \times N$ system. In the large scale context, this is not desirable and various approaches have been proposed to cope with the curse of dimensionality in MM quadratic methods [58, 59, 63, 121, 43, 222, 56]. In particular, to limit the dependence of the MM algorithm on the dimension of the problem, block alternating approaches have been developed. In these schemes, at each iteration only a subset of the variables are updated [121], giving rise to so-called block MM algorithms, that we describe hereafter.

Define a partition $\mathbb{T}$ of $[\![1, N]\!]$. The block MM approach requires to build a majorizing surrogate for the restriction $f_{(\mathcal{S})}(\cdot, \boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^N$ and block index $\mathcal{S} \in \mathbb{T}$. Let us assume the existence of a

mapping $\boldsymbol{A} : \boldsymbol{x} \in \mathbb{R}^N \mapsto \boldsymbol{A}(\boldsymbol{x}) \in \mathbb{S}^N_{++}$ such that for all $\mathcal{S} \in \mathbb{T}$, $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ and $\boldsymbol{x} \in \mathbb{R}^N$

$$Q_{\mathcal{S}}(\boldsymbol{v}, \boldsymbol{x}) = f(\boldsymbol{x}) + \langle \nabla_{(\mathcal{S})} f(\boldsymbol{x}), \boldsymbol{v} - \boldsymbol{x}_{(\mathcal{S})} \rangle + \frac{1}{2} \left\| \boldsymbol{v} - \boldsymbol{x}_{(\mathcal{S})} \right\|^2_{\boldsymbol{A}_{(\mathcal{S})}(\boldsymbol{x})}, \tag{5.2}$$

fulfills the majorizing condition

$$Q_{\mathcal{S}}(\boldsymbol{v}, \boldsymbol{x}) \geq f_{(\mathcal{S})}(\boldsymbol{v}, \boldsymbol{x}). \tag{5.3}$$

Note that, by (5.2),

$$Q_{\mathcal{S}}(\boldsymbol{x}_{(\mathcal{S})}, \boldsymbol{x}) = f_{(\mathcal{S})}(\boldsymbol{x}_{(\mathcal{S})}, \boldsymbol{x}). \tag{5.4}$$

The existence of such mapping $\boldsymbol{A}$ can be ensured under mild assumptions. For instance, it is satisfied as soon as f is Lipschitz differentiable. Morever, [65, Remark 2.4] shows that, as soon as the above mapping holds for $\mathcal{S} = \mathbb{R}^N$, it stays valid for any block subset $\mathcal{S} \subset [\![1, N]\!]$. Examples of constructions of majorization mappings have been extensively discussed in [223, 59, 222] for optimization problems arising in the fields of inverse problems, image processing and telecommunication.

Once the block majorization approximations (5.2) satisfying (5.3) are built, the block MM (B2M) algorithm reads [128] (also called BSUM in [121]):

$$(\forall k \in \mathbb{N}) \quad \begin{cases} \text{Choose } \mathcal{S}^k \in \mathbb{T}, \\ \boldsymbol{x}^{k+1}_{(\mathcal{S}^k)} \in \underset{\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}^k|}}{\arg\min}\, Q_{\mathcal{S}^k}(\boldsymbol{v}, \boldsymbol{x}^k), \\ \boldsymbol{x}^{k+1}_{(\overline{\mathcal{S}^k})} = \boldsymbol{x}^k_{(\overline{\mathcal{S}^k})}. \end{cases} \tag{B2M}$$

Hereabove, $\left(\mathcal{S}^k\right)_{k \in \mathbb{N}}$ is a sequence of subsets (i.e., blocks) chosen in the predefined partition $\mathbb{T}$. The most current strategy is to adopt a cyclic rule, where each element of $\mathbb{T}$ is selected sequentially until the end of the partition list, and then the loop is repeated until convergence of the algorithm. A more flexible option is to adopt a so-called quasi-cyclic (or acyclic) rule where each $\mathcal{S} \in \mathbb{T}$ must be updated at least once per K iterations period.

The interest of scheme (B2M) and more generally block coordinate methods notably lies in the large scale context involving a very huge N, for which dealing with all the coordinates of the current iterate may be too high time consuming and even infeasible due to memory limitations. However, block MM methods require a sequential update of the blocks and thus, by construction, might require many iterations to reach convergence. To limit this issue, (block) diagonal mappings have been considered for instance in [222, 43]. The underlying idea is to choose the mapping so that the inner minimization problem in (B2M) is separable, and thus can be performed in parallel over the entries of the selected block. This yields the so-called block parallel MM schemes that take advantage of recent technological advances in parallel computing on multicore architectures. In particular, these methods can tailor the number of available processors to the computational load. However, such block diagonal structure may be detrimental to the approximation quality of the surrogates, and thus reduce again the practical convergence rate. In the present chapter, we opt for not making any extra structural assumption on the majorization mapping, thanks to the introduction of two catalizers into (B2M), namely (i) a subspace acceleration approach, (ii) a distributed asynchronous update strategy, that we describe hereafter.

5.2.3 *Subspace acceleration*

Our first catalyst is to introduce a subspace acceleration [223], in (B2M). This strategy has been initially introduced for full-batch MM algorithms (i.e., without any block coordinate strategy) in [58]. Convergence analysis can be found in [59, 57, 63, 62, 60] under various situations. We recently extended this strategy to cope with block coordinate updates with the form of (B2M) [56], leading to the B2MS (**B**lock **MM** **S**ubspace) scheme that we present hereafter.

Starting with the (B2M) iteration, the subspace acceleration consists in performing the minimization of the majorization function within the current block $\mathcal{S}^k$ in a constrained vectorial subspace spanned by a small number $M_k \geq 1$ of search directions. This reads:

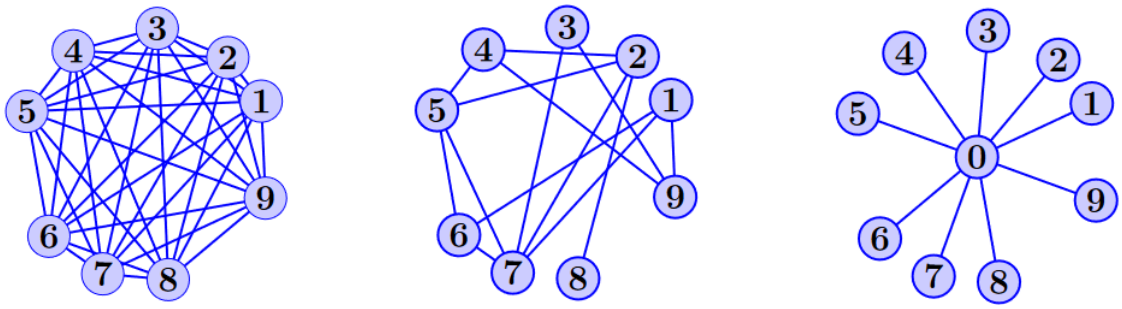
$$(\forall k \in \mathbb{N}) \quad \begin{cases} \text{Choose } \mathcal{S}^k \in \mathbb{T}, \\ \text{Choose } \boldsymbol{D}^k \in \mathbb{R}^{M_k \times |\mathcal{S}^k|}, \\ \boldsymbol{v}^k \in \underset{\boldsymbol{v} \in \mathbb{R}^{M_k}}{\arg\min}\, Q_{\mathcal{S}^k}(\boldsymbol{x}^k_{(\mathcal{S}^k)} + \boldsymbol{D}^k \boldsymbol{v}, \boldsymbol{x}^k), \\ \boldsymbol{x}^{k+1}_{(\mathcal{S}^k)} = \boldsymbol{x}^k_{(\mathcal{S}^k)} + \boldsymbol{D}^k \boldsymbol{v}^k, \\ \boldsymbol{x}^{k+1}_{(\overline{\mathcal{S}^k})} = \boldsymbol{x}^k_{(\overline{\mathcal{S}^k})}, \end{cases} \quad . \tag{B2MS}$$

Hereabove, for every $k \in \mathbb{N}$, $\boldsymbol{D}^k \in \mathbb{R}^{M_k \times |\mathcal{S}^k|}$ is the so-called subspace matrix. It stacks, row-wise, $M_k \geq 1$ vectors of dimension $|\mathcal{S}^k|$, spanning a vectorial subspace within which we seek for a minimizer of the majorization function $Q_{\mathcal{S}^k}(.\,, \boldsymbol{x}^k)$ (i.e., our next iterate). The advantage is to reduce again the dimensionality of the inner MM problems, without jeopardizing the convergence rate [62]. Several choices for the subspace matrix are discussed in [58, 63, 60]. Intensive comparisons in the fields of inverse problems, image processing and machine learning (e.g., [102, 57]), have shown the superiority of the so-called memory gradient subspace which seems to reach the best compromise between simplicity and efficiency. In the context of (B2MS), this amounts to defining, for every $k \in \mathbb{N}$, the memory gradient matrix $\boldsymbol{D}^k = \left[\nabla_{(\mathcal{S}^k)} f(\boldsymbol{x}^k), \boldsymbol{x}^k_{(\mathcal{S}^k)} - \boldsymbol{x}^{k-1}_{(\mathcal{S}^k)}\right]$ (with the convention $\boldsymbol{x}_{-1} = \boldsymbol{0}_N$), so that $M_k = 2$. When combined with a block diagonal majorization mapping, (B2MS) becomes equivalent to the BP3MG method considered in [49] for 3D image deblurring. The convergence properties of (B2MS) have recently been studied in [56].

5.2.4 *Block Delayed Majorize-Minimize Memory Gradient (BD3MG)*

The second catalyst we introduce is the main contribution of this chapter, namely the introduction of a distributed asynchronous update rule within (B2MS). Our motivation is to make the algorithm well suited to an implementation on a multi-core / multi-processor architecture, while not being endangered by potential communication delays within the computing units. Let us consider a computing architecture with C units (or cores), each of them being able to communicate (i.e., send or receive) information to a master node. The architecture thus considered is forming a *star* graph as presented in Figure (5.1c). The two other graph topologies are discarded from this present study (see, for example, [1] for an efficient distributed method running on a generic hypergraph topology).

The proposed method BD3MG is presented in Algorithms 5.1-5.2, describing the iterations of the master (i.e., node 0) and a given worker/node $c \in [\![1, C]\!]$, respectively. Let us describe these two



(a) Complete graph (b) Connected graph (c) Star graph

Figure 5.1: Examples of graph topologies. The graph in (c) is encompassed by our framework.

algorithms. Each computation node $c \in [\![1, C]\!]$ updates (independently from the other nodes) a subset of coordinates $\mathcal{S}_c \in \mathbb{T}$ (which can change over the process) by applying an MM iteration including a memory gradient acceleration and thus "books" its running block $\mathcal{S}_c$ so that no other worker overwrites the associated coordinates. Conversely, any other $\mathcal{S} \in \mathbb{T} \backslash \{\mathcal{S}_c\}$ remains free to be updated by other workers. Communication steps are performed in order to maintain convergence to a minimizer of the globally shared objective function f and to control the propagation of errors. Basically, even if the other workers are still busy on their tasks, every time a worker $c \in [\![1, C]\!]$ ends one MM iteration on its running block $\mathcal{S}_c$, it sends a feedback to the master. As a response, the latter updates it with most recent available information, and assigns it a new task.

We denote $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ the sequence of iterates gathered by the master loop. For any given node index $c \in [\![1, C]\!]$ and $k \in \mathbb{N}$, $\mathcal{S}_c^k$ denotes the block of coordinates processed by worker c during step k. We impose, by construction, that two nodes do not update the same block of coordinates at the same time, so that we ensure the no-overlap condition

$$(\forall k \in \mathbb{N}) \quad (\forall (c, c') \in [\![1, C]\!]^2, c \neq c') \quad \mathcal{S}_c^k \cap \mathcal{S}_{c'}^k = \emptyset. \tag{5.5}$$

At iteration $k \in \mathbb{N}$, worker $c^k \in [\![1, C]\!]$, updates the block of coordinates $\mathcal{S}_{c^k}^k$ and sends to the master a vector $\boldsymbol{d}_k$ of size $\left|\mathcal{S}_{c^k}^k\right|$. The corresponding indexes of variable $\boldsymbol{x}^k$ within block $\mathcal{S}_{c^k}^k$ are then incremented with $\boldsymbol{d}_k$ while the others remain unchanged, thus defining $\boldsymbol{x}^{k+1}$. The master then defines a new set of coordinates $\mathcal{S}_{c^k}^{k+1}$ to be treated by worker c^k, so as to satisfy the non-overlap rule. The master informs worker c^k of this new running set of coordinates, and sends him the most recent information $\boldsymbol{x}^{k+1}$ and the difference $(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k)_{(\mathcal{S}_{c^k}^{k+1})}$. Meanwhile, the other workers keep processing their allocated indexes. The master then waits until a new worker (possibly the same one) sends a new increment.

Let us now make a focus on the worker loop described in Algorithm 5.2. Remark that, even if worker c has access to some properties of function f (i.e., the expression for its gradient and for its majorizing approximation $(Q_{\mathcal{S}})_{\mathcal{S}\in\mathbb{S}}$), it is not informed about the work done by the master or those of the other workers . It can only rely on the data the master sends to it to perform its local task. From the viewpoint of the worker, a triplet set $(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d}) \in \mathbb{R}^N \times \mathbb{T} \times \mathbb{R}^{|\mathcal{S}|}$ is received from the master and must be used to perform its MM update with memory gradient acceleration. The worker is in charge

Algorithm 5.1. *BD3MG algorithm - Master loop*

Initialization.

(a) *Set* $k = 0$ *and* $\boldsymbol{x}^0 \in \mathbb{R}^N$.

(b) *Set* $\mathcal{S}_1^0, ..., \mathcal{S}_C^0 \in \mathbb{T}$ *such that* $\forall (c, c^{'}) \in [\![1, C]\!]^2,\ \mathcal{S}_c^0 \cap \mathcal{S}_{c'}^0 = \emptyset$.

(c) 0-*th transmission: For every* $c \in [\![1, C]\!]$, *send* $\left(\boldsymbol{x}^0, \mathcal{S}_c^0, 0_{|\mathcal{S}_c^0|}\right)$ *to worker* c

While a stopping criterion is not met:

(Wait for a feedback from a worker)

(a) $(k+1)$-*th reception: Receive* $\boldsymbol{d}_k$ *from a worker* c^k.

(b) *Update* $\boldsymbol{x}^{k+1}_{\left(\mathcal{S}^k_{c^k}\right)} = \boldsymbol{x}^k_{\left(\mathcal{S}^k_{c^k}\right)} + \boldsymbol{d}_k$ *and* $\boldsymbol{x}^{k+1}_{\left(\overline{\mathcal{S}^k_{c^k}}\right)} = \boldsymbol{x}^k_{\left(\overline{\mathcal{S}^k_{c^k}}\right)}$.

(c) *Set* $\mathcal{S}_1^{k+1}, \ldots, \mathcal{S}_C^{k+1} \in \mathbb{T}$ *such that* $\mathcal{S}_{c^k}^{k+1} \in \mathbb{T} \backslash \{\mathcal{S}_c^k\}_{c \neq c^k}$ *and,* $(\forall c \in [\![1, C]\!] \backslash c^k)$, $\mathcal{S}_c^{k+1} = \mathcal{S}_c^k$.

(d) $(k+1)$-*th transmission: Send* $\left(\boldsymbol{x}^{k+1}, \mathcal{S}_{c^k}^{k+1}, (\boldsymbol{x}^{k+1} - \boldsymbol{x}^k)_{\left(\mathcal{S}_{c^k}^{k+1}\right)}\right)$ *to worker* c^k.

(e) $k = k + 1$

End While

Output. Vector $\boldsymbol{x}^k$.

of first building the new memory gradient matrix

$$\boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d}) = [-\nabla_{(\mathcal{S})} f(\boldsymbol{x}) \mid \boldsymbol{d}] \in \mathbb{R}^{|\mathcal{S}| \times 2}. \tag{5.6}$$

and then compute the MM increment $\boldsymbol{d}^{'} \in \mathbb{R}^{|\mathcal{S}|}$ defined as

$$\boldsymbol{d}^{'} = \boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})\boldsymbol{u} \tag{5.7}$$

$$\text{with } \boldsymbol{u} \in \underset{\boldsymbol{v} \in \mathbb{R}^2}{\arg\min}\, Q_{\mathcal{S}}(\boldsymbol{x} + \boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})\boldsymbol{v}, \boldsymbol{x}). \tag{5.8}$$

Note that the uniqueness of the solution for problem (5.7)-(5.8) is not guaranteed in general. To overcome such an obstacle, we follow the strategy in [58], and retain the pseudo-inverse solution given by

$$\boldsymbol{u} = -\left(\boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})^\top \boldsymbol{A}_{(\mathcal{S})}(\boldsymbol{x})\ \boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})\right)^\dagger \boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})\ \nabla_{(\mathcal{S})} f(\boldsymbol{x}), \tag{5.9}$$

where † referes to the Moore-Penrose pseudo-inverse. Such solution notably verifies the normal equation

$$\left\langle \nabla_{(\mathcal{S})} f(\boldsymbol{x}), \boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})\boldsymbol{u} \right\rangle = -\|\boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})\boldsymbol{u}\|^2_{\boldsymbol{A}_{(\mathcal{S})}(\boldsymbol{x})}. \tag{5.10}$$

Algorithm 5.2. *BD3MG algorithm - Worker loop*

(Wait for a feedback from the master)

(a) `Receive` $(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})$ *from Master.*

(b) $\boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d}) = [-\nabla_{(\mathcal{S})} f(\boldsymbol{x}) \mid \boldsymbol{d}]$.

(c) Compute $\nabla_{(\mathcal{S})} f(\boldsymbol{x})$ *and* $\boldsymbol{A}_{(\mathcal{S})}(\boldsymbol{x})$.

(d) $\boldsymbol{u} = -\left(\boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})^\top \boldsymbol{A}_{(\mathcal{S})}(\boldsymbol{x})\ \boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})\right)^\dagger \boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})\ \nabla_{(\mathcal{S})} f(\boldsymbol{x})$.

(e) `Send` $\boldsymbol{d}' = \boldsymbol{D}(\boldsymbol{x}, \mathcal{S}, \boldsymbol{d})\boldsymbol{u}$ *to the Master.*

5.2.5 Distributed structure of BD3MG

We first have to make an important remark, regarding the communication load in terms of memory, in between the master and the workers. Consider a worker associated to the block index $\mathcal{S}$. According to Algorithm 5.2, the worker receives three quantities, namely $\boldsymbol{x}$ of N real values, the set of integer indexes $\mathcal{S}$ with cardinality $|\mathcal{S}|$ and the vector $\boldsymbol{d}$ of $|\mathcal{S}|$ real values. The sent vector $\boldsymbol{d}'$ is again made of $|\mathcal{S}|$ real values. Clearly, the main memory load is related to the reception of vector $\boldsymbol{x}$. One should however notice that the worker only uses $\boldsymbol{x}$ to compute $\nabla_{(\mathcal{S})} f(\boldsymbol{x})$ and $\boldsymbol{A}_{(\mathcal{S})}(\boldsymbol{x})$. In most situations encountered in inverse problems of imaging, f shows some inherent separable structure, so that both of these quantities only depend on a subset of entries of vector $\boldsymbol{x}$ that can be of small cardinality compared to N. The practical implementation of Algorithm 5.2 should account for this specific situation, in order to avoid memory saturation and important communication delays. We give a detailed analysis for this aspect, in the case of our experimental example, in the Section 5.5.1.4.

The proposed distributed structure of BDM3G follows a star graph. Practically, it means that one of the computing unit has a higher load, in terms of memory, since it must process the full vector $\boldsymbol{x}$ of size N, while the memory load of the workers is limited, as we discussed hereabove. This can be viewed as a limitation for the proposed method. The extension of our analysis to the case of a hypergraph distributed framework would require to be more specific about the structure of function f (in the line of the study of [1]), which might reduce the versatility of the algorithm. Up to our knowledge, this analysis is not straightforward and is thus left as future work.

5.2.6 Equivalent form for BD3MG

The way we introduced our scheme BD3MG in the previous subsection was "implementation-oriented". In order to study its convergence behaviour, we must exhibit an equivalent form of it, mimicking the one of its non distributed counterpart, (B2MS). To do so, it is necessary to formalize the information gap between the master and the workers during the iterative process.

As we have already mentioned, all the information available to a worker (except those on f and $(Q_{\mathcal{S}})_{\mathcal{S}\in\mathbb{T}}$) is sent to it by the master only after it produces a feedback. For a given $k \in \mathbb{N}$, worker c^k does not receive any information between the $(k+1)$-reception and the previous one it made. During this time, its counterparts $c \in [\![1, C]\!] \backslash \{c^k\}$ may have performed additional updates to the master without c^k being informed. This results in an information mismatch, that we propose to formalize through a vector $\boldsymbol{x}^{\iota_k}$ where

$$(\forall k \in \mathbb{N}) \quad \iota_k = \begin{cases} 0 \text{ if } k = 0, \\ \max\left(\left\{\ell \in [\![1, k]\!] \mid c^{\ell-1} = c^k\right\} \cup \{0\}\right), \text{ otherwise.} \end{cases} \tag{5.11}$$

This vector corresponds to the iteration index of the working variable of worker c^k, which does not necessarily matches with the vector $\boldsymbol{x}^k$ manipulated by the master.
Let us list herebelow some situations of interest given the value of ι_k at some iteration $k \in \mathbb{N}$:

- If $\iota_k = 0$, and $k > 0$, it means that $\left\{\ell \in [\![1, k]\!] \mid c^{\ell-1} = c^k\right\}$ is an empty set. Hence, the worker c^k never returned any feedback to the master before the iteration k. Note that $\iota_0 = 0$ by construction.

- If $\iota_k = k$, we thus have $c^{k-1} = c^k$. Hence, worker c^k was in charge of the two most recent updates, namely the $(k+1)$-th and the k-th ones. As a consequence, to prepare the $(k+1)$-th update, worker c^k received vector $\boldsymbol{x}^k$ from the master.

- More generally, if $\iota_k > 0$, it follows that worker c^k at least returned one feedback to the master before iteration k. And we have the relation $c^{\iota_k - 1} = c^k$.

Moreover, the non-overlap rule translates into

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}^{\iota_k}_{(\mathcal{S}^k_{c^k})} = \boldsymbol{x}^k_{(\mathcal{S}^k_{c^k})}. \tag{5.12}$$

For instance, if $\iota_k = k - 1$ for some $k > 1$, this indicates that $c^{k-2} = c^k$ and $c^{k-1} \neq c^k$. The worker c^k thus proceeded to the $(k-1)$-th and $(k+1)$-th reception of the master while the k-th was made by another $\tilde{c}^k$ who received the vector $\boldsymbol{x}^k$ (from the master). However, since worker c^k was still processing block $\mathcal{S}^{k-1}_{c^k}$, the master was not able to update the associated coordinate for computing $\boldsymbol{x}^k$ from $\boldsymbol{x}^{k-1}$ for worker c^k, i.e $\boldsymbol{x}^{k-1}_{(\mathcal{S}^k_{c^k})} = \boldsymbol{x}^k_{(\mathcal{S}^k_{c^k})}$, which is typical from an asynchronous scheme.

More generally, when it comes to dealing with asynchronous algorithms, the use of a specific indexes with similar roles than our ι_k $(k \in \mathbb{N})$ is often necessary to build a theoretical delay model and thus to formulate an equivalent scheme being more compact and easier to analyse [76].

With this aim in mind, let us introduce the shorter notations

$$(\forall k \in \mathbb{N}) \quad \begin{cases} \mathcal{B}^k = \mathcal{S}^k_{c^k}, \\ \boldsymbol{D}^k = \boldsymbol{D}\left(\boldsymbol{x}^{\iota_k}, \mathcal{B}^k, \left(\boldsymbol{x}^{\iota_k} - \boldsymbol{x}^{\iota_k - 1}\right)_{(\mathcal{B}^k)}\right), \end{cases} \tag{5.13}$$

and $\boldsymbol{D}^k = \boldsymbol{D}\left(\boldsymbol{x}^{\iota_k}, \mathcal{B}^k, \left(\boldsymbol{x}^{\iota_k} - \boldsymbol{x}^{\iota_k - 1}\right)_{(\mathcal{B}^k)}\right)$ with convention $\boldsymbol{x}^{-1} = \boldsymbol{0}_N$. Then, the master/worker BD3MG loops from Algorithms 5.1-5.2 can be rewritten equivalently in a single compact scheme as:

$$(\forall k \in \mathbb{N}) \quad \begin{cases} \text{Let } c^k \in [\![1, C]\!], \\ \boldsymbol{u}^k = -\left((\boldsymbol{D}^k)^\top \boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^{\iota_k})\, \boldsymbol{D}^k\right)^\dagger (\boldsymbol{D}^k)^\top \nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k}), \\ \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} = \boldsymbol{x}^k_{(\mathcal{B}^k)} + \boldsymbol{D}^k \boldsymbol{u}^k, \\ \boldsymbol{x}^{k+1}_{(\overline{\mathcal{B}^k})} = \boldsymbol{x}^k_{(\overline{\mathcal{B}^k})}, \end{cases} \tag{5.14}$$

where we noticed that (5.12) now reads (using (5.13))

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}^{\iota_k}_{(\mathcal{B}^k)} = \boldsymbol{x}^k_{(\mathcal{B}^k)}. \tag{5.15}$$

For every $k \in \mathbb{N}$, according to (5.14), $\boldsymbol{u}_k$ still reads (5.9) and thus verifies (5.10) with $\boldsymbol{D}^k = \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}$. The optimality equation can be rewritten as:

$$\left(\nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k})\right)^\top \left(\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\right) = -\left\|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\right\|^2_{\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^{\iota_k})}. \tag{5.16}$$

The two next Sections are dedicated to establish the convergence of the iterates produced by (5.14).

5.2.7 Link with existing works

Let us discuss the links between our proposed scheme BD3MG and existing methods from the literature. When $\iota_k = k$ for any $k \in \mathbb{N}$ in BD3MG, the algorithm identifies with our block alternating scheme B2MS [56] where the blocks of variables were updated sequentially in a non parallel (thus, not asynchronous) manner. This present chapter can thus be viewed as an extension of the framework and of the convergence analysis of [56] to the distributed setting. Other related methods are [76, 45, 154], and our convergence analysis relies on similar tools than the one from [76]. Assuming zero-valued non-smooth terms in [76, 45, 154] (i.e., the objective function is differentiable), these methods identify with particular instances of BD3MG that (i) would not implement any subspace acceleration (i.e., $\boldsymbol{D}_k = \boldsymbol{I}_N$ in (5.14)), (ii) would rely on the simple Lipschitz-based majorization metric (i.e., $\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^{\iota_k}) = \mathcal{L}\boldsymbol{I}_{|\mathcal{B}^k|}$ in (5.14)) in the case of [76]. As a consequence, assuming differentiability of all terms, our convergence analysis presented in the next section thus also covers the schemes of [76, 45, 154]. Up to our knowledge, this chapter is the first to show convergence of the iterates of a distributed MM algorithm involving generic quadratic surrogates and subspace acceleration, in the non-convex setting. Finally, we would like to point out that the 3MG update performed in Alg. 5.2 identifies with a non-linear conjugate gradient (NLCG) update, for a specific (and closed form) pair of stepsize and conjugacy parameters (see discussion in [60, Sec. 1]). Therefore, the work in this chapter can also be understood as the first convergence analysis of a distributed NLCG method in the non-convex setting. A comparative study will be conducted in our experimental section to illustrate the superiority of BD3MG with respect to the aforementioned existing methods in terms of convergence speed.

5.3 Assumptions and preliminary results

5.3.1 Assumptions

In order to analyse the asymptotic behaviour of the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ generated by scheme (5.14), we introduce technical assumptions both on function f and on the parameters of BD3MG method.

Assumption 5.1. *Function f is coercive, continuously differentiable on $\mathbb{R}^N$, and has a $\mathcal{L}$-Lipschitzian gradient with $\mathcal{L} > 0$, i.e.*

$$(\forall(\boldsymbol{x},\boldsymbol{y}) \in (\mathbb{R}^N)^2) \qquad \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq \mathcal{L}\|\boldsymbol{x}-\boldsymbol{y}\|. \tag{5.17}$$

Assumption 5.1 ensures the existence of solutions for Problem (5.1) (by coercivity). Moreover, (5.17) in Assumption 5.1 guarantees the existence of a quadratic function (5.2) satisfying (5.3), setting $\boldsymbol{A} : \boldsymbol{x} \mapsto \mathcal{L}\boldsymbol{I}_N$. Another direct consequence is

$$(\forall \mathcal{S} \in \mathbb{T})(\forall(\boldsymbol{x},\boldsymbol{y}) \in (\mathbb{R}^N)^2) \qquad \|\nabla_{(\mathcal{S})} f(\boldsymbol{x}) - \nabla_{(\mathcal{S})} f(\boldsymbol{y})\| \leq \mathcal{L}\|\boldsymbol{x}-\boldsymbol{y}\|, \tag{5.18}$$

since $\|\nabla_{(\mathcal{S})} f(\boldsymbol{x}) - \nabla_{(\mathcal{S})} f(\boldsymbol{y})\| \leq \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|$ for all $\mathcal{S} \in \mathbb{T}$ and $(\boldsymbol{x},\boldsymbol{y}) \in (\mathbb{R}^N)^2$.

Assumption 5.2. *(Boundedness of delay) For every $k \in \mathbb{N}$, and every worker $c^k \in [\![1, N]\!]$, the set $\mathcal{S}_c^k$ is not empty and there exists $\tau \in \mathbb{N}^*$ such that*

$$(\forall k \geq \tau) \quad [\![1, N]\!] = \bigcup_{j=k-\tau}^{k-1} \mathcal{B}^j, \tag{5.19}$$

using the notation in (5.13).

Assumption 5.2 gives an upper bound on the delay τ. Each of the block of variables should be updated within a time frame of at most τ iterations and thus the workers must follow a certain regularity. Such a condition follows a similar goal than quasi-cyclic rule frequently assumed in block coordinate methods [56, 121]. From Assumption 5.2, we deduce the following Proposition, which appears fundamental for the rest of our convergence study. It guarantees that, for a given $k \in \mathbb{N}$, the vector treated by worker c^k before its feedback (i.e the $(k+1)$-th master's reception) is not "too old" compared to the iteration index.

Proposition 5.1. *Under Assumption 5.2, for every $k \geq \tau$, the index ι_k given in* (5.11) *belongs to* $[\![k-\tau+1, k]\!]$.

Proof. Let $k \geq \tau$, where $\tau > 0$ is defined in Assumption 5.2. Inequality $\iota_k \leq k$ directly comes from Definition (5.11). We prove the lower bound on ι_k by contradiction. Let us suppose that $\iota_k \leq k - \tau$. Two situations may arise.

Case 1: $\iota_k = 0$. By definition, $c^0, \ldots, c^{k-1} \neq c^k$ and an easy induction gives $\mathcal{S}^0_{c^k} = \ldots = \mathcal{S}^k_{c^k}$. Non-overlap rule (5.5) with $c^0, \ldots, c^{k-1} \neq c^k$ yields

$$(\forall j \in [\![0, k-1]\!]) \quad \mathcal{S}^j_{c^k} \cap \mathcal{S}^j_{c^j} = \mathcal{S}^k_{c^k} \cap \mathcal{B}^j \tag{5.20}$$

$$= \emptyset. \tag{5.21}$$

Since $\mathcal{S}^k_{c^k}$ is non empty by Assumption 5.2, condition (5.20) ensures the existence of some $i_k \in [\![1, N]\!]$ verifying $i_k \notin \bigcup_{j=0}^{k-1} \mathcal{B}^j$ contradicting $\bigcup_{j=k-\tau}^{k-1} \mathcal{B}^j = [\![1, N]\!]$, as $k \geq \tau$.

Case 2: $\iota_k > 0$. We have $c^{\iota_k - 1} = c^k$ and a finite induction leads to

$$(\forall j \in [\![\iota_k, k]\!]) \quad \mathcal{S}^{\iota_k}_{c^k} = \mathcal{S}^{\iota_k}_{c^{\iota_k - 1}} = \mathcal{S}^{j}_{c^{\iota_k - 1}} = \mathcal{S}^{j}_{c^k}. \tag{5.22}$$

Majoration $\iota_k \leq k - \tau$ implies that

$$(\forall j \in [\![k - \tau, k]\!]) \quad \mathcal{S}^{\iota_k}_{c^k} = \mathcal{S}^{j}_{c^k}. \tag{5.23}$$

Non-overlap rule (5.5) with $c^{k-\tau}, \ldots, c^{k-1} \neq c^k$ then gives

$$(\forall j \in [\![k - \tau, k - 1]\!]) \quad \mathcal{S}^{j}_{c^k} \cap \mathcal{S}^{j}_{c^j} = \mathcal{S}^{\iota_k}_{c^k} \cap \mathcal{B}^j \tag{5.24}$$

$$= \emptyset. \tag{5.25}$$

Since $\mathcal{S}^{\iota_k}_{c^k}$ is non empty, Condition (5.24) thus ensures the existence of $i_k \in [\![1, N]\!]$ verifying $i_k \notin \bigcup_{j=k-\tau}^{k-1} \mathcal{B}^j$ which contradicts $\bigcup_{j=k-\tau}^{k-1} \mathcal{B}^j = [\![1, N]\!]$.

□

Assumption 5.3. *(Curvature of majorizing matrix)*

(i) The mapping $\boldsymbol{A} : \boldsymbol{x} \in \mathbb{R}^N \mapsto \boldsymbol{A}(\boldsymbol{x}) \in \mathbb{S}^N_{++}$ *is such that* (5.3) *holds. Moreover, there exists* $\overline{\nu} > 0$ *such that, for all* $\mathcal{S} \in \mathbb{T}$ *and* $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$,

$$0 \prec \boldsymbol{A}_{(\mathcal{S})}(\boldsymbol{v}) \preceq \overline{\nu} \boldsymbol{I}_{|\mathcal{S}|}. \tag{5.26}$$

(ii) There exists $\underline{\nu} > 0$ *such that, for all* $k \in \mathbb{N}$,

$$\boldsymbol{\Gamma}^k_c = \boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^{\iota_k}) - \frac{1}{2}\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^k) \succeq \left(\frac{\mathcal{L}\sqrt{\tau}(1+\tau)}{2} + \underline{\nu} \right) \boldsymbol{I}_{|\mathcal{B}^k|}. \tag{5.27}$$

Assumption 5.3(i) is standard in optimization literature dealing with MM methods involving quadratic surrogates [58]. Assumption 5.3(ii) assumes that the spectrum of the difference between delayed and exact majorizing matrices of the partial quadratic majoring functions is strictly greater than a certain constant. This hypothesis controls the length of the MM increments performed by each worker. It aims at ensuring consistency between the asynchronous updates, by directly relating the worst-case curvature of the function f (parameterized by the Lipschitz constant $\mathcal{L}$) and the worst-case communication delay (parameterized by the constant τ). Condition (5.27) is key to ensure a condition descent for the general process generated by BD3MG scheme (see subsection 5.4.1). Assumption 5.3(ii) becomes redundant with Assumption 5.3(i) in the case when no delay occurs (i.e., $\tau = 0$). A detailed constructive example on how to meet Assumption 5.3(ii) will be provided in our experimental Section 5.5.

5.3.2 Technical lemmas

We conclude this section by presenting some preliminary results that will be useful for our convergence analysis.

Lemma 5.1. *Under Assumption 5.2, for every $k \geq \tau$,*

$$\left\|\boldsymbol{x}^k - \boldsymbol{x}^{\iota_k}\right\|^2 \leq \tau \sum_{j=k-\tau+1}^{k} \left\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\right\|^2. \tag{5.28}$$

Proof. Let $k \in \mathbb{N}$. If $\iota_k = k$, inequality (5.28) is trivial. For the rest of the proof we thus suppose $\iota_k \leq k-1$. According to the definition of the euclidean norm we have

$$\left\|\boldsymbol{x}^k - \boldsymbol{x}^{\iota_k}\right\|^2 = \sum_{n=1}^{N} \left(x_n^k - x_n^{\iota_k}\right)^2. \tag{5.29}$$

Then, for all $n \in [\![1, N]\!]$, the Jensen's inequality leads to

$$\left(x_n^k - x_n^{\iota_k}\right)^2 = \left(\sum_{j=\iota_k+1}^{k} \left(x_n^j - x_n^{j-1}\right)\right)^2 \leq (k - \iota_k) \sum_{j=\iota_k+1}^{k} \left(x_n^j - x_n^{j-1}\right)^2. \tag{5.30}$$

Moreover, Proposition 5.1 ensures that ι_k belongs to $[\![k-\tau+1, k]\!]$. As a consequence

$$(\forall n \in [\![1, N]\!]) \quad \left(x_n^k - x_n^{\iota_k}\right)^2 \leq \tau \sum_{j=k-\tau+1}^{k} \left(x_n^j - x_n^{j-1}\right)^2. \tag{5.31}$$

We then replace (5.31) in (5.29), which yields

$$\left\|\boldsymbol{x}^k - \boldsymbol{x}^{\iota_k}\right\|^2 \leq \tau \sum_{j=k-\tau+1}^{k} \sum_{n=1}^{N} \left(x_n^j - x_n^{j-1}\right)^2. \tag{5.32}$$

Relation (5.28) directly comes from the identification of the inner sum of (5.32) as $\left\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\right\|^2$ for all $j \in [\![k-\tau+1, k]\!]$. □

Lemma 5.1 provides a bound on the residual between $\boldsymbol{x}^k$ and the delayed vector $\boldsymbol{x}^{\iota_k}$ updated by worker c^k at iteration $k \in \mathbb{N}$. The right term in (5.28) can be understood as the extra information available to the master, when compared to the one available to worker c^k. This Lemma will allow to establish a descent condition on the BD3MG process in the next Section.

Lemma 5.2. *Under Assumptions 5.1 and 5.3(i), for every $k \in \mathbb{N}$,*

$$\|\nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k})\|^2 \leq \overline{\nu}^2 \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2. \tag{5.33}$$

Proof. Let $k \in \mathbb{N}$. Let us analyse the quantity $f(\boldsymbol{x}^{\iota_k}) - Q_{\mathcal{B}^k}(\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)}, \boldsymbol{x}^{\iota_k})$.

On the one hand, function $\Psi_k : \alpha \in \mathbb{R} \mapsto Q_{\mathcal{B}^k}(\boldsymbol{x}^k_{(\mathcal{B}^k)} - \boldsymbol{D}^k \alpha \boldsymbol{e}, \boldsymbol{x}^{\iota_k})$ with $\boldsymbol{e} = (1,0)^\top$ is a second degree convex polynomial with a unique minimizer that reads

$$\widehat{\alpha}_k = \frac{\|\nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k})\|^2}{\|\nabla_{(\mathcal{B}^k)} f\left(\boldsymbol{x}^{\iota_k}\right)\|^2_{\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^{\iota_k})}}. \tag{5.34}$$

Since $\boldsymbol{u}^k$ is a minimizer of $Q_{\mathcal{B}^k}(\boldsymbol{x}^{\iota_k}_{(\mathcal{B}^k)} + \boldsymbol{D}^k ., \boldsymbol{x}^{\iota_k}) = Q_{\mathcal{B}^k}(\boldsymbol{x}^k_{(\mathcal{B}^k)} + \boldsymbol{D}^k ., \boldsymbol{x}^{\iota_k})$, with $\boldsymbol{x}^{\iota_k}_{(\mathcal{B}^k)} = \boldsymbol{x}^k_{(\mathcal{B}^k)}$ by Equation (5.11), we deduce that

$$Q_{\mathcal{B}^k}(\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)}, \boldsymbol{x}^{\iota_k}) \leq \Psi_k(\widehat{\alpha}_k) = f(\boldsymbol{x}^{\iota_k}) - \frac{1}{2}\widehat{\alpha}_k \|\nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k})\|^2. \tag{5.35}$$

From Assumption 5.3(i), $\widehat{\alpha}_k$ verifies $\widehat{\alpha}_k \geq \overline{\nu}^{-1}$. Equation (5.35) can thus be rewritten as

$$f(\boldsymbol{x}^{\iota_k}) - Q_{\mathcal{B}^k}(\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)}, \boldsymbol{x}^{\iota_k}) \geq \frac{1}{2\overline{\nu}} \|\nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k})\|^2. \tag{5.36}$$

On the other hand, using (5.15) from Definition (5.2), and Equation (5.16) yield

$$\begin{aligned}
&f(\boldsymbol{x}^{\iota_k}) - Q_{\mathcal{B}^k}(\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)}, \boldsymbol{x}^{\iota_k}) \\
&= \left\langle \nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k}) \ , \ \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^{\iota_k}_{(\mathcal{B}^k)} \right\rangle + \frac{1}{2} \left\| \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^{\iota_k}_{(\mathcal{B}^k)} \right\|^2_{\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^{\iota_k})} \\
&= \left\langle \nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k}) \ , \ \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^{k}_{(\mathcal{B}^k)} \right\rangle + \frac{1}{2} \left\| \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^{k}_{(\mathcal{B}^k)} \right\|^2_{\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^{k})} \\
&= \frac{1}{2} \left\| \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^{k}_{(\mathcal{B}^k)} \right\|^2_{\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^{\iota_k})}.
\end{aligned} \tag{5.37}$$

The combination of (5.36) and (5.37) leads to

$$\|\nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k})\|^2 \leq \overline{\nu} \left\| \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^{k}_{(\mathcal{B}^k)} \right\|^2_{\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^{\iota_k})}. \tag{5.38}$$

Finally, Equation (5.33) comes using Assumption 5.3(i), and in particular,

$$\left\| \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^{k}_{(\mathcal{B}^k)} \right\|^2_{\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^{\iota_k})} \leq \overline{\nu} \left\| \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^{k}_{(\mathcal{B}^k)} \right\|^2 \tag{5.39}$$

$$= \overline{\nu} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2. \tag{5.40}$$

□

Lemma 5.2 generalizes the decreasing behavior of standard MM schemes [59, 56] to the asynchronous context. It is not directly invoked in our main convergence proof but serves as an intermediary to show the following technical result.

Lemma 5.3. *Under Assumptions 5.1 and 5.3(i), for all $k \geq 2\tau$,*

$$\|\nabla f(\boldsymbol{x}^k)\| \leq \mathcal{L}\tau \sum_{j=k-2\tau+1}^{k} \|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\| + \overline{\nu} \sum_{j=k-\tau}^{k} \|\boldsymbol{x}^{j+1} - \boldsymbol{x}^j\|. \tag{5.41}$$

Proof. Let $k \geq 2\tau$. Assumption 5.2 allows us to bound the gradient of f at $\boldsymbol{x}^k$, as

$$\|\nabla f(\boldsymbol{x}^k)\|^2 \leq \sum_{\ell=k-\tau}^{k-1} \|\nabla_{(\mathcal{B}^\ell)} f(\boldsymbol{x}^k)\|^2 \leq \left(\sum_{\ell=k-\tau}^{k-1} \|\nabla_{(\mathcal{B}^\ell)} f(\boldsymbol{x}^k)\| \right)^2. \tag{5.42}$$

Let us extract the root of the above terms, and use triangular and gradient-Lipschitz inequalities, leading to

$$\begin{aligned} \|\nabla f(\boldsymbol{x}^k)\| &\leq \sum_{\ell=k-\tau}^{k-1} \|\nabla_{(\mathcal{B}^\ell)} f(\boldsymbol{x}^k) - \nabla_{(\mathcal{B}^\ell)} f(\boldsymbol{x}^{\iota_\ell})\| + \sum_{\ell=k-\tau}^{k-1} \|\nabla_{(\mathcal{B}^\ell)} f(\boldsymbol{x}^{\iota_j})\| \\ &\leq \mathcal{L} \sum_{\ell=k-\tau}^{k-1} \|\boldsymbol{x}^k - \boldsymbol{x}^{\iota_\ell}\| + \sum_{j=k-\tau}^{k-1} \|\nabla_{(\mathcal{B}^j)} f(\boldsymbol{x}^{\iota_j})\|. \end{aligned} \tag{5.43}$$

For all $\ell \in [\![k-\tau, k-1]\!]$, by Proposition 5.1, $\iota_\ell \geq \ell - \tau + 1 \geq k - 2\tau$. Thus,

$$\|\boldsymbol{x}^k - \boldsymbol{x}^{\iota_\ell}\| \leq \sum_{j=k-2\tau+1}^{k} \|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\|. \tag{5.44}$$

The right term of (5.44) does not depend on index ℓ. Using (5.44) and inequality (5.33) finally proves the result. □

Lemma 5.3 is useful as it provides a bound on the gradient at step $\boldsymbol{x}^k$ only depending on the $2\tau+1$ past iterates $\boldsymbol{x}^k, \ldots, \boldsymbol{x}^{k-2\tau}$.

Lemma 5.4. *Let $(u^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$ be two sequences of positive real numbers. If there exists $P \in \mathbb{N}$ and $k^* \geq P$ such that*

$$(\forall k \geq k^*) \quad u^k \leq r \sum_{j=k-P}^{k-1} u^j + v^{k-1}, \tag{5.45}$$

with $r < 1/P$ and $\sum_{k=0}^{+\infty} v^k < +\infty$, then $\sum_{k=0}^{+\infty} u^k < +\infty$.

Proof. Summing (5.45) from k^* to $n \geq k^*$ leads to

$$\sum_{k=k^*}^{n} u^k \leq r \sum_{k=k^*}^{n} \sum_{j=k-P}^{k-1} u^j + \sum_{k=k^*}^{n} v^{k-1}, \tag{5.46}$$

with

$$\sum_{k=k^*}^{n}\sum_{j=k-P}^{k-1} u^j = \sum_{k=k^*}^{n}\sum_{j=1}^{P} u^{k-j} = \sum_{j=1}^{P}\sum_{k=k^*-j}^{n-j} u^k \leq \sum_{j=1}^{P}\sum_{k=0}^{n} u^k. \tag{5.47}$$

Plugging (5.47) into (5.46), yields

$$\sum_{k=k^*}^{n} u^k \leq rP \sum_{k=k*}^{n} u^k + \left(rP \sum_{k=0}^{k^*-1} u^k + \sum_{k=k^*}^{n} v^{k-1} \right) \leq rP \sum_{k=k*}^{n} u^k + \left(rP \sum_{k=0}^{k^*-1} u^k + \sum_{k=0}^{+\infty} v^k \right), \tag{5.48}$$

that is $(1-rP)\sum_{k=k^*}^{n} u^k \leq rP \sum_{k=0}^{k^*-1} u^k + \sum_{k=0}^{+\infty} v^k$. With $0 < 1-rP < 1$, we deduce the summability of $\left(u^k\right)_{k\in\mathbb{N}}$. □

Lemma 5.4 is a technical result to ensure the convergence of some real series. Several variants of inequality (5.45) have been used to prove the finite length of iterative processes and then their convergence [76, 27].

5.4 Convergence results

Let us now state our main theoretical results, that relate to the convergence properties of BD3MG iterates. Our proof line is reminiscent of [76, 27] and follows similar steps that we summarize hereafter. First, starting from the majoration property (5.3) and using Lemma 5.1, we will establish a descent inequality. The latter is the key point of the rest of the proof. In particular, it will allow to show convergence of $(f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$. Then, Lemma 5.3 will ensure that $(\nabla f(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges to $\mathbf{0}_N$, and usual topological properties will serve to show that the set of cluster points $\mathcal{C}(\boldsymbol{x}^k)_{k\in\mathbb{N}})$ of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ lies in the set of stationary point of f. Finally, we will exhibit a Lyapunov function from our descent inequality and will resort to the Kurdyka-Łojasewicz (KL) inequality [27] to prove our main theorem, showing the convergence of the BDM3G iterates and providing a rate of convergence.

5.4.1 Descent inequality

Proposition 5.2. *Under Assumptions 5.1-5.2-5.3, there exists a positive sequence $(\nu_k)_{k\geq\tau}$ such that*

$$(\forall k \geq \tau) \quad f(\boldsymbol{x}^{k+1}) + \nu_{k+1} \leq f(\boldsymbol{x}^k) + \nu_k - \underline{\nu}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2. \tag{5.49}$$

Proof. By definition of the majorization function (5.3), for every $k \in \mathbb{N}$,

$$f(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{x}^k) + \left\langle \nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^k),\ \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)} \right\rangle + \frac{1}{2}\|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\|^2_{\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^k)}. \tag{5.50}$$

Decomposing the scalar product term then yields, for every $k \in \mathbb{N}$,

$$\begin{aligned} f(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{x}^k) + R_k + \left\langle \nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k}),\ \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)} \right\rangle + \frac{1}{2}\|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\|^2_{\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^k)}, \\ \text{with } R_k = \left\langle \nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^k) - \nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k}),\ \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)} \right\rangle. \end{aligned} \tag{5.51}$$

Let τ defined as in Assumption 5.2. A majoration of R_k for every $k \geq \tau$ comes by using successively Cauchy-Schwartz inequality, $\mathcal{L}$ gradient-Lipschitz inequality from Assumption 5.1, and Lemma 5.1:

$$\begin{aligned}(\forall k \geq \tau) \quad R_k &\leq \mathcal{L}\|\boldsymbol{x}^k - \boldsymbol{x}^{\iota_k}\| \, \|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\| \\ &\leq \frac{\mathcal{L}}{2\sqrt{\tau}}\|\boldsymbol{x}^k - \boldsymbol{x}^{\iota_k}\|^2 + \frac{\mathcal{L}\sqrt{\tau}}{2}\|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\|^2, \\ &\leq \frac{\mathcal{L}\sqrt{\tau}}{2}\sum_{j=k-\tau+1}^{k}\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\|^2 + \frac{\mathcal{L}\sqrt{\tau}}{2}\|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\|^2. \end{aligned} \tag{5.52}$$

We then set, for all $k \geq \tau$, $\nu_k = \frac{\mathcal{L}\sqrt{\tau}}{2}\sum_{j=k-\tau+1}^{k}(j-k+\tau)\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\|^2$. Since $\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)} = \boldsymbol{x}^{k+1} - \boldsymbol{x}^k$, (5.52) also reads

$$\begin{aligned}(\forall k \geq \tau) \quad R_k &\leq \nu_k - \nu_{k+1} + \frac{\mathcal{L}\tau\sqrt{\tau}}{2}\|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\|^2 + \frac{\mathcal{L}\sqrt{\tau}}{2}\|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\|^2, \\ &= \nu_k - \nu_{k+1} + \frac{\mathcal{L}\sqrt{\tau}(1+\tau)}{2}\|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\|^2. \end{aligned} \tag{5.53}$$

Moreover, Equation (5.16) ensures that

$$(\forall k \geq \tau) \quad \left\langle \nabla_{(\mathcal{B}^k)} f(\boldsymbol{x}^{\iota_k}),\ \boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - x^k_{(\mathcal{B}^k)} \right\rangle = -\|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\|^2_{\boldsymbol{A}_{(\mathcal{B}^k)}(\boldsymbol{x}^{\iota_k})}. \tag{5.54}$$

Replacing both (5.53) and (5.54) in (5.51) gives, for all $k \geq \tau$,

$$\begin{aligned} f(\boldsymbol{x}^{k+1}) + \nu_{k+1} &\leq f(\boldsymbol{x}^k) + \nu_k + \frac{\mathcal{L}\sqrt{\tau}(1+\tau)}{2}\|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\|^2 - \|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\|^2_{\boldsymbol{\Gamma}^k_c} \\ &= f(\boldsymbol{x}^k) + \nu_k - \|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\|^2_{\boldsymbol{\Gamma}^k_c - \frac{\mathcal{L}\sqrt{\tau}(1+\tau)}{2}\boldsymbol{I}_{|\mathcal{B}^k|}}, \end{aligned} \tag{5.55}$$

with $\boldsymbol{\Gamma}^k_c$ defined in Assumption 5.3(ii). (5.49) is a direct consequence of Assumption 5.3(ii) remarking that $\|\boldsymbol{x}^{k+1}_{(\mathcal{B}^k)} - \boldsymbol{x}^k_{(\mathcal{B}^k)}\| = \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|$. □

5.4.2 General behaviour

We now state our first convergence Theorem for BD3MG algorithm.

Theorem 5.1. *Under Assumptions 5.1-5.2-5.3, sequence* $\left(f(\boldsymbol{x}^k)\right)_{k\in\mathbb{N}}$ *generated by BD3MG converges to a finite limit* f^∞*. Moreover,* $\left(\nabla f(\boldsymbol{x}^k)\right)_{k\in\mathbb{N}}$ *converges to* $\mathbf{0}_N$*.*

Proof. Coercivity of f (Assumption 5.1) and (5.49) guarantee that $\left(f(\boldsymbol{x}^k) + \nu_k\right)_{k\in\mathbb{N}}$ is a decreasing and lower-bounded sequence. It is thus converging to a real value f^∞. Equation (5.49) then directly leads to $\sum_{k=0}^{+\infty}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 < +\infty$. On the first hand, using the same notation $(\nu_k)_{k\in\mathbb{N}}$ introduced in our proof of Proposition 5.2, we have

$$(\forall k \geq \tau) \quad \nu_k \leq \frac{\mathcal{L}\tau\sqrt{\tau}}{2}\sum_{j=k-\tau+1}^{+\infty}\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\|^2. \tag{5.56}$$

Thus, the sequence $(\nu_k)_{k\in\mathbb{N}}$ converges to 0 and so, by Proposition 5.2, $\left(f(\boldsymbol{x}^k)\right)_{k\in\mathbb{N}}$ goes to f^∞. On the other hand, Lemma 5.3 gives

$$(\forall k \geq 2\tau) \quad \|\nabla f(\boldsymbol{x}^k)\| \leq \mathcal{L}\tau \sum_{j=k-2\tau+1}^{+\infty} \|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\| + \sum_{j=k-\tau}^{+\infty} \|\boldsymbol{x}^{j+1} - \boldsymbol{x}^j\|, \tag{5.57}$$

which enables to conclude that $\left(\nabla f(\boldsymbol{x}^k)\right)_{k\in\mathbb{N}}$ converges to $\mathbf{0}_N$.

□

Proposition 5.3. *Under Assumptions 5.1-5.2-5.3, $\mathcal{C}\left((\boldsymbol{x}^k)_{k\in\mathbb{N}}\right)$, defined as the set of accumulation points of $\left(\boldsymbol{x}^k\right)_{k\in\mathbb{N}}$, is non empty , compact and is a subset of the set of stationary points of f. Moreover, f takes value f^∞ on $\mathcal{C}\left((\boldsymbol{x}^k)_{k\in\mathbb{N}}\right)$.*

Proof. Coercivity of f (by Assumption 5.1) and convergence of $\left(f(\boldsymbol{x}^k)\right)_{k\in\mathbb{N}}$ to f^∞ (by Theorem 5.1) show that $\left(\boldsymbol{x}^k\right)_{k\in\mathbb{N}}$ is a bounded sequence and $\mathcal{C}\left((\boldsymbol{x}^k)_{k\in\mathbb{N}}\right)$ is non empty and compact. Convergence of $\left(\nabla f(\boldsymbol{x}^k)\right)_{k\in\mathbb{N}}$ to $\mathbf{0}_N$ (by Theorem 5.1) guarantees that every point $\boldsymbol{x}^* \in \mathcal{C}\left((\boldsymbol{x}^k)_{k\in\mathbb{N}}\right)$ is a stationary point of f. Moreover, using again $\left(f(\boldsymbol{x}^k)\right)_{k\in\mathbb{N}}$ converging to f^∞ yields $f^\infty = f(\boldsymbol{x}^*)$ for every $\boldsymbol{x}^* \in \mathcal{C}\left((\boldsymbol{x}^k)_{k\in\mathbb{N}}\right)$ which concludes the proof. □

5.4.3 Lyapunov-based asymptotical analysis

In order to establish the convergence of the iterates of BD3MG, we will follow an analysis relying on the use of a Lyapunov function. Such proof technique has also been used in [76, 244, 235]. The idea is to exhibit a function, related to the loss function f but non necessarily equals to it, that decreases monotonically along the iterative process. Given (5.49), a natural choice is

$$L : \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_0 \\ \vdots \\ \boldsymbol{Z}_\tau \end{pmatrix} \in \mathbb{R}^{(\tau+1)N} \mapsto f(\boldsymbol{Z}_0) + \frac{\mathcal{L}\sqrt{\tau}}{2} \sum_{\ell=1}^{\tau} (\tau - \ell + 1)\|\boldsymbol{Z}_\ell - \boldsymbol{Z}_{\ell-1}\|^2. \tag{5.58}$$

Let us denote, for every $k \geq \tau$, $\boldsymbol{Z}^k = \begin{pmatrix} \boldsymbol{x}^k \\ \vdots \\ \boldsymbol{x}^{k-\tau} \end{pmatrix} \in \mathbb{R}^{(\tau+1)N}$, with $\boldsymbol{x}^k$ the k-th BD3MG iterate. Then, the descent condition from Proposition 5.2 can be rewritten as

$$(\forall k \geq \tau) \quad L(\boldsymbol{Z}^{k+1}) \leq L(\boldsymbol{Z}^k) - \underline{\nu}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2. \tag{5.59}$$

The structure of L allows to build an upper bound of its gradient norm along the iterates, where the bound depends only on the differences of the past iterates:

Lemma 5.5. *There exists $\rho > 0$ such that*

$$(\forall k \geq \tau) \quad \|\nabla L(\boldsymbol{Z}^k)\| \leq \rho \sum_{j=k-2\tau+1}^{k} \|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\|. \tag{5.60}$$

Proof. Function L is differentiable. The expression of its gradient is

$$\left(\forall \boldsymbol{Z} \in \mathbb{R}^{(\tau+1)N}\right) \quad \nabla L(\boldsymbol{Z}) = \boldsymbol{g}_0 + \mathcal{L}\sqrt{\tau}\sum_{\ell=1}^{\tau}(\tau-\ell+1)\boldsymbol{g}_\ell, \tag{5.61}$$

$$\text{where } \boldsymbol{g}_0 = \begin{pmatrix} \nabla f(\boldsymbol{Z}_0) \\ \boldsymbol{0}_{\tau N} \end{pmatrix} \text{ and } (\forall \ell \in [\![1,\tau]\!]) \quad \boldsymbol{g}_\ell = \begin{pmatrix} \boldsymbol{0}_{(\ell-1)N} \\ \boldsymbol{Z}_{\ell-1} - \boldsymbol{Z}_\ell \\ \boldsymbol{Z}_\ell - \boldsymbol{Z}_{\ell-1} \\ \boldsymbol{0}_{(\tau-\ell)N} \end{pmatrix}. \tag{5.62}$$

Let us apply twice the Jensen inequality for the square of the norm and then the majoration $\tau-\ell+1 \leq \tau$ for $1 \leq \ell \leq \tau$. This yields

$$\begin{aligned}\left(\forall \boldsymbol{Z} \in \mathbb{R}^{(\tau+1)N}\right) \quad \|\nabla L(\boldsymbol{Z})\|^2 &\leq 2\|\boldsymbol{g}_0\|^2 + 2(\mathcal{L}\tau)^2\sum_{\ell=1}^{\tau}(\tau-\ell+1)^2\|\boldsymbol{g}_\ell\|^2 \\ &= 2\|\nabla f(\boldsymbol{Z}_0)\|^2 + 4(\mathcal{L}\tau)^2\sum_{\ell=1}^{\tau}(\tau-\ell+1)^2\|\boldsymbol{Z}_\ell - \boldsymbol{Z}_{\ell-1}\|^2 \\ &\leq 2\|\nabla f(\boldsymbol{Z}_0)\|^2 + 4\mathcal{L}^2\tau^4\sum_{\ell=1}^{\tau}\|\boldsymbol{Z}_\ell - \boldsymbol{Z}_{\ell-1}\|^2. \end{aligned} \tag{5.63}$$

Using $\sqrt{a^2+b^2} \leq a+b$ for the two quantities at the right of (5.63) and then standard norm majoration inequalities, we get:

$$\left(\forall \boldsymbol{Z} \in \mathbb{R}^{(\tau+1)N}\right) \quad \|\nabla L(\boldsymbol{Z})\| \leq \sqrt{2}\|\nabla f(\boldsymbol{Z}_0)\| + 2\mathcal{L}\tau^2\sum_{\ell=1}^{\tau}\|\boldsymbol{Z}_\ell - \boldsymbol{Z}_{\ell-1}\|. \tag{5.64}$$

The application of (5.64) to sequence $(\boldsymbol{Z}^k)_{k\in\mathbb{N}}$ leads to

$$(\forall k \geq \tau) \quad \|\nabla L(\boldsymbol{Z}^k)\| \leq \sqrt{2}\|\nabla f(\boldsymbol{x}_k)\| + 2\mathcal{L}\tau^2\sum_{j=k-\tau+1}^{k}\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\|. \tag{5.65}$$

By Lemma 5.3 and (5.65), we finally deduce that

$$\begin{aligned}(\forall k \geq 2\tau) \quad \|\nabla L(\boldsymbol{Z}^k)\| &\leq \sqrt{2}\mathcal{L}\tau\sum_{j=k-2\tau+1}^{k}\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\| + \sqrt{2}\overline{\nu}\sum_{j=k-\tau}^{k}\|\boldsymbol{x}^{j+1} - \boldsymbol{x}^j\| \\ &\quad + 2\mathcal{L}\tau^2\sum_{j=k-\tau+1}^{k}\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\| \\ &\leq \left(\sqrt{2}L\tau + \sqrt{2}\overline{\nu} + 2L\tau^2\right)\sum_{j=k-2\tau+1}^{k}\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\|, \end{aligned} \tag{5.66}$$

which concludes the proof taking $\rho = \sqrt{2}L\tau + \sqrt{2}\overline{\nu} + 2L\tau^2$. □

The following analysis makes use of recent theoretical results around the KL inequality [9, 27] that we recall hereafter. For every $\zeta > 0$, we denote by Φ_ζ the set of concave functions $\varphi\colon [0,\zeta) \mapsto \mathbb{R}_+$ verifying :

- $\varphi(0) = 0$.
- $\varphi \in C^1((0,\zeta))$ and is continuous in 0.
- $\forall s \in (0,\zeta), \varphi'(s) > 0$.

We are then ready to introduce the so-called KL property. [9, 27]

Definition 5.1. *[KL property] A differentiable function $g : \mathbb{R}^d \to \mathbb{R}$, with $d \geq 1$, satisfies the Kurdyka-Łojasiewicz (KL) property on $E \subset \mathbb{R}^d$ if, for every $\boldsymbol{z} \in E$ and every bounded neighborhood V of $\boldsymbol{z}$, there exist $\zeta > 0$ and $\varphi \in \Phi_\zeta$ such that every $\boldsymbol{x} \in E \cap \{\boldsymbol{x} \text{ s.t. } |g(\boldsymbol{x}) - g(\boldsymbol{z})| < \zeta\}$,*

$$\|\nabla g(\boldsymbol{x})\| \varphi'(|g(\boldsymbol{x}) - g(\boldsymbol{z})|) \geq 1. \tag{5.67}$$

We also recall the following Lemma:

Lemma 5.6. *[Uniform KL property [27, Lemma 6]] Let C a compact set of $\mathbb{R}^d$ and $g : \mathbb{R}^d \to \mathbb{R}$ a differentiable function satisfying KL property on C and constant on the latter. Then, there exist $\epsilon, \zeta > 0$ and $\varphi \in \Phi_\zeta$ such that every $\overline{\boldsymbol{x}} \in C$ and all $\boldsymbol{x} \in \mathbb{R}^d$ satisfying both $d(\boldsymbol{x}, C) < \epsilon$, $0 < g(\boldsymbol{x}) - g(\overline{\boldsymbol{x}}) < \zeta$, we have*

$$\|\nabla g(\boldsymbol{x})\| \varphi'(|g(\boldsymbol{x}) - g(\overline{\boldsymbol{x}})|) \geq 1. \tag{5.68}$$

Proposition 5.4. *Under Assumptions 5.1-5.2-5.3, if L defined in (5.58) fulfills the KL property on $\mathbb{R}^{(\tau+1)N}$ then, considering $g = L$, $C = \mathcal{C}\left((\boldsymbol{Z}^k)_{k\in\mathbb{N}}\right)$ with $L(C) = \{f^\infty\}$, there exists ϵ^L, ζ^L and $\phi^L \in \Phi_{\zeta^L}$ such that L satisfies (5.68).*

Proof. This is a direct consequence of Lemma 5.6. Continuity of L is clear. We still have to verify the compactness of $\mathcal{C}\left((\boldsymbol{Z}^k)_{k\in\mathbb{N}}\right)$ and that L is constant valued on that set. $\mathcal{C}\left((\boldsymbol{Z}^k)_{k\in\mathbb{N}}\right)$ is closed. Moreover, it is straightforward to show that this set is included in the Cartesian product $\left[\mathcal{C}\left((\boldsymbol{x}^k)_{k\in\mathbb{N}}\right)\right]^{\tau+1}$, where $\mathcal{C}\left((\boldsymbol{x}^k)_{k\in\mathbb{N}}\right)$ is compact. $\mathcal{C}\left((\boldsymbol{Z}^k)_{k\in\mathbb{N}}\right)$ is thus bounded and, finally, it is compact.
Let $\boldsymbol{Z} \in \mathcal{C}\left((\boldsymbol{Z}^k)_{k\in\mathbb{N}}\right)$. We have $L(\boldsymbol{Z}^k) = f(\boldsymbol{x}^k) + \nu_k$ for all $k \in \mathbb{N}$. From our proof of Theorem 5.1, it follows that sequence $\left(L(\boldsymbol{Z}^k)\right)_{k\in\mathbb{N}}$ converges to f^∞. Continuity of L finally ensures that $f^\infty = L(\boldsymbol{Z})$. This proves that f is constant valued on $\mathcal{C}\left((\boldsymbol{Z}^k)_{k\in\mathbb{N}}\right)$ (and equals to f^∞). □

5.4.4 Convergence of the iterates

We are now ready to state our second convergence Theorem for BD3MG algorithm, characterizing the convergence of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$.

Theorem 5.2. *Let assume that Assumptions 5.1-5.2-5.3 hold. Assume furthermore that the Lyapunov function L in (5.58) satisfies the KL property on $\mathbb{R}^{(\tau+1)N}$. Then, sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is of finite length, i.e :*

$$\sum_{k=0}^{+\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| < +\infty, \tag{5.69}$$

and converges to a stationary point of f.

Proof. Let us start considering the case when there exists some $k_0 \in \mathbb{N}$ where $L(\boldsymbol{Z}^{k_0}) = f^\infty$. Since $\big(L(\boldsymbol{Z}^k)\big)_{k\in\mathbb{N}}$ is decreasing sequence converging to f^∞ (see proof of Proposition 5.4), it follows that $L(\boldsymbol{Z}^k) = f^\infty$ for all $k \geq k_0$. (5.59) then gives

$$(\forall k \geq k_0) \quad \|\boldsymbol{x}^{k+1} - \boldsymbol{x}_k\|^2 \leq \underline{\nu}^{-1} \Big(L(\boldsymbol{Z}^k) - L(\boldsymbol{Z}^{k+1})\Big) = 0, \tag{5.70}$$

ensuring that $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ has a finite length and $\boldsymbol{x}^k$, $k \geq 0$, is a stationary point of f.

We now suppose that, for all $k \in \mathbb{N}$, $L(\boldsymbol{Z}^{k_0}) \neq f^\infty$. We aim at exhibiting a uniform KL inequality on sequence $\big(L(\boldsymbol{Z}^k)\big)_{k\in\mathbb{N}}$. To do so, let us peruse the quantities $\epsilon^L, \eta^L, \varphi^L$ arising from Proposition 5.4. On the one hand, the decrease of $\big(L(\boldsymbol{Z}^k)\big)_{k\in\mathbb{N}}$ implies that, for all $k \in \mathbb{N}$, $L(\boldsymbol{Z}^k) > f^\infty$. The set $\mathcal{C}\left((\boldsymbol{x}^k)_{k\in\mathbb{N}}\right)$ is non empty (see proof of Proposition 5.3), so is the set $\mathcal{C}\left((\boldsymbol{Z}^k)_{k\in\mathbb{N}}\right)$. Let $\boldsymbol{Z} \in \mathcal{C}\left((\boldsymbol{Z}^k)_{k\in\mathbb{N}}\right)$ an element of such set i.e., a cluster point of $(\boldsymbol{Z}^k)_{k\in\mathbb{N}}$. From Proposition 5.4, $L(\boldsymbol{Z}) = f^\infty$. Hence, $L(\boldsymbol{Z}^k) - L(\boldsymbol{Z}) > 0$ for all $k \in \mathbb{N}$.
On the other hand, $\big(L(\boldsymbol{Z}^k)\big)_{k\in\mathbb{N}}$ converges to $f^\infty = L(\boldsymbol{Z})$. The boundedness of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ also ensures this of $(\boldsymbol{Z}^k)_{k\in\mathbb{N}}$.
We deduce the existence of some $k_1 \geq 2\tau$ such that

$$(\forall k \geq k_1) \quad 0 < L(\boldsymbol{Z}^k) - L(\boldsymbol{Z}) < \eta^L, \quad d\Big(\boldsymbol{Z}^k, \mathcal{C}\Big((\boldsymbol{Z}^k)_{k\in\mathbb{N}}\Big)\Big) < \epsilon^L. \tag{5.71}$$

From Proposition 5.4, the uniform KL property on L holds i.e.,

$$(\forall k \geq k_1) \quad \|\nabla L(\boldsymbol{Z}^k)\| \ \left(\varphi^L\right)' \Big(L(\boldsymbol{Z}^k) - L(\boldsymbol{Z})\Big) \geq 1. \tag{5.72}$$

Moreover, setting $\Delta^k = \varphi^L\left(L(\boldsymbol{Z}^k) - L(\boldsymbol{Z})\right) - \varphi^L\left(L(\boldsymbol{Z}^{k+1}) - L(\boldsymbol{Z})\right)$ for all $k \in \mathbb{N}$, concavity of φ^L and (5.59) ensure that

$$\begin{aligned}(\forall k \geq k_1)\ \Delta^k &\geq \left(\varphi^L\right)' \Big(L(\boldsymbol{Z}^k) - L(\boldsymbol{Z})\Big)\Big(L(\boldsymbol{Z}^k) - L(\boldsymbol{Z}^{k+1})\Big) \\ &\geq \overline{\nu}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 \left(\varphi^L\right)' \Big(L(\boldsymbol{Z}^k) - L(\boldsymbol{Z})\Big).\end{aligned} \tag{5.73}$$

The combination of the latter with (5.72) leads to

$$(\forall k \geq k_1) \quad \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 \leq \overline{\nu}^{-1}\Delta^k \|\nabla L(\boldsymbol{Z}^k)\|. \tag{5.74}$$

By Lemma 5.5, we can upper bound the gradient term in (5.74). This gives

$$(\forall k \geq k_1) \quad \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 \leq \rho\overline{\nu}^{-1}\Delta^k \sum_{j=k-2\tau+1}^{k} \|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\|. \tag{5.75}$$

Passing to the root and using the classical identity $\sqrt{ab} \leq a/c + bc/4$, with $a = \sum_{j=k-2\tau+1}^{k} \|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\|$ for all $k \geq k_1$, $b = \Delta^k$, both positive for all $k \geq k_1$ and some $c > 0$ is generic, leads to

$$(\forall k \geq k_1) \quad \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \leq \frac{\sqrt{\rho\overline{\nu}}^{-1/2}}{c} \sum_{j=k-2\tau+1}^{k} \|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\| + \frac{c\sqrt{\rho\overline{\nu}}^{-1/2}}{4}\Delta^k. \tag{5.76}$$

Since $(\Delta^k)_{k\in\mathbb{N}}$ is summable (as a telescopic sequence), we can apply Lemma 5.4 with some $c > 2\tau\sqrt{\rho\overline{\nu}}^{-1/2}$ so that $2\tau\frac{\sqrt{\rho\overline{\nu}}^{-1/2}}{c} < 1$. This shows that sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ has a finite length.

This finite length property entails that $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is a Cauchy sequence and thus a converging one. The final conclusion directly comes from Proposition 5.1, ensuring that every accumulation point of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is a stationary point of f. □

5.4.5 Discussion

Under the KL condition for the Lyapunov function L defined in (5.58), we were able to demonstrate the convergence of sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ to a stationary point of f. Let us notice that f satisfying KL property does not necessary imply that L does. Still, our assumption on L can be verified in practice for a wide class of functions f. For instance, following the discussion in [76, section 6], if f is semi-algebraic [23, 27], then the required condition on L in Theorem 5.2 is satisfied, with function $\varphi^L = \kappa(.)^{1-\theta}$ for a some $(\kappa, \theta) \in \mathbb{R}_+^* \times (0, 1)$. Such situation will be met in our experimental settings in Section 5.5. Extending Theorem 5.2 to any KL function f would be an interesting avenue for future work but up to our knowledge, it does not seem straightforward.

5.5 Application to 3D image restoration

5.5.1 Problem statement

5.5.1.1 *Observation model.*

We focus on the inverse problem of restoring a vectorized 3D volume $\overline{\boldsymbol{x}}$ of size $N = N_{\mathsf{X}} \times N_{\mathsf{Y}} \times N_{\mathsf{Z}}$ given blurry and noisy observation $\boldsymbol{y} \in \mathbb{R}^N$. We consider a depth-variant 3D blur operator $\boldsymbol{H} \in \mathbb{R}^{N\times N}$ associated to kernels with support size $M = M_{\mathsf{X}} \times M_{\mathsf{Y}} \times M_{\mathsf{Z}}$, and additive i.i.d. Gaussian noise with standard deviation $\sigma > 0$, so that the observed volume is related to $\overline{\boldsymbol{x}}$ through,

$$\boldsymbol{y} = \boldsymbol{H}\overline{\boldsymbol{x}} + \boldsymbol{b}, \tag{5.77}$$

with vector $\boldsymbol{b} \in \mathbb{R}^N$ accounting for the noise. The goal is to solve the inverse problem of estimating $\overline{\boldsymbol{x}}$ given $\boldsymbol{y}$ and $\boldsymbol{H}$. Depth-variant blurs are commonly encountered in 3D microscopy [193, 116, 136, 130], due to optical aberrations. They are particular cases of spatially-variant blurs [48, 170]. The degradation operator $\boldsymbol{H}$ raises specific challenges due to its high computational cost. Several strategies have been investigated in the case of 2D spatially variant blur maps encountered for instance in astronomical imaging [82, 83, 93]. The extension to 3D maps of these methods is however not covered

up to our knowledge. This motivates the use of a distributed optimization approach for solving the inverse problem (5.77).

5.5.1.2 *Objective function*

We adopt a variational strategy, which consists in seeking for an estimate of $\overline{\boldsymbol{x}}$ that minimizes a penalized least squares criterion f. A hybrid regularization term is employed incorporating prior knowledge on the smoothness and the range of the sought solution. The objective function reads:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad f(\boldsymbol{x}) = \sum_{s=1}^{S} f_s(\boldsymbol{L}_s \boldsymbol{x}), \tag{5.78}$$

where, for every $s \in \{1, \ldots, S\}$, $\boldsymbol{L}_s \in \mathbb{R}^{P_s \times N}$, $P_s \in \mathbb{N}^*$, and f_s is a function from $\mathbb{R}^{P_s}$ to $\mathbb{R}$. $f_1 \circ \boldsymbol{L}_1$ represents the data fidelity term while the other terms are regularization terms. Here, we set $S = 4$ and

- $P_1 = N$, $\boldsymbol{L}_1 = \boldsymbol{H}$, $f_1 = \frac{1}{2}\| \cdot - \boldsymbol{y}\|^2$,
- $P_2 = N$, $\boldsymbol{L}_2 = \boldsymbol{I}_N$, $f_2 = \eta\, d^2_{[x_{\min}, x_{\max}]^N}$,
- $P_3 = 2N$, $\boldsymbol{L}_3 = [(\boldsymbol{V}^{\mathsf{X}})^\top\ (\boldsymbol{V}^{\mathsf{Y}})^\top]^\top$, $f_3 = \lambda \sum_{n=1}^{N} \sqrt{[.]_n^2 + [.]_{N+n}^2 + \delta^2}$,
- $P_4 = N$, $\boldsymbol{L}_4 = \boldsymbol{V}^{\mathsf{Z}}$, $f_4 = \kappa\| \cdot \|^2$.

Hereabove, $(\eta, \lambda, \delta, \kappa) \in (0, +\infty)^4$ are hyper-parameters. The linear operators $\boldsymbol{V}^{\mathsf{X}}, \boldsymbol{V}^{\mathsf{Y}}, \boldsymbol{V}^{\mathsf{Z}} \in \mathbb{R}^{N \times N}$ are discrete gradient operators along X (horizontal), Y (vertical), and Z (longitudinal) directions of the 3D volume. Function $d^2_{[x_{\min}, x_{\max}]^N}$ states for the squared distance to set $[x_{\min}, x_{\max}]^N \subset \mathbb{R}^N$, with $(x_{\min}, x_{\max}) \in \mathbb{R}^2$ minimal and maximal bounds on the sought intensity values. The later term can be viewed as an exterior penalty function [60]. [1].

5.5.1.3 *majorization mapping.*

In order to implement BD3MG, we must build a majorization mapping ensuring the majorization condition (5.3). First, let us notice that, according to (5.78), the gradient of f reads

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad \nabla f(\boldsymbol{x}) = \sum_{s=1}^{S} \boldsymbol{L}_s^\top \varphi_s(\boldsymbol{L}_s \boldsymbol{x}), \tag{5.79}$$

with, for every $s \in \{1, \ldots, S\}$, $\varphi_s : \mathbb{R}^{P_s} \to \mathbb{R}^{P_s}$ the gradient operator of f_s. Then, function f fits within the class of half-quadratic majorizing constructions initially introduced in [108, 109] and later analysed in [176, 5, 58]. A general structure for the majorization mapping of (5.78) is thus

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad \boldsymbol{A}(\boldsymbol{x}) = \sum_{s=1}^{S} \boldsymbol{L}_s^\top \underset{1 \le p \le P_s}{\mathrm{Diag}} \left\{ [\omega_s(\boldsymbol{L}_s \boldsymbol{x})]_p \right\} \boldsymbol{L}_s, \tag{5.80}$$

[1] Function $\boldsymbol{x} \in \mathbb{R}^N \mapsto d_E^2(\boldsymbol{x})$ is 2-Lipschitz differentiable a soon as E is non-empty closed and convex set [14]. Denoting by p_E the orthogonal projection operator, its gradient then corresponds to $\boldsymbol{x} \in \mathbb{R}^N \mapsto 2(\boldsymbol{x} - P_E(\boldsymbol{x}))$.

where, for every $s \in \{1, \dots, S\}$, $\omega_s : \mathbb{R}^{P_s} \to]0, +\infty[^{P_s}$ is a majorizing potential that depends on the properties of $(f_s)_{1 \leq s \leq S}$ [63, Tab. I]. In our case, for every $s \in \{1, \dots, 4\}$, each of these terms is f_s is β_s-Lipschitz differentiable with

$$\begin{cases} \beta_1 = 1, \\ \beta_2 = 4\eta, \\ \beta_3 = \lambda \delta^{-1}, \\ \beta_4 = 2\kappa. \end{cases} \tag{5.81}$$

Then, from descent lemma [19], a valid choice is $\omega_s(\cdot) = \alpha \beta_s \mathbf{1}_{P_s}$ with some $\alpha \geq 1$ [58]. We adopt this simple strategy for functions f_1, f_2 and f_4, which yields

$$\begin{cases} \omega_1(\cdot) = \alpha \mathbf{1}_N \\ \omega_2(\cdot) = 4\alpha\eta \mathbf{1}_N \\ \omega_4(\cdot) = 2\alpha\kappa \mathbf{1}_N. \end{cases} \tag{5.82}$$

Regarding function f_3, a more sophisticated majorization is adopted, inherited from half-quadratic strategies [5, 176]:

$$(\forall \boldsymbol{v} \in \mathbb{R}^{2N}) \quad \omega_3(\boldsymbol{v}) = \lambda \begin{bmatrix} \left(1/\sqrt{v_n^2 + v_{N+n}^2 + \delta^2}\right)_{1 \leq n \leq N} \\ \left(1/\sqrt{v_n^2 + v_{N+n}^2 + \delta^2}\right)_{1 \leq n \leq N} \end{bmatrix}. \tag{5.83}$$

A quadratic majorization function satisfying (5.3) for a given block $\mathcal{S} \in \mathbb{T}$ can then be obtained using (5.2) with (5.79) and (5.80).

5.5.1.4 *Distributed implementation*

We implement BD3MG algorithm as presented in Section 5.2.4. Our code is available at [50]. We split the 3D volume into 2D slices along the depth axis $z \in \{1, \dots, N_{\mathsf{Z}}\}$, and consider each 2D slice as an individual block upon which workers can compute an update. Assuming a lexicographic ordering of the voxels, this means that the following partition is adopted:

$$\mathbb{T} = \{ \, [\![(i-1)N_{\mathsf{X}}N_{\mathsf{Y}} + 1, iN_{\mathsf{X}}N_{\mathsf{Y}}]\!] \mid 1 \leq i \leq N_{\mathsf{Z}} \, \}. \tag{5.84}$$

BD3MG is implemented on a star graph of workers with a specific master aggregating the current solution. For a given number of active cores $C_{\text{tot}} = C + 1$ of the computer (or of the cluster), one is used as the master process to manage the computation split between the workers while all the $C (= C_{\text{tot}} - 1)$ others, are computing updates asynchronously on planar blocks (i.e., Algorithm 5.2). We initially set, for every $c \in \{1, \dots, C\}$, $\mathcal{S}_c^0$ corresponding to the index set of the $((c-1)\lfloor \frac{N_{\mathsf{Z}}}{C} \rfloor + 1)$-th 2D slice in the volume. Then, at each iteration k, the master requires worker c^k to process the 2D slice with index set $\mathcal{S}_{c^k}^{k+1}$, by applying a first-in, first-out basis. The worker c^k hence computes the update for the 2D slice that has been modified the longest time ago, assuming it is available (i.e., not processed in the same time by another worker). A cyclic block update is used as default choice, if several blocks are available (this typically arises in the first iterations). Furthermore, the master controls that each slice has been updated at least once every τ iterations. Regarding data exchange, as emphasized in

Section 5.2.5, in practice, it is not necessary to share the full vector $\boldsymbol{x}$ with all the workers. Consider a worker update associated to the block $\mathcal{S} \in \mathbb{T}$. The worker has to compute $\nabla_{(\mathcal{S})} f(\boldsymbol{x})$ and $\boldsymbol{A}_{(\mathcal{S})}(\boldsymbol{x})$. Because of the structure of (5.78), these quantities actually only depend on a subpart of vector $\boldsymbol{x}$, defined by $(x_n)_{n \in \mathbb{V}_{\mathcal{S}}}$, with $\mathbb{V}_{\mathcal{S}} \subset [\![1, N]\!]$ a set which has low cardinality compared to the full volume size N. Let us explicit this set for our practical example. The key ingredients to account for are (i) the presence of null entries in the linear operators $(\boldsymbol{L}_s)_{1 \leq s \leq S}$, (ii) the (almost) separability of operators $(\varphi_s, \omega_s)_{1 \leq s \leq S}$. We introduce the following sets, for every $s \in \{1, \dots, S\}$,

$$(\forall n \in \{1, \dots, N\}) \quad \mathrm{col}_{n,s} = \{p \in \{1, \dots, P_s\} \quad \text{s.t.} \quad (\boldsymbol{L}_s)_{p,n} \neq 0\}, \tag{5.85}$$

$$(\forall p \in \{1, \dots, P_s\}) \quad \mathrm{row}_{p,s} = \{n \in \{1, \dots, N\} \quad \text{s.t.} \quad (\boldsymbol{L}_s)_{p,n} \neq 0\}, \tag{5.86}$$

Moreover the separable structures of $(\varphi_s, \omega_s)_{1 \leq s \leq S}$ ensure that for every $s \in \{1, \dots, S\}$ and $p \in \{1, \dots, P_s\}$, there exists a subset $\mathcal{V}_{s,p} \subset [\![1, P_s]\!]$ as well as two functions $\widetilde{\varphi}_{s,p} : \mathbb{R}^{|\mathcal{V}_{s,p}|} \to \mathbb{R}$ and $\widetilde{\omega}_{s,p} : \mathbb{R}^{|\mathcal{V}_{s,p}|} \to (0, +\infty)$ such that

$$(\forall \boldsymbol{v} \in \mathbb{R}^{P_s}) \quad \begin{cases} [\varphi_s(\boldsymbol{v})]_p = \widetilde{\varphi}_{s,p}(\boldsymbol{v}_{(\mathcal{V}_{s,p})}), \\ [\omega_s(\boldsymbol{v})]_p = \widetilde{\omega}_{s,p}(\boldsymbol{v}_{(\mathcal{V}_{s,p})}). \end{cases} \tag{5.87}$$

Considering this, we can now rewrite the expressions $\nabla_{(\mathcal{S})} f(\boldsymbol{x})$ and $\boldsymbol{A}_{(\mathcal{S})}(\boldsymbol{x})$ as

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad \nabla_{(\mathcal{S})} f(\boldsymbol{x}) = ([\nabla f(\boldsymbol{x})]_i)_{i \in \mathcal{S}}, \tag{5.88}$$

with, for every $i \in \mathcal{S}$,

$$[\nabla f(\boldsymbol{x})]_i = \sum_{s=1}^{S} \left[\boldsymbol{L}_s^\top \varphi_s(\boldsymbol{L}_s \boldsymbol{x}) \right]_i, \tag{5.89}$$

$$= \sum_{s=1}^{S} \sum_{p=1}^{P_s} (\boldsymbol{L}_s)_{p,i} \left[\varphi_s(\boldsymbol{L}_s \boldsymbol{x}) \right]_p, \tag{5.90}$$

$$= \sum_{s=1}^{S} \sum_{p \in \mathrm{col}_{i,s}} (\boldsymbol{L}_s)_{p,i} \left[\varphi_s(\boldsymbol{L}_s \boldsymbol{x}) \right]_p, \tag{5.91}$$

$$= \sum_{s=1}^{S} \sum_{p \in \mathrm{col}_{i,s}} (\boldsymbol{L}_s)_{p,i} \widetilde{\varphi}_{s,p}([\boldsymbol{L}_s \boldsymbol{x}]_{(\mathcal{B}_{s,p})}), \tag{5.92}$$

$$= \sum_{s=1}^{S} \sum_{p \in \mathrm{col}_{i,s}} (\boldsymbol{L}_s)_{p,i} \widetilde{\varphi}_{s,p} \left(\left[\sum_{n=1}^{N} (\boldsymbol{L}_s)_{\ell,n} x_n \right]_{\ell \in \mathcal{V}_{s,p}} \right), \tag{5.93}$$

$$= \sum_{s=1}^{S} \sum_{p \in \mathrm{col}_{i,s}} (\boldsymbol{L}_s)_{p,i} \widetilde{\varphi}_{s,p} \left(\left[\sum_{n \in \mathrm{row}_{\ell,s}} (\boldsymbol{L}_s)_{\ell,n} x_n \right]_{\ell \in \mathcal{V}_{s,p}} \right), \tag{5.94}$$

Similar computation shows that, for every $(i, j) \in \mathcal{S}^2$,

$$[\boldsymbol{A}(\boldsymbol{x})]_{i,j} = \sum_{s=1}^{S} \sum_{p \in \mathrm{col}_{i,s} \cap \mathrm{col}_{j,s}} (\boldsymbol{L}_s)_{p,i} (\boldsymbol{L}_s)_{p,j} \widetilde{\omega}_{s,p} \left(\left[\sum_{n \in \mathrm{row}_{\ell,s}} (\boldsymbol{L}_s)_{\ell,n} x_n \right]_{\ell \in \mathcal{V}_{s,p}} \right). \tag{5.95}$$

Hence, (5.94)-(5.95) reflect the fact that the only coordinates of the vector $\boldsymbol{x}$ that are manipulated to compute the gradient and majorization mapping related to block $\mathcal{S}$, belong to $\mathbb{V}_{\mathcal{S}}$ where

$$\mathbb{V}_{\mathcal{S}} = \bigcup_{i\in\mathcal{S}} \bigcup_{s\in\{1,\dots,S\}} \bigcup_{p\in \mathrm{col}_{i,s}} \bigcup_{\ell\in\mathcal{V}_{s,p}} \mathrm{row}_{\ell,s}. \tag{5.96}$$

Since matrices $(\boldsymbol{L}_s)_{1\leq s\leq S}$ are very sparse and functions $(\varphi_s, \omega_s)_{1\leq s\leq S}$ close to separable ones, the cardinality of the involved sets in (5.96) is small so that the memory load for each communication in between master and worker is limited.

5.5.1.5 *Validity of Assumptions.*

Let us discuss the validity of Assumptions 5.1, 5.2 and 5.3 for the considered problem and implementation.

5.5.1.5.1 Assumption 5.1. Function f in (5.78) is differentiable. Moreover, it has a $\mathcal{L}$-Lipschitzian gradient with $\mathcal{L} = \sum_{s=1}^{S} \beta_s |||\boldsymbol{L}_s|||^2$, where $|||\cdot|||$ denotes the spectral norm over matrices and $(\beta_s)_{1\leq s\leq S}$ were given in the previous subsection. According to [203, Prop. 2.5], a sufficient condition for f to be coercive is $\ker(\boldsymbol{H}) = \{\boldsymbol{0}_N\}$. This latter is verified in our experiments, since $\boldsymbol{H}$ is a full-rank operator. Thus, Assumption 5.1 holds.

5.5.1.5.2 Assumption 5.2. This assumption relates to the practical implementation of BD3MG and requires every subset of variables to be updated within a finite number of iterations. In practice, we introduced a safety check in the master loop, that introduces an idle time if a slice has not been updated within the last τ iterations with τ a predefined value. In our implementation, each worker is in average in charge of $\frac{N_Z}{C}$ 2D slices, of the volume. We thus set $\tau = 2\left\lceil \frac{N_Z}{C} \right\rceil$, that is each worker is allowed to spend, in average twice more time to update one slice than another. Given our block selection rule, with balanced computational load per slide, and relying on first-in, first-out, this situation could only arise if a worker experienced a major delay, which never occurred in our experiments.

5.5.1.5.3 Assumption 5.3. This assumption relates to the majorization mapping. To check this assumption, we proceed in three steps. On the one hand, we have,

$$(\forall \boldsymbol{x}\in\mathbb{R}^N) \quad \boldsymbol{A}(\boldsymbol{x}) \succeq \boldsymbol{L}_2^\top \operatorname*{Diag}_{1\leq p\leq P_2} \{\omega_2(\boldsymbol{L}_2\boldsymbol{x})\}\, \boldsymbol{L}_2 \succeq \alpha\eta \boldsymbol{I}_N. \tag{5.97}$$

On the other hand, according to definition (5.80) and those of $\omega_1, \dots, \omega_4$

$$(\forall \boldsymbol{x}\in\mathbb{R}^N) \quad \boldsymbol{A}(\boldsymbol{x}) \preceq \left(\sum_{s=1}^{S} |||\boldsymbol{L}_s|||^2 \max_{1\leq p\leq P_s} [\omega_s(\boldsymbol{L}_s\boldsymbol{x})]_p \right) \boldsymbol{I}_N \preceq \overline{\nu}\boldsymbol{I}_N, \tag{5.98}$$

with

$$\overline{\nu} = \alpha \left(\sum_{s=1}^{S} \beta_s |||\boldsymbol{L}_s|||^2 \right). \tag{5.99}$$

Considering (5.97), (5.98)-(5.99) and the fact that any sub-matrix $\mathcal{M}_{(\mathcal{S})}$ ($\mathcal{S} \subset [\![1, N]\!]$) of a (symmetric) positive matrix $\boldsymbol{M}$ remains positive, the chosen mapping $\boldsymbol{A}$ thus respects conditions imposed by Assumption 5.3(i). Moreover, for all $(\boldsymbol{x}, \boldsymbol{y}) \in (\mathbb{R}^N)^2$,

$$
\begin{aligned}
&\boldsymbol{A}(\boldsymbol{x}) - \frac{1}{2}\boldsymbol{A}(\boldsymbol{y}) \\
&= \sum_{s=1}^{S} (\boldsymbol{L}_s)^\top \underset{1\leq p\leq P_s}{\mathrm{Diag}} \left\{ \left([\omega_s(\boldsymbol{L}_s\boldsymbol{x})]_p - \frac{1}{2}[\omega_s(\boldsymbol{L}_s\boldsymbol{y})]_p \right) \right\} \boldsymbol{L}_s \qquad (5.100) \\
&= \frac{\alpha}{2} \sum_{s\in\{1,2,4\}} (\boldsymbol{L}_s)^\top \boldsymbol{L}_s + (\boldsymbol{L}_3)^\top \underset{1\leq p\leq P_3}{\mathrm{Diag}} \left\{ \left([\omega_3(\boldsymbol{L}_3\boldsymbol{x})]_p - \frac{1}{2}[\omega_3(\boldsymbol{L}_3\boldsymbol{y})]_p \right) \right\} \boldsymbol{L}_3 \\
&\succeq \frac{\alpha}{2}\boldsymbol{L}_2^\top \boldsymbol{L}_2 + (\boldsymbol{L}_3)^\top \underset{1\leq p\leq P_3}{\mathrm{Diag}} \left\{ \left([\omega_3(\boldsymbol{L}_3\boldsymbol{x})]_p - \frac{1}{2}[\omega_3(\boldsymbol{L}_3\boldsymbol{y})]_p \right) \right\} \boldsymbol{L}_3 \\
&\succeq \eta(\alpha)\boldsymbol{I}_N \text{ with } \eta(\alpha) = \frac{\alpha}{2} - \frac{8\lambda}{2\delta}, \qquad (5.101)
\end{aligned}
$$

as $|||\boldsymbol{L}_3|||^2 = 8$. Under the same previous remark on the block positivity preservation, Assumption 5.3(ii) is verified considering α large enough (i.e so as for $\eta(\alpha)$ to strictly exceed bound $\frac{\mathcal{L}\sqrt{\tau}(1+\tau)}{2}$). In practice, we opt for $\alpha = 1.1 \times \left(\mathcal{L}\sqrt{\tau}(1+\tau) + \frac{8\lambda}{\delta}\right)$. The associated $\underline{\nu}$ in (5.27) is $\underline{\nu} = \eta(\alpha) - \frac{\mathcal{L}\sqrt{\tau}(1+\tau)}{2}$).

5.5.1.5.4 Convergence result In a nutshell, Assumptions 5.1-5.2-5.3 are fulfilled in our experiments, so that Theorem 5.1 holds. Moreover, function f is semi-algebraic, hence so is the Lyapunov function L (see discussion in Sec. 5.4.5). Thus, Theorem 5.2 holds.

5.5.2 Comparative analysis on a controlled scenario

We first set $\overline{\boldsymbol{x}}$ as the 3D microscopic image `FlyBrain` [194] with size $N = N_{\mathsf{X}} \times N_{\mathsf{Y}} \times N_{\mathsf{Z}} = 256 \times 256 \times 57$. The linear operator $\boldsymbol{H}$ models a 3D depth-varying Gaussian blur. For each depth $z \in \{1, \ldots, N_{\mathsf{Z}}\}$, the blur kernel is characterized by different variance and rotation parameters $(\sigma_{\mathsf{X}}(z), \sigma_{\mathsf{Y}}(z), \sigma_{\mathsf{Z}}(z), \varphi_{\mathsf{Y}}(z), \varphi_{\mathsf{Z}}(z))$, following the model from [236]. In practice, the values of these five parameters are chosen randomly through a uniform distribution over $[0, 3] \times [0, 3] \times [0, 4] \times [0, 2\pi] \times [0, 2\pi]$, sampled independently for every z. The support of the blur is then truncated to reach a kernel size of $M = M_{\mathsf{X}} \times M_{\mathsf{Y}} \times M_{\mathsf{Z}} = 5 \times 5 \times 11$, which appears large enough to avoid spurious ringing effects. A zero-mean white Gaussian noise with standard deviation $\sigma = 4 \times 10^{-2}$ is then added to the blurred volume. The regularization parameters $(\lambda, \delta, \kappa, \eta) = (1, 1, 10^{-1}, 10^{-3})$ are chosen empirically so as to maximize the Signal-to-Noise Ratio (SNR) of the restored volume. Moreover, we set $(x_{\min}, x_{\max}) = (0, 1)$, equal to the range of the ground truth image. In order to illustrate the acceleration induced by the proposed BD3MG, we run a comparative analysis between different versions of the optimization scheme, in the spirit of an ablation study. Namely, we propose to compare BD3MG with three methods listed hereafter.

- The 3MG algorithm [58, 59] is considered as the baseline. At each iteration, this algorithm builds the majorization mapping as in Sec. 5.5.1.3 and computes memory gradient updates on the full volume, without any parallelization.

- The Asynchronous Block Gradient Descent (ABGD) algorithm implements the method from [177]. It performs parallel asynchronous gradient descent updates over the slices of the volume. We adopt here the same parallelization settings as for our BD3MG. Updates correspond to the standard gradient descent on the selected planar blocks, using a fixed step-size μ ensuring convergence of the iterative scheme, namely $\mu = 0.99/(1 + \kappa + 2\lambda/\delta + 2\kappa)$.
- The BP3MG algorithm from [43, 56] runs a synchronous version of BD3MG algorithm. The master process carries out the main loop of [43, Alg 4.3]. At each iteration $k \in \mathbb{N}$, it selects C block indices (following a cyclic rule) and sends to each worker $c \in [\![1, C]\!]$ the required data allowing it to update $\mathcal{S}_c^k$, the associated block. Workers process their block in parallel, wait for each other to finish their tasks, combine their respective updates into a unique vector $(\boldsymbol{x}^j)_{j \in \mathcal{S}_1^k \cup \ldots \cup \mathcal{S}_C^k}$ and finally send the latter to the master. The majorization mapping is set as a block diagonal matrix, allowing synchronous parallel updates, as described in [43]. This approach could be interpreted as a special case of BD3MG with a single worker (potentially composed of several subworkers) sending its update (potentially composed of several sub-updates) to the central process $\mathcal{S}^k = \{\mathcal{S}_c^k\}_{c \in C}$. Thanks to the specific structure of the majorization mapping in BP3MG, there is no mismatch in information between central process and workers in this synchronous version, the delay vector i_k always equals k. Nonetheless, the block diagonal form of the majorization mapping of BP3MG is at the price of a lower quality of approximation of the cost function, which might result in slower convergence.

All methods are implemented in *Python* using the built-in *Multiprocessing* library as well as *Numpy* and *Scipy* for both data manipulation and scientific computing. The experiments of this section are conducted on an Intel® Xeon(R) W-2135 CPU with $C_{\text{tot}} = 12$ cores clocked at 3.70GHz. All the versions were initialized with $\boldsymbol{x}^0 = \mathbf{0}_N$ leading to an initial value $f(\boldsymbol{x}^0) = 91292.92$. For every iteration $k \in \mathbb{N}^*$, we monitor the cost function $f(\boldsymbol{x}^k)$, the normalized increment $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|/\|\boldsymbol{x}^k\|$, the signal to noise ratio (SNR, in dB) defined as

$$\text{SNR} = 20 \log_{10} \left(\frac{\|\overline{\boldsymbol{x}}\|}{\|\overline{\boldsymbol{x}} - \boldsymbol{x}^k\|} \right), \tag{5.102}$$

and the reconstruction error $\|\overline{\boldsymbol{x}} - \boldsymbol{x}^k\|$. The evolution of these metrics along time for the tested algorithms is displayed on Figure 5.2. We then set a stopping criterion $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \leq \widetilde{\varepsilon}\|\boldsymbol{x}^k\|$. The obtained solution is denoted as $\boldsymbol{x}^{\mathsf{f}}$. We display in Table 5.1 the metrics for the stopping criterion threshold $\widetilde{\varepsilon} = 10^{-3}$. Table 5.1 and Figure 5.2 show that BD3MG exhibits a faster practical convergence than its competitors. Both BD3MG and ABGD are asynchronous distributed schemes, and the former implements an accelerated version of the gradient descent involved in the latter. The MM metric and the subspace scheme in BD3MG act as catalizers, improving the convergence rate compared to ABGD which relies on a simple steepest descent with fixed stepsize. BP3MG and BD3MG are based on the same inherent optimization scheme 3MG. However, BP3MG uses a simplified block diagonal majorization mapping, and imposes synchronous updates, which might yield idle times. These differences can explain why BD3MG converges faster than BP3MG. Finally, 3MG does not exploit the multicore structure of the computing architecture, and thus shows higher computational time.

Slices of the reconstructed volume are displayed in Figure 5.3, revealing fine details of the image recovered by the restoration procedure.

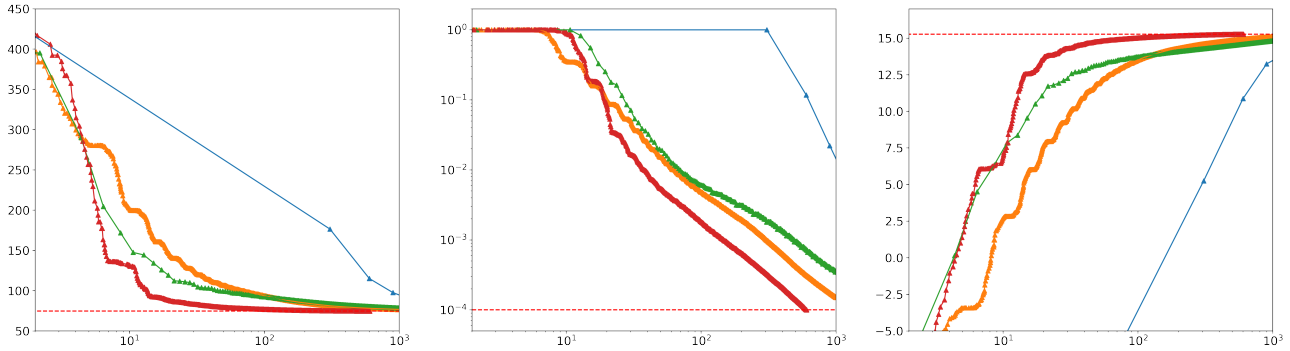


Figure 5.2: Evolution of quantitative metrics along time (in seconds), for algorithms 3MG (blue), ABGD (orange), BP3MG (green) and BD3MG (red), for `FlyBrain` restoration. Evolution of reconstruction error $\|\boldsymbol{x}^k - \overline{\boldsymbol{x}}\|$ (left), relative increment $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|/\|\boldsymbol{x}^k\|$ (middle), and SNR in dB (right).

Version	//	Asy.	MM	SNR (dB)	$f(\boldsymbol{x}^{\mathrm{f}})$	Error	Time ($\times$ Acc.)
3MG	✗	✗	✓	14.72	1266.04	79.28	1683.79 (1)
ABGD	✓	✓	✗	15.11	1268.80	76.73	305.76 (5.51)
BP3MG	✓	✗	✓	15.21	1264.08	75.33	489.99 (3.44)
BD3MG	✓	✓	✓	**15.26**	**1261.59**	**74.72**	**147.16 (11.44)**

Table 5.1: Characteristics and performances of compared algorithms on the `Flybrain` restoration task, for reaching the stopping criterion with $\widetilde{\varepsilon} = 10^{-3}$. "//" = Parallel, "Asy." = Asynchronous, "MM" = Majorize-Minimize scheme. Time is in seconds and "$\times$ Acc." is the acceleration ratio with respect to 3MG running time.

5.5.3 Effect of an imbalanced computing power

In order to further demonstrate the advantages of BD3MG over its synchronous counterpart BP3MG, we tested the methods under different computing environments by synthetically modeling stochastic delays in the computing loop of workers. More specifically, the same restoration task and computer characteristics than in the previous section is considered, again with $C_{\mathrm{tot}} = 12$ active cores. We introduce artificial perturbation in the computing environment by randomly "freezing" some worker processes for a certain amount of time (i.e., delay) following the three scenarios below:

- **Type I**: One of the workers is consistently affected by a delay that follows a uniform distribution $\mathcal{U}([0,1])$ (in sec.). The other cores are not affected by any delay.
- **Type II**: Two worker cores are not affected by any delay while the others 9 agents are delayed in the following fashion:

  3 cores hold a delay following a uniform distribution $\mathcal{U}([0,1])$.

  3 cores hold a delay following a uniform distribution $\mathcal{U}([0,0.5])$.

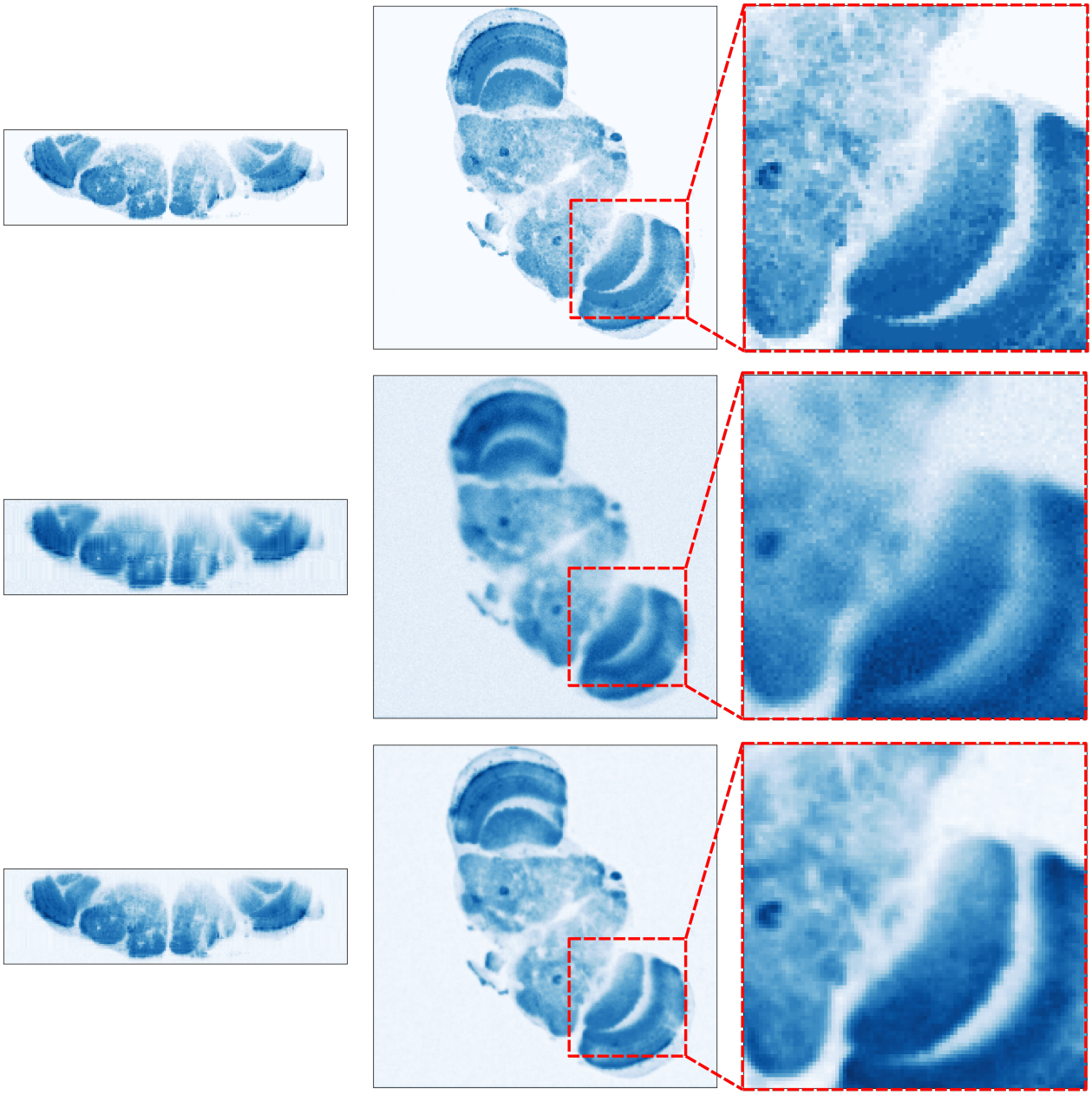


Figure 5.3: Restoration results of `Flybrain`: ground truth volume (top), degraded version (middle), and results of BD3MG restoration (bottom). Visual comparisons along the $\mathsf{X}-\mathsf{Z}$ axis (left) the $\mathsf{X}-\mathsf{Y}$ axis (middle) and zoomed details (right). The optimization process recovers fine details of the original volume that were lost in its degraded version.

3 cores hold a delay following a uniform distribution $\mathcal{U}([0, .25])$.

- **Type III**: All worker cores are affected by a delay that follows a uniform distribution $\mathcal{U}([0, 1])$.

Method (Scenario)	SNR (dB)	$f(\boldsymbol{x}^f)$	Time (s.)
3MG (no delay)	18.132	1247.01	1683.79
BP3MG (Type I)	17.941	1247.14	623.07
BD3MG (Type I)	18.679	1246.04	211.34
BP3MG (Type II)	17.941	1247.14	707.92
BD3MG (Type II)	18.681	1246.03	220.65
BP3MG (Type III)	17.941	1247.14	752.83
BD3MG (Type III)	18.670	1246.02	219.90

Table 5.2: Performances of BP3MG and BD3MG under imbalanced computed power, for reaching the stopping criterion with $\widetilde{\varepsilon} = 5 \times 10^{-4}$ for `Flybrain` restoration. We additionally provide results for the vanilla 3MG algorithm for sake of comparison.

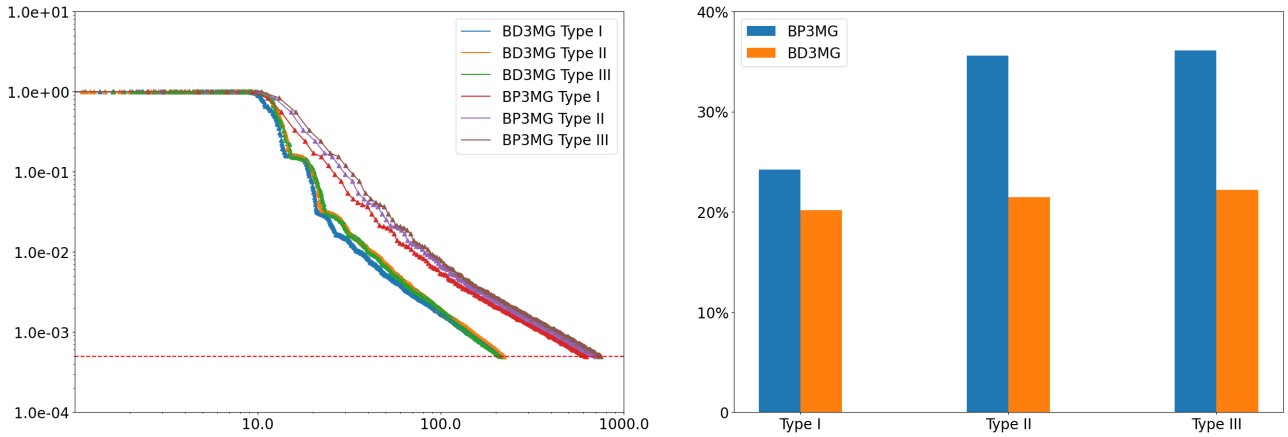


Figure 5.4: Numerical comparisons between BD3MG and BP3MG for `FlyBrain` restoration under imbalanced computing power: evolution of the relative increment $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|/\|\boldsymbol{x}^k\|$ along time (in sec.) for each of the three experimental settings in log-log scale (left), and averaged ratio of workers CPU idle time over the entire optimization process for each scenario (right).

The results are summarized in Table 5.2 and Figure 5.4. We also report the results of a plain, not delayed, 3MG implementation, for the sake of comparison. In all three scenarios, BD3MG outperforms its synchronous version BP3MG, in terms of computation time while reaching similar final criterion value and SNR. The criteria decrease is faster for BD3MG which can be explained by two main differences with BP3MG. First, the majorization mapping of BD3MG performs a tighter approximation of the cost function than BP3MG, thus leading intrinsically to improved convergence rate. Second, BD3MG is asynchronous by essence and thus it is resilient to communication delays as soon as they are bounded, as shown by our convergence analysis. In contrast, BP3MG simply waits for all workers to finalize their update, to force the synchronicity, which yields slowdown in case of delayed workers. A more efficient and dynamic handling of the workload is performed in BD3MG, as shown in Figure 5.4

where CPU idle time is consistently lower for BD3MG than for BP3MG. We note that in asymmetric settings such as (Type II) and (Type III), BD3MG proved to be particularly efficient in reducing the synchronicity constraint of BP3MG for "fast" workers. The comparable results for BD3MG on all three scenarios further suggest that the proposed algorithm is robust to an imbalance in the computing power of workers. Moreover, despite the delayed feedbacks of the workers, both BP3MG and BD3MG remain competitive with respect to the vanilla 3MG, which shows the great interest of a parallel friendly algorithmic structure in this context.

5.5.4 Scalability assessment.

In order to assess the scalability properties of BD3MG, we further analyse the speed-up generated by the number of cores available. We consider the restoration problem of the 3D image `Aneurysm` [140] of size $N = N_{\mathsf{X}} \times N_{\mathsf{Y}} \times N_{\mathsf{Z}} = 256 \times 256 \times 154$, under the same degradation operator and noise level than in the previous example. Figure 5.5 presents the acceleration ratio between the required computation time for a single active worker versus the computation time of up to 30 active workers in reaching the stopping criterion with $\widetilde{\varepsilon} = 10^{-3}$. The regularization parameters are set empirically to $(\lambda, \delta, \kappa, \eta) = (1, 1, 10^{-1}, 10^{-3})$ to maximize the final SNR and the same blur kernel than in the previous subsection is used. The computations were performed using HPC resources from the **Oscar** - Ocean State Center for Advanced Resources of the Center for Computation and Visualization, black University. The hardware is an Intel Corei9 CPU with up to 48 physical cores at 3.3 GHz GHz and 300G of RAM. Results found in Figure 5.5 illustrate the great potential of scalability of the proposed algorithm. As the number of core increases, a mild saturation effect is observed (in agreement with Amdahl's law [183]).

5.5.5 Application to real data from multiphoton microscopy

We finally illustrate the performance of BD3MG on a restoration task of real multiphoton microscopy data specifically acquired for this experiment. Multiphoton microscopy is an interesting solution for the 3D and submicrometric characterization of biomedical structures, it is label-free and contactless [113]. Such a solution takes advantage of optical sectioning, an optical property resulting from the nonlinear optical processes involved. 3D images are produced with sub-micrometer resolution without slicing the sample. We use an instrumental acquisition pipeline relying on a commercial system from Olympus (BX61WI) coupled with a multiphoton water immersion objective (Olympus XLPLN25XWMP, 25×, NA 1.05). A laser system, emitting femtosecond pulses centred at 810 nm with 10 nm of spectral bandwidth, is used for production of the nonlinear phenomena of second harmonic generation (SHG) and two-photon fluorescence (TPF). The biomedical sample is made of a whole mouse muscle, the Extensor digitorum longus (EDL), isolated from tendon to tendon. Sub-micrometric fluorescent microspheres emitting in the green range are included into the EDL and spread homogeneously all along the whole muscle structure. Under such an experimental protocol, the production of two 3D images is obtained. The first channel contains the SHG from the myosin of the muscle and the second channel displayed the TPF of microspheres used for calibrating the instrumental PSF. A hundred of 2D image slices of SHG and TPF are produced, with 0.1 μm resolution along depth axis Z and $0.049\mu\text{m} \times 0.049\mu\text{m}$ resolution over $\mathsf{X} - \mathsf{Y}$ horizontal-vertical axis. The acquisition recording starts 140 μm under the sample surface

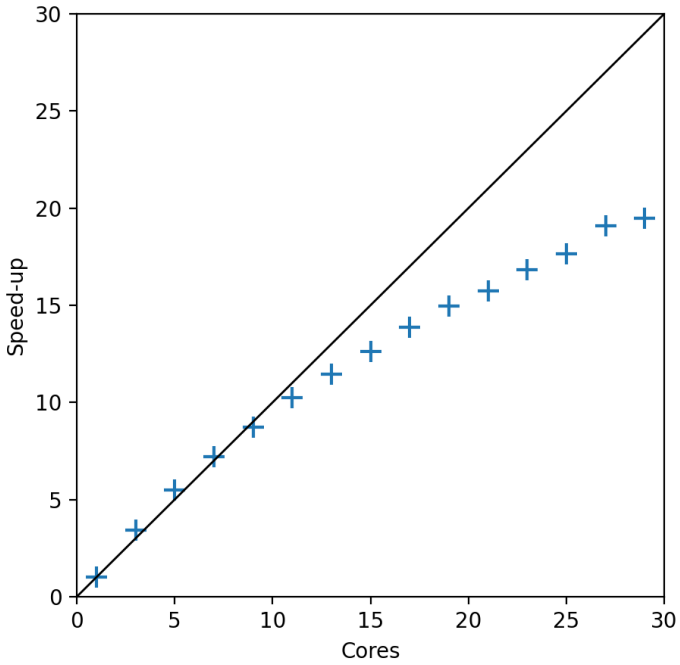


Figure 5.5: Speed-up ratio of the computation time for 1 to 30 cores for BD3MG for the restoration of `Aneurysm`.

for a total sample thickness of 180 µm. For this range of depth, the imaging of biological samples is degraded by scattering effects. Both raw volumes (i.e., SHG and TPF) dimension have $2048 \times 2048 \times 100$ voxels, from which we extract a subpart with size $N = N_{\mathsf{X}} \times N_{\mathsf{Y}} \times N_{\mathsf{Z}} = 256 \times 256 \times 100$ voxels for the purpose of our study.

We follow the computational pipeline FAMOUS previously introduced in [147]. We estimate a depth-variant Gaussian PSF field within the 3D microscopic volume by applying the 3D Gaussian fitting algorithm FIGARO from [67] to volume of interests extracted from the second image channel, displaying fluorescence of calibrated microbeads. Each volume of interest is selected through an automatic search of connected components within a filtered and binarized version of the observed volume. Then, FIGARO method is ran, yielding parameters (i.e., mean, covariance, scaling, shift) of a 3D Gaussian shape. This allows to build, through a simple interpolation strategy, a model for a depth-variant PSF with truncated support of size $M = M_{\mathsf{X}} \times M_{\mathsf{Y}} \times M_{\mathsf{Z}} = 21 \times 21 \times 21$ (see more details in [147, Sec.2.4]). Since no ground truth is available, the regularization parameters $(\lambda, \delta, \kappa, \eta) = (10^2, 2, 10, 10^{-3})$, are selected by retrospective visual inspection. The reconstruction shown in Figure 5.6 exhibits clear contrasts and sharpness properties. Comparative videos of the original and restored volume are available at [50]. The native signal from the raw image was presenting a high level of noise and blur due to the presence of scattering elements all along the 140 µm of sample depth. Thanks to the proposed restoration strategy, the localisation of the myosin in the muscle sample is made possible, and the spatial organization of this protein into the down side of the EDL is revealed. The volume restoration

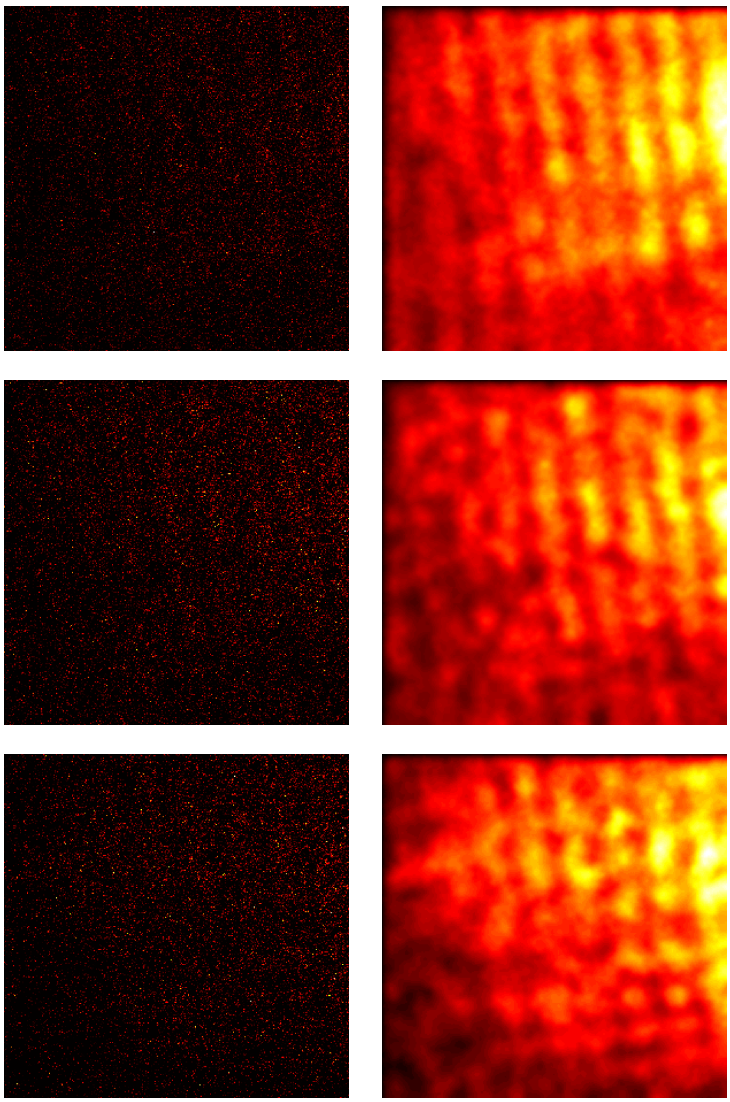


Figure 5.6: Slices $(12,5\mu m \times 12,5\mu m)$ for depths $z = 5, 25$ and 70 (from top to bottom) of the original acquisition (left) and after restoration (right). The comparisons show that the definition of the muscular structure has been enhanced by the reconstruction.

took 305 seconds and ~ 2000 iterations on a $C_{\text{tot}} = 12$ cores setting, when setting $\widetilde{\varepsilon} = 10^{-3}$.

5.6 Conclusion

In this chapter, we have presented a new block distributed Majorize-Minimize algorithm, BD3MG, devised to tackle large-size differentiable optimization problems met in a wide range of applications. Our main contribution lies in a distributed asynchronous formulation that allows for delays in the current

solution computed between workers, while securing convergence guarantees under mild assumptions. Our new algorithm BD3MG has been tested in the context of 3D image restoration with depth-variant blur. Experimental results underlined the speedup potential of this method and its concrete applicability in the field of fluorescence microscopy. Future work will be dedicated to extension to more general distributed graph topologies.

Chapter 6

Introduction to stochastic differentiable optimization

Contents

6.1 Introduction

The complexity of most natural or even artificial phenomena requires infinitely precise knowledge to be perfectly controlled and reproducible. On a human scale and whatever the tools used, the access to such sources of information is impossible and it is therefore necessary to admit the existence of so-call random events which escape any form of prediction. In a way, the introduction of the notion of chance is based on a failure: human beings have neither the time nor the means to deal with all the problems they face using a deterministic approach.

On a mathematical viewpoint and similarly with a black box model, the random behaviours interfering with the problem of interest are grouped behind the veil of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In more concrete terms, σ-algebra $\mathcal{F}$ is associated to all the necessary information we shall consider for the study to be conducted while probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ acts as a quantification tool on the latter. An event, i.e. an element of $\mathcal{F}$, corresponds more precisely to a grouping of elementary productions $\omega \in \Omega$. The field of Euclidean probabilistic unconstrained optimization, more commonly known as Euclidean stochastic (unconstrained) optimization, is thus dedicated to find a minimizer of a cost function $F : \mathcal{H} \to \mathbb{R}$ defined on a real-finite dimensional Hilbert space $\mathcal{H}$ by using an algorithm whose outputs are likely to change even though the operating conditions set by the user (initial point, stepsize...) are identical. More specifically, we aim to solve the generic problem

$$\text{Find } \boldsymbol{x}_s \in \mathcal{H} \text{ s.t. } F(\boldsymbol{x}_s) \leq F(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathcal{H}, \tag{6.1}$$

using an approximation sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$. In contrast to the deterministic case, the terms of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ are no longer vectors of $\mathcal{H}$ but random variables $\mathbf{x}_k : \omega \in \Omega \mapsto \mathbf{x}_k(\omega) \in \mathcal{H}$ $(k \in \mathbb{N})$ defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in the Borel space $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$. A sequence of random variables as $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is generally found under the name of *stochastic process* in the literature. Instead of focusing on a simple sequence of $\mathcal{H}^{\mathbb{N}}$ as we previously made, the mathematical objects we study here thus group together the set of so-call trajectories, i.e. of elementary productions $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ $(\omega \in \Omega)$. Historically, the theory of stochastic process naturally follows the one from Kolmogorov on probabilities [134, 86] and its application to optimization initially comes from the more general field of stochastic approximation whose origins trace back to the works of H. Robbins and S. Monro [201].

Example 6.1. *Let* $\mathbf{v}$ *be a random variable defined on* $(\Omega, \mathcal{F}, \mathbb{P})$ *with values in a measurable space* $(E, \mathcal{A})$*,* $j : \mathcal{H} \times E \to \mathbb{R}$ *be a measurable application for which* $j(\boldsymbol{x}, \mathbf{v})$ *is integrable for every* $\boldsymbol{x} \in \mathcal{H}$ *and consider the minimization of function* $F : \boldsymbol{x} \mapsto \mathbb{E}(j(\boldsymbol{x}, \mathbf{v}))$*.*

In the case where j *is differentiable regarding its first variable, denoting by* $\nabla j^{(1)}$ *its partial gradient and assuming that the conditions for applying Leibniz integral rule derivative are met [214],* F *is differentiable and its gradient can be written as* $\nabla F : \boldsymbol{x} \in \mathcal{H} \to \mathbb{E}\left(\nabla j^{(1)}(\boldsymbol{x}, \mathbf{v})\right)$*. When the distribution of* $\mathbf{v}$ *is unknown, the curvature information on* F *therefore cannot be known in an exact manner and shall require the use of stochastic framework. The construction of an attached stochastic process in a way to approximate a solution of* (6.1) *can be made using multiple strategies, most of which will be described throughout this chapter.*

The presentation we propose follows a similar structure to those of Chapter 2 to facilitate the reader's understanding and to better highlight the existing analogies and differences between probabilistic and deterministic frameworks. To such an extent, we introduce in section 6.2 the most usual

family of algorithms to deal with stochastic differentiable optimization, the stochastic gradient one. Section 6.3 proposes a review of the classical mathematical tools to conduct asymptotic analysis considering a generic stochastic algorithm. Section 6.4 gives the conclusion of this chapter.

6.2 The class of stochastic gradient methods

Similarly to the deterministic differentiable optimization field with the general class of descent methods, those of probabilistic differentiable optimization also possesses its family of algorithms widely described in the literature. The latters are grouped under the name of stochastic gradient methods. By contrast with formulation (2.3) which is able to encompass the largest part of deterministic existing methods, it is quite challenging to find a unique analog formulation for the stochastic setting. Most of those proposed are mainly based on the even more general field of the stochastic approximation [201, 89]. For presentation purposes, we consider in this section processes $(\mathbf{x}_k)_{k\in\mathbb{N}}$ generated through a scheme of the form of

$$\begin{aligned} &\mathbf{x}_0 = \boldsymbol{x}_0 \in \mathcal{H}, \\ &(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k, \end{aligned} \tag{6.2}$$

where $(\mathbf{g}_k)_{k\in\mathbb{N}}$ is a process of directions of $\mathcal{H}$ intended to approximate those of the true gradient $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$. In the same way as (2.3), $(\alpha_k)_{k\in\mathbb{N}}$ is a deterministic positive stepsize sequence. The latter is also regularly found under the name of *learning rate* in the machine learning community [169].

6.2.1 Overview on stochastic gradient approximation constructions

6.2.1.1 *Gradient approximation models through sampling strategies*

Beyond the scope of optimization, sampling approximation field is initially dedicated to identify some properties (distribution, moments...) of a given random variable $\mathbf{v}$ of $(\Omega, \mathcal{F}, \mathbb{P})$ from several realizations of the latter. To such an extent, the Law of Large Numbers (LLN) [22] is classically the most fundamental result of this theory:

Theorem 6.1. *(Law of Large Number, strong version) Let $(\mathbf{v}_k)_{k\in\mathbb{N}}$ be an stochastic process of integrable independant and identically distributed random variables of $(\Omega, \mathcal{F}, \mathbb{P})$. Then*

$$\mathbb{P}\left(\left\{\omega \in \Omega \mid \frac{1}{k+1}\sum_{i=0}^{k} \mathbf{v}_k(\omega) \underset{k\to+\infty}{\longrightarrow} \mathbb{E}(\mathbf{v}_0)\right\}\right) = 1. \tag{6.3}$$

LLN stipulates that the "vast" majority of trajectories (we will notably speak about *almost-sure convergence* in the rest of this manuscript) here associated to the empirical mean process converges to a common deterministic quantity, $\mathbb{E}(\mathbf{v}_0)$, namely the expectation associated with the common law of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ family. Applied to an optimization context, Theorem 6.1 is first and foremost an approximation tool to build a relevant process $(\mathbf{g}_k)_{k\in\mathbb{N}}$ in various scenarios depending from the nature of the stochasticity. Two situations are under the scope of the most recent works; the first one corresponds

to those when the cost function F is of probabilistic nature (Example 6.2) while, in the second one, F admits a deterministic closed-from but with a too complex structure to be itself or its gradient exactly computed (Example 6.3).

Example 6.2. *(Expected Risk) We place ourselves in the situation of Example 6.1 where we aim to minimize the differentiable function* $F : \boldsymbol{x} \mapsto \mathbb{E}(j(\boldsymbol{x}, \mathbf{v}))$ *with* $\nabla F : \boldsymbol{x} \mapsto \mathbb{E}(\nabla j^{(1)}(\boldsymbol{x}, \mathbf{v}))$ *as associated gradient. Then, for any* $\boldsymbol{x} \in \mathcal{H}$*, LLN makes natural the use of approximation of the form of*

$$\nabla F(\boldsymbol{x}) \simeq \frac{1}{P} \sum_{p=1}^{P} \nabla j^{(1)}(\boldsymbol{x}, \mathbf{v}_p) \ (P \geq 1). \tag{6.4}$$

As a consequence and if the distribution of $\mathbf{v}$ *is known, one can build the process* $(\mathbf{g}_k)_{k\in\mathbb{N}}$ *as*

$$(\forall k \in \mathbb{N}) \quad \mathbf{g}_k := \frac{1}{P_k} \sum_{p=1}^{P_k} \nabla j^{(1)}(\mathbf{x}_k, \mathbf{v}_{k,p}) \ (P_k \geq 1), \tag{6.5}$$

where $\mathbf{v}_{k,1}, \ldots, \mathbf{v}_{k,P_k}$ *correspond to a sample of* $\mathbf{v}$ *(i.e. a family of independant random variables following the distribution of* $\mathbf{v}$*) of a given size* P_k*. Estimation* (6.4) *is generally referred in the literature as a Stochastic Average Approximation (SAA) whose theoretical aspect are typically detailed in [219].*

Example 6.3. *(Empirical Risk) We here consider* $F : \boldsymbol{x} \mapsto \frac{1}{M} \sum_{i=1}^{M} f_i(\boldsymbol{x})$ *written as a sum of* M *functions* $(f_i)_{i\leq M}$ $(M \geq 1)$*. To the extent the computation of the true gradient* $\nabla F : \boldsymbol{x} \mapsto \frac{1}{M} \sum_{i=1}^{M} f_i(\boldsymbol{x})$ *requires to calculate those of every* f_i *function, it may therefore be highly time-consuming depending on the value of* M*. To overcome such an issue, one strategy consists in adopting a so-call "mini-batch" approach [36] at a given iteration* $k \in \mathbb{N}$*, indices* $i \in \{1, \ldots, M\}$ *to be consider only belong to a randomly chosen subset (i.e. a batch)* $\mathbf{I}_k \subset \{1, \ldots, M\}$*. Considering this,* $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$ *is thus approximated as* $(\mathbf{g}_k)_{k\in\mathbb{N}}$*, where:*

$$(\forall k \in \mathbb{N}) \quad \mathbf{g}_k := \frac{1}{|\mathbf{I}_k|} \sum_{i\in\mathbf{I}_k} \nabla f_i(\mathbf{x}_k), \tag{6.6}$$

the size of $|\mathbf{I}_k|$ *generally remaining small compared with* M*. In practice, the consideration of a unique sample (i.e.* $|\mathbf{I}_k| = 1$ *for all* $k \in \mathbb{N}$*) may promote interesting theoretical guarantees as soon as it is chosen following a distribution (typically a uniform one) which intuitively ensures the data of any* ∇f_i $(1 \leq i \leq N)$ *being regularly taken into account all along the process [18, 212].*

Since sampling approximation only uses partial gradient information, it is expected for the resulting stochastic algorithm to converge slower in term of number of iterations than its deterministic counterpart using directly ∇F instead. Typically, in the strongly convex case and considering a constant stepsize, the convergence speed of steepest descent method remains linear while the use of a mini-batch approach with a unique sample (see example 6.3) only gives a sublinear rate in expectation [35]. In fact, a deterministic algorithm would be more efficient regarding the number of total required iterations by contrast with a stochastic approach possessing a smaller cost-per-iteration ; intuitively it is always more economical to settle for an approximation of an object instead of the whole object itself . As a consequence, the interest of sampling gradient approximation techniques especially lies in a large scale optimization context for which a simple computation of the full gradient may be too energy consuming to be correctly proceeded.

6.2.1.2 *Generic Additive model*

For first-order methods, the information provided by F being concentrated in the algorithm only through the gradient term, a noise model adapted to a general framework, working with a generic structure of cost function, is of the additive type:

$$(\forall k \in \mathbb{N}) \quad \mathbf{g}_k := \nabla F(\mathbf{x}_k) + \epsilon_k. \tag{6.7}$$

Otherwise stated, $(\epsilon_k)_{k\in\mathbb{N}}$ is the process which simply corresponds the approximation error between the estimate $(\mathbf{g}_k)_{k\in\mathbb{N}}$ and the true gradient $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$. Such a probabilistic framework is the original one at the root of stochastic approximation theory [201, 135, 143]. Historically, [201] built this kind of algorithm in a way of finding the unique zero of a regression function while [135] applied this approach to an optimization context, considering the latter regression function as a gradient object whose zeros are sought for. The interest of decomposition (6.7) mainly lies in its simplicity which enables to conduct theoretical reasoning in a relatively easy way. The gradient term is here "deterministic" in the sense that it is entirely dependent on the current state of the process; considering $k \in \mathbb{N}$, all the uncertainty is concentrated in $(\epsilon_k)_{k\in\mathbb{N}}$ and $\nabla F(\mathbf{x}_k)$ simply remains $\sigma(\mathbf{x}_k)$-measurable.

6.2.2 Stepsize choice

Similarly with the deterministic setting, the ways of adjusting the stepsize $(\alpha_k)_{k\in\mathbb{N}}$ are numerous and depend on the information available on the curvature of F (Lipschitz continuity of the gradient, convexity, strongly convexity...). In general, such a setting follows a relatively uniform pattern and $(\alpha_k)_{k\in\mathbb{N}}$ is constructed so as to verify

$$\sum_{k=0}^{+\infty} \alpha_k = +\infty, \quad \sum_{k=0}^{+\infty} \alpha_k^2 < +\infty. \tag{6.8}$$

Rule (6.8) can be justified considering two points of view. The first one is purely physical and makes the link with differential equations field [104]. In such a context, $(\alpha_k)_{k\in\mathbb{N}}$ can be interpreted as a time increments sequence $(T_{k+1} - T_k)_{k\in\mathbb{N}}$ and so (6.8) indicates that the evolution time $(T_k)_{k\in\mathbb{N}}$ tends to infinity and is of finite energy. The second one is directly relative to the structure of (6.2) ; $(\alpha_k)_{k\in\mathbb{N}}$ plays the role of a control term so as to compensate the stochasticity effect of $(\mathbf{g}_k)_{k\in\mathbb{N}}$. (6.8) is thus the result of a trade-off. On the one hand, condition $\sum_{k=0}^{+\infty} \alpha_k^2 < +\infty$ ensures, to a certain extent, the decay of $(\alpha_k)_{k\in\mathbb{N}}$ and its convergence to zero to attenuate the noise fluctuations. On the other hand, $(\alpha_k)_{k\in\mathbb{N}}$ should not be too small otherwise too little evolution of the process would be observed.

Condition $\sum_{k=0}^{+\infty} \alpha_k^2 < +\infty$ does not appear as the most natural way to balance the noise effect contrary with $\alpha_k \underset{k\to\infty}{\longrightarrow} 0$. However, it has revealed to promote easier asymptotical properties in view of the evolution of the theoretical material of stochastic optimization [202, 89]. More recent works have started to only consider $\alpha_k \underset{k\to\infty}{\longrightarrow} 0$ in their assumptions as a relaxed alternative and are able to promote interesting asymptotical guarantees [104]. Finally, when the stepsize is not chosen to decrease to zero but constant equal to $\alpha > 0$ instead, [186, 85] highlighted the fact that the convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to a stationary point remains very challenging due to the Markovian nature of scheme (6.2). Typically in a strongly convex setting, $(\mathbf{x}_k)_{k\in\mathbb{N}}$ tends to oscillate around $\boldsymbol{x}_s$, the unique minimizer of F, according to a certain distribution and with an amplitude of $\sqrt{\alpha}$ as order of magnitude.

6.2.3 *Acceleration techniques*

6.2.3.1 *Variance reduction strategies*

Most of the recent works on stochastic gradient methods have directly based their asymptotical studies on the analysis of quantities $\left(\mathbb{E}[\|\mathbf{x}_k - \boldsymbol{x}_s\|^2]\right)_{k\in\mathbb{N}}$ or $(\mathbb{E}[F(\mathbf{x}_k) - F(\boldsymbol{x}_s)])_{k\in\mathbb{N}}$ considering $\boldsymbol{x}_s$ as a solution of (6.1). In other words, the theoretical performances of an algorithm are mainly judged on the rate of convergence of its attached variance.

To best meet this evaluation criterion, variance reduction-based strategies consist in tuning the parameters of (6.2), even if it requires changing its structure so that $\mathbb{E}\left[\|\mathbf{x}_k - \boldsymbol{x}_s\|^2\right]$ and/or $\mathbb{E}[F(\mathbf{x}_k) - F(\boldsymbol{x}_s)] \leq \varepsilon$ $(k \in \mathbb{N}, \varepsilon > 0)$ are reached with a complexity C_ε as small as possible. This kind of approach relies on two different levels:

- The assumptions on F. In the case where F admits multiple global minimizers, it generates an ambiguity on $\boldsymbol{x}_s$ to consider and thus on the variance to adopt. More generally, if F admits multiple stationary points, the convergence of the variance to zero may be compromised. In order to overcome such issues, variance reduction algorithms generally required to work in a convex or even strongly convex setting [171].

- The choice of stepsize. Obtaining a precision on the variance as small as desired after a finite number of iterations already implies the variance to converge to zero. The strong convexity assumption makes possible to achieve such a result but it also requires for the stepsize to converge to zero not to slowly in which case it would risk completely masking the behaviour of the variance. In general $(\alpha_k)_{k\in\mathbb{N}}$ is taken as proportional to $(n^{-\gamma})_{k\in\mathbb{N}}$ where $\gamma \in [0, 1]$ [168].

In some works the term of variance considered differs from the one we previously introduce. For instance, they involve $(\overline{\mathbf{x}}_k)_{k\in\mathbb{N}} = \left(1/(k+1)\sum_{i=0}^{k}\mathbf{x}_i\right)_{k\in\mathbb{N}}$ the Ruppert-Polyak averaging sequence instead of directly $(\mathbf{x}_k)_{k\in\mathbb{N}}$ for accuracy purposes [189, 12].

Due to the necessity of manipulating high-dimensional data in the field of statistical learning, most of the usual algorithms have been developped to deal with functions F under the form of an empirical risk (see Example 6.3). Their updates generally remain of the form of (6.2) but the corresponding sequence $(\mathbf{g}_k)_{k\in\mathbb{N}}$ is no longer dedicated to directly approximate $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$ at every iteration. Typically, the class of stochastic averaging methods build process $(\mathbf{g}_k)_{k\in\mathbb{N}}$ exploiting part of the information of the current state (i.e. those of $k \in \mathbb{N}$), but also of some previous states (i.e. relative to $k-1, k-2, ...$). While [173, 238] initially used the information of all past gradients $\nabla F(\mathbf{x}_k), \ldots, \nabla F(\mathbf{x}_0)$, most recents works instead focus on updating a single component with a known expectation at each iteration [78].

Method	Criterion ($k \in \mathbb{N}$)	Stepsize	Convergence rate	$C_\varepsilon / \dim(\mathcal{H})$
GD [172]	$\mathbb{E}[F(\boldsymbol{x}_k) - F(\boldsymbol{x}_s)]$	$\frac{2}{L+\mu}$	$\mathcal{O}\left(\exp(-\frac{\mu}{Lk})\right)$	$\frac{LM}{\mu}\log(\frac{1}{\varepsilon})$
SGD	$\mathbb{E}[F(\overline{\mathbf{x}}_k) - F(\boldsymbol{x}_s)]$	$\frac{1}{Lk}$	$\mathcal{O}\left(\frac{L}{\mu k}\right)$	$\frac{L}{\varepsilon\mu}$
SAG [206]	$\mathbb{E}[F(\mathbf{x}_k) - F(\boldsymbol{x}_s)]$	$\frac{1}{16L}$	$\mathcal{O}\left(\left[1 - \min(\frac{1}{8n}, \frac{\mu}{16L}\right]^k\right)$	$\left(\frac{L}{\mu} + M\right)\log(\frac{1}{\varepsilon})$
SVRG [131]	$\mathbb{E}[F(\mathbf{x}_k) - F(\boldsymbol{x}_s)]$	$\alpha > 0$	$\mathcal{O}\left(\left[\frac{1}{\mu\alpha(1-2L\alpha)} + \frac{2L\alpha}{1-2L\alpha}\right]^k\right)$	$\left(\frac{L}{\mu} + M\right)\log(\frac{1}{\varepsilon})$
SAGA [78]	$\mathbb{E}[\|\mathbf{x}_k - \boldsymbol{x}_s\|]$	$\frac{1}{3L}$	$\mathcal{O}\left(\left[1 - \min(\frac{1}{4n}, \frac{\mu}{3L}\right]^k\right)$	$\left(\frac{L}{\mu} + M\right)\log(\frac{1}{\varepsilon})$

Table 6.1: Performances of some usual variance-reduction methods for the minimization of a L-Lipschitz continuous gradient and μ-strongly convex empirical risk $F : \boldsymbol{x} \mapsto \frac{1}{M}\sum_{i=1}^M f_i(\boldsymbol{x})$. Deterministic steepest descent and SGD serving here as baseline algorithms. As introduced previously, $\overline{\mathbf{x}}_k$ ($k \in \mathbb{N}$) denotes the Polyak-Ruppert average while C_ε corresponds to the complexity required to have a criterion smaller than ε ($\varepsilon > 0$).

As a way of illustration, Table 6.1 highlights the interest of variance-reduction techniques in the strongly-convex setting. They allow to reduce the complexity from a linear to a logarithmic scale despite only capturing partial information on the full gradient. In addition to the algorithms mentioned, we can add those of [79, 218] possessing similar convergence rates and complexities but considering a criterion with a little more elaborated structure.

6.2.3.2 *Preconditioned method*

A popular class of gradient-based stochastic algorithms is those incorporating a preconditioning step:

$$\begin{aligned} &\mathbf{x}_0 = \boldsymbol{x}_0 \in \mathcal{H}, \\ &(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{B}_k \mathbf{g}_k. \end{aligned} \tag{6.9}$$

Their update remains similar of those of (6.2) in the sense they involve a gradient approximation process as well as a stepsize but also requires the use of an additionnal process $(\mathbf{B}_k)_{k\in\mathbb{N}}$ of linear operators in a way to better capture information on the curvature of F. The ways for building the preconditioners are numerous and we here try to give an overview of the most popular approaches.

The first category encompasses those considering $(\mathbf{B}_k)_{k\in\mathbb{N}}$ as an approximation of the second order information of F. Their use is motivated by the encouraging results obtained by their deterministic counterparts. When F is twice continuously differentiable, it is natural to build $(\mathbf{B}_k)_{k\in\mathbb{N}}$ in a way to approximate the Hessian process $\left(\nabla^2 F(\mathbf{x}_k)\right)_{k\in\mathbb{N}}$ be it through sub-sampling techniques [41, 36] or by extension of the non-probabilistic framework [42, 213].

To the extent the computation of second order objects generally tends to highly increase the per-iteration cost of an algorithm [36], another strategy consists in restricting $(\mathbf{B}_k)_{k\in\mathbb{N}}$ to a process of diagonal operators which amounts to only re-scale the direction process $(\mathbf{g}_k)_{k\in\mathbb{N}}$. The greatest challenge of re-scaling algorithms relies in the choice of the information to preserve. While some works prefer keeping the true [15] or an estimation [33] of the diagonal curvature of the Hessian, others privilege

techniques putting the algorithm progress on a same equal footing in every direction. The latters have notably become relatively popular among supervised deep learning community for which the linear operators under study are easily near singular [36]. Let us mention ADAGRAD [87], RMSprop [120] and ADAM [137], probably the most widespread approaches in this field. Their re-scaling steps are all based on a normalization of the second moment of $(\mathbf{g}_k)_{k\in\mathbb{N}}$ using its magnitude component by component. ADAM differs from its two counterparts as it also incorporates an additional moment step in the computation of $(\mathbf{g}_k)_{k\in\mathbb{N}}$ in a way of bias reduction.

6.3 Theoretical background to deal with stochastic setting

This section can be considered as the stochastic counterpart of section 2.4 of Chapter 2. As the notion of supermartingale generalizes that of decreasing sequence in the probabilistic framework, it becomes possible to extend that of descent condition (see Chapter 2 subsection 2.3.2) in a similar way. Moreover, since the theory of martingales leads quite naturally to almost-sure convergence results [86], various asymptotic properties based on this mode of convergence follow logically from this and are now regarded as essential elements of the stochastic literature.

6.3.1 On convergence of stochastic schemes in general

Whether it is deterministic or stochastic, the primary objective of a minimization process always remains the same, namely to restore a global minimizer or at least a stationary point $\boldsymbol{x}^*$ of the cost function considered, as precise as possible. In the case where the process at stake is deterministic and is in fact similar to a sequence of vectors of $\mathcal{H}$, global convergence can be seen as a natural performance criterion. Since a stochastic process no longer involves deterministic vectors but random variables, the notion of global convergence becomes inappropriate in a probabilistic setting and shall be redefined to remain consistent with the mathematical objects manipulated. Since a stochastic process is no more than a family of trajectories, the most intuitive way to extent the global convergence definition is to use the general notion of almost-sure convergence instead.

Definition 6.1. *(Almost-sure convergence) A stochastic process $(\mathbf{x}_k)_{k\in\mathbb{N}}$ of random variables $\mathbf{x}_k : \Omega \to \mathcal{H}$ $(k \in \mathbb{N})$ is said to converge almost-surely (a.s.) to a random variable $\mathbf{x}_\infty : \Omega \to \mathcal{H}$ if*

$$\mathbb{P}\left(\left\{\omega \in \Omega \mid \mathbf{x}_k(\omega) \underset{k\to+\infty}{\longrightarrow} \mathbf{x}(\omega)\right\}\right) = 1. \tag{6.10}$$

If so, $\boldsymbol{x}_\infty$ is named the almost-sure limit of $(\mathbf{x}_k)_{k\in\mathbb{N}}$.

More generally, $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is said to converge almost-surely if there exists a random variable to which the latter converges a.s..

The measurability being preserved by passing to the simple limit, a process $(\mathbf{x}_k)_{k\in\mathbb{N}}$ converges almost-surely every time the set of $\omega \in \Omega$ for which trajectory $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ admits a finite limit is of probability one. The rest of our study is dedicated to the exhibition of a theoretical toolbox in order to obtain convergence results of this nature. As we shall see in the following, this choice is especially

motivated by the close proximity of our framework to that of discrete-time martingales [234]. However, we recall for the reader the most common modes of convergence encountered in the literature.

Definition 6.2. *A stochastic process* $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *is said to converge to a random variable* $\mathbf{x}_\infty : \Omega \to \mathcal{H}$

(i) *In* L^p *norm* $(p \geq 1)$ *if* $\mathbb{E}(\|\mathbf{x}_k - \mathbf{x}_\infty\|^p) \underset{k\to+\infty}{\longrightarrow} 0$.

(ii) *In probability if* $\mathbb{P}\left(\{\omega \in \Omega \mid \|\mathbf{x}_k(\omega) - \mathbf{x}_\infty(\omega)\| \geq \varepsilon\}\right) \underset{k\to+\infty}{\longrightarrow} 0$ *for all* $\varepsilon > 0$.

(iii) *In distribution if* $\mathbb{E}(g(\mathbf{x}_k)) \underset{k\to+\infty}{\longrightarrow} \mathbb{E}(g(\mathbf{x}_\infty))$ *for all* $g : \mathcal{H} \to \mathbb{R}$ *continuous and bounded function.*

In particular, L^p and almost-sure convergences, separately, imply convergence in probability and convergence in probability implies convergence in distribution. There also exist some partial reciprocals that we will not go into in detail as this would take us away from the initial discussion. We invite the curious reader to consult the reference book [22] for more details on this topic.

6.3.2 Probabilistic version of descent concept

We introduced in Chapter 2 (subsection 2.4.2) the notion of (l, r)-descent condition for a deterministic sequence. The goal of this subsection is to propose an alternative version so as to remain consistent with the stochastic framework.

6.3.2.1 *Almost-Sure vs In Expectation approaches*

The first intuitive approach would be to rely on a definition of (l, r)-descent condition directly based on the trajectories of the process thus keeping the analogy with almost-sure convergence. In such a way, $(\mathbf{x}_k)_{k\in\mathbb{N}}$ may said to satisfy a (l, r)-descent condition if the set of $\omega \in \Omega$ for which $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ verify an (l, r)-descent condition is a probability-one event. However, such an approach faces a major obstacle, namely the ω-dependence of the Lyapunov function and the residual. This disadvantage is relatively compromising insofar as these two objects live in spaces with relatively complex topological structures (l is an application from $\mathcal{H}^{\mathbb{N}}$ to $\mathbb{R}^{\mathbb{N}}$ while r lies in $\mathbb{R}^{\mathbb{N}}$) and in fact obtaining even their measurability is relatively delicate (this typically requires to define a σ-algebra on the space of applications from $\mathcal{H}^{\mathbb{N}}$ to $\mathbb{R}^{\mathbb{N}}$). Even if the latter property is ensured, the structure of the resulting random variables will make them difficult to manipulate anyway.

A second approach, the opposite of the first, would consist in only focusing on behaviour of the process at stake in average; instead on directly investigating on the trajectories of $(\mathbf{x}_k)_{k\in\mathbb{N}}$, only the evolution of the first moment information, i.e. of the expectation, is scrutinised here. Although this approach has the advantage of ensuring the decay of l in expectation, it does not allow to make the link with what we are really looking for, i.e. the setting up of almost sure type behaviours. More generally, the transition to expectation tends to mask a large part of the stochastic information we need.

In a nutshell, the two introduced approaches induce the use of a quantity of information that is far too unbalanced, either in excess (for the first approach, by taking an interest in all the trajectories) or in default (for the second approach, by using only the first moment) for the use that one wishes to make of it. It is therefore necessary to adopt an alternative strategy with a better theoretical trade-off.

6.3.2.2 *Stochastic $(l, \mathbf{r})$-descent condition as a first step*

In contrast to the expectation operator which reduces the properties of a random variable to a scalar quantity, the conditional expectation one appears to be more flexible insofar as it sends back an estimate of the said variable with respect to a certain quantity of the observed information [22]. For such a reason, it stands as a legitimate tool to build probabilistic version of the (l, r)-descent condition.

Definition 6.3. *(Stochastic descent condition) Consider $(\mathcal{F}_k)_{k\in\mathbb{N}}$ a filtration defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $l : \mathcal{H}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ be an application (defined on the space of sequences of $\mathcal{H}$ and with values in that of sequences of $\mathbb{R}$) and $\mathbf{r} := (\mathbf{r}_k)_{k\in\mathbb{N}}$ a $\mathcal{F}_k$-measurable non-negative process. A stochastic process $(\mathbf{x}_k)_{k\in\mathbb{N}}$ on $\mathcal{H}$ verifies a $(l, \mathbf{r})$-descent condition (regarding $(\mathcal{F}_k)_{k\in\mathbb{N}}$) if there exists $k_0 \in \mathbb{N}$ for which $(\mathbf{L}_k)_{k\in\mathbb{N}} = l\left((\mathbf{x}_k)_{k\in\mathbb{N}}\right)$ is an integrable process and satisfies*

$$(\forall k \geq k_0) \quad \mathbb{E}[\mathbf{L}_{k+1}|\mathcal{F}_k] \leq \mathbf{L}_k - \mathbf{r}_k \quad \text{a.s..} \tag{6.11}$$

- *In such a context, we will commonly speak of a Lyapunov application to name l (relative to $(l, \mathbf{r})$-descent condition) and of a residual sequence to mention $\mathbf{r}$ (relative to $(l, \mathbf{r})$-descent condition).*
- *$(\mathbf{x}_k)_{k\in\mathbb{N}}$ is said to verify a simple descent condition (relative to f) if there exits a $\mathcal{F}_k$-measurable residual process $\mathbf{r}$ for which $(\mathbf{x}_k)_{k\in\mathbb{N}}$ verifies a $(l, \mathbf{r})$-descent condition regarding $l : (\boldsymbol{y}_k)_{k\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}} \mapsto (F(\boldsymbol{y}_k))_{k\in\mathbb{N}}$, i.e. $l_k = F(\mathbf{x}_k)$ a.s. for any k starting from a certain rank.*

Filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ attached to a $(l, \mathbf{r})$-descent accounts for all past information of the process. We can make a parallel with a gambling game situation. Considering $(\mathbf{L}_k)_{k\in\mathbb{N}}$ as a player ressource, descent inequality (6.11) simply indicates that it is expected for the latter and from a fixed turn, to suffer in average from an impoverishment whose amounts $(\mathbf{r}_k)_{k\in\mathbb{N}}$ only depends on the past of the game. In technical terms, (6.11) induces that Lyapunov process $(\mathbf{L}_k)_{k\in\mathbb{N}}$ follows a supermartingale behaviour and thus, in the case where the latter is almost-surely non-negative, the Doob's "Forward" convergence theorem [234] ensures its almost-sure convergence to an almost-sure finite random variable. Even more precisely, the following result also provides valuable information on the residual process $(\mathbf{r}_k)_{k\in\mathbb{N}}$.

Proposition 6.1. *Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a stochastic process following a $(l, \mathbf{r})$-descent condition regarding a filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ and for which the attached Lyapunov process $(\mathbf{L}_k)_{k\in\mathbb{N}}$ is almost-surely non-negative. Then, $(\mathbf{L}_k)_{k\in\mathbb{N}}$ almost-surely converges to an almost-sure finite random variable and $\sum_{k=0}^{+\infty} \mathbf{r}_k < +\infty$ almost surely.*

Proof. As mentioned, the first point is a direct consequence of Doob's "Forward" convergence theorem [234]. To also obtain the almost-sure summability of $(\mathbf{r}_k)_{k\in\mathbb{N}}$, we need to refine the latter result by considering process $(\mathbf{L}'_k)_{k\in\mathbb{N}}$ such that

$$(\forall k \geq 1) \quad \mathbf{L}'_k := \mathbf{L}_k + \sum_{i=0}^{k-1} \mathbf{r}_i. \tag{6.12}$$

Descent inequality (6.11) ensures that $\mathbf{r}_k \leq |\mathbf{L}_k| + |\mathbb{E}[\mathbf{L}_{k+1}|\mathcal{F}_k]|$ and $\mathbf{r}_k$ is thus integrable for all $k \geq k_0$. The use of conditional expectation to $\mathbf{r}_k$ $(k \geq k_0)$ is then legitimate and $\mathbb{E}[\mathbf{r}_k|\mathcal{F}_k] = \mathbf{r}_k$ due to its $\mathcal{F}_k$-measurability. By linearity and using $(\mathbf{L}'_k)_{k\in\mathbb{N}}$, (6.11) can be rewritten as

$$(\forall k > k_0) \quad \mathbb{E}[\mathbf{L}'_{k+1}|\mathcal{F}_k] \leq \mathbf{L}'_k \quad \text{a.s..} \tag{6.13}$$

$(\mathbf{L}'_k)_{k\in\mathbb{N}}$ therefore also follows a supermartingale behaviour and since it is basically non-negative, we deduce its almost-sure convergence to an almost-sure finite random variable. The latter point combined with definition (6.12) and non-negativity of $(\mathbf{L}_k)_{k\in\mathbb{N}}$ finally conduct process $\left(\sum_{i=0}^{k-1}\mathbf{r}_i\right)_{k\geq 1}$ to admit a finite limit almost-surely which concludes the proof. □

Proposition 6.1 can be seen as the stochastic analog of the deterministic result discussed in subsection 2.3.3. In such a case, a monotone convergence argument was sufficient to conclude about the summability of the residuals while here it is necessary to use more general tools.

Example 6.4. *We aim to minimize F using the stochastic gradient scheme $\mathbf{x}_{k+1}=\mathbf{x}_k-\alpha\mathbf{g}_k$ $(\alpha>0)$ for all $k\in\mathbb{N}$. Considering canonical filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}=(\sigma(\mathbf{x}_0,...\mathbf{x}_k))_{k\in\mathbb{N}}$ we admit that $(\mathbf{g}_k)_{k\in\mathbb{N}}$ is a conditionally unbiased gradient approximation whose conditional variance is bounded w.r.t the exact derivative, i.e. for any $k\in\mathbb{N}$, $\mathbb{E}[\mathbf{g}_k|\mathcal{F}_k]=\nabla F(\mathbf{x}_k)$ and $\mathbb{E}[\|\mathbf{g}_k\|^2|\mathcal{F}_k]\leq C\|\nabla F(\mathbf{x}_k)\|^2$ $(C>0)$ almost surely. Then, $(\mathbf{x}_k)_{k\in\mathbb{N}}$ follows a simple descent condition as soon as F is L-Lipschitz continuous gradient and stepsize α chosen so as to verify $\alpha<2/(LC^2)$:*

$$(\forall k\in\mathbb{N})\quad \mathbb{E}[F(\mathbf{x}_{k+1})|\mathcal{F}_k]\leq F(\mathbf{x}_k)-\alpha\left(1-\frac{\alpha LC^2}{2}\right)\|\nabla F(\mathbf{x}_k)\|^2\ \text{ a.s..} \tag{6.14}$$

The proof of (6.14) relies on the usual descent Lemma 2.4 and can be found in [212]. If F is bounded below and considering positive process $(F(\mathbf{x}_k)-F_{inf})_{k\in\mathbb{N}}$ (where F_{inf} is the minimal value of F), Proposition 6.1 can be invoked; $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ converges almost surely while $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$ converges almost surely, more precisely, to zero, due to $\sum_{k=0}^{+\infty}\|\nabla F(\mathbf{x}_k)\|^2<+\infty$a.s..

6.3.2.3 *Quasi-supermartingale*

The previous subsection highlighted the importance of conditioning theory for the asymptotical study of stochastic algorithms. Although the relative simplicity of stochastic descent inequality (6.11) has the advantage of being very easily interpretable in term of convergence guarantees (typically through Proposition 6.1), it also remains quite restrictive to the extent that only few schemes are likely to verify such a behavior. One of the main reason is that manipulating objects of probabilistic nature actually tends to promote occurrences of additional or multiplicative specific terms which highly complicate the structure of the Lyapunov function l to be considered. Such an issue invites to propose a relaxed alternative of the initial stochastic descent condition through the almost-supermatingale concept of H.Robbins and D.Siegmund [202] .

Definition 6.4. *(Almost-supermartingale) Consider a filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ defined on probability space $(\Omega,\mathcal{F},\mathbb{P})$. Let $l:\mathcal{H}^{\mathbb{N}}\to\mathbb{R}^{\mathbb{N}}$ be an application and three non-negative $\mathcal{F}_k$- measurable processes $\mathbf{u}:=(\mathbf{u}_k)_{k\in\mathbb{N}},\boldsymbol{v}:=(\boldsymbol{v}_k)_{k\in\mathbb{N}},\boldsymbol{w}:=(\mathbf{w}_k)_{k\in\mathbb{N}}$. A stochastic process on $(\mathbf{x}_k)_{k\in\mathbb{N}}$ of $\mathcal{H}$ is said to be a $(l,\mathbf{u},\mathbf{v},\mathbf{w})$-almost-supermartingale (regarding $(\mathcal{F}_k)_{k\in\mathbb{N}}$) if there exists $k_0\in\mathbb{N}$ for which $(\mathbf{L}_k)_{k\in\mathbb{N}}=l\left((\mathbf{x}_k)_{k\in\mathbb{N}}\right)$ is an integrable process and satisfies*

$$(\forall k\geq k_0)\quad \mathbb{E}(\mathbf{L}_{k+1}|\mathcal{F}_k)\leq(1+\mathbf{u}_k)\mathbf{L}_k+\boldsymbol{v}_k-\mathbf{w}_k\ \text{ a.s..} \tag{6.15}$$

- *In such a context, we will commonly speak of a Lyapunov application to name l and of a residual sequence to mention $\mathbf{w}$.*

- $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *is said to be a simple almost-supermartingale (relative to* f*) if there exist three non-negative* $\mathcal{F}_k$*-measurable processes* $\mathbf{u},\mathbf{v},\mathbf{w}$ *for which* $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *is a* $(l,\mathbf{u},\mathbf{v},\mathbf{w})$*-almost-supermartingale with* $l:(\boldsymbol{y}_k)_{k\in\mathbb{N}}\in\mathcal{H}^{\mathbb{N}}\mapsto(F(\boldsymbol{y}_k))_{k\in\mathbb{N}}$*, i.e.* $\mathbf{L}_k=F(\mathbf{x}_k)$ *a.s. for any* k *starting from a certain rank.*

The class of $(l,\mathbf{u},\mathbf{v},\mathbf{w})$-almost-supermartingale processes especially encompasses those following a $(l,\mathbf{r})$-descent condition in the particular situation where $\mathbf{w}=\mathbf{r}$ and $\mathbf{u},\mathbf{v}$ can be chosen as zero almost surely. Similarly with Proposition 6.1 for stochastic descent condition, there exists a key asymptotical result associated to the very notion of almost-supermatingale. This can be seen as the most fundamental almost-sure convergence theorem in the field of stochastic approximation:

Theorem 6.2. *(Robbins-Siegmund) Let* $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *be a non-negative* $(l,\mathbf{u},\mathbf{v},\mathbf{w})$*-almost-supermartingale for which* $\sum_{k=0}^{+\infty}\mathbf{u}_k<+\infty$ *and* $\sum_{k=0}^{+\infty}\mathbf{v}_k<+\infty$ *almost surely. Then, the attached Lyapunov process* $(\mathbf{L}_k)_{k\in\mathbb{N}}$ *almost surely converges to a random variable* $\mathbf{L}_\infty$*. Moreover,* $\mathbf{L}_\infty$ *and* $\sum_{k=0}^{+\infty}\mathbf{w}_k$ *are finite almost-surely.*

Proof. The proof strategy relies on the same argument as those considered to establish Proposition 6.1 writing (6.15) under a supermartingale form. The complete proof can be found in its original version in [202] or more recently in [89]. □

As such, Robbins-Siegmund Theorem 6.2 can be interpreted as a relaxed version of Proposition 6.1 stipulating that the addition or multiplication of summable "spurious term" $(\mathbf{u},\mathbf{v})$ in the descent condition finally does not compromise the almost-sure convergence of the Lyapunov function as well as the summability of the residual. The construction of a generalization of the descent condition (6.11) was historically conducted in intermediate steps. In this context, Theorem 6.2 can be seen as an extension of the work of [112] which was the first one to take into consideration terms of type of $\mathbf{u},\mathbf{v}$ (but of deterministic nature) interfering in a stochastic descent.

Example 6.5. *Our goal is here to minimize a differentiable function* F *admitting a stationary point* $\boldsymbol{x}^*$ *which the monotonic relation* $\langle\nabla F(\boldsymbol{x}),\boldsymbol{x}-\boldsymbol{x}^*\rangle<0$ *for all* $\boldsymbol{x}\in\mathcal{H}$*. To do so, we here consider a scheme of the form of* $\mathbf{x}_{k+1}=\mathbf{x}_k-\alpha_k\mathbf{g}_k$ $(\alpha_k>0)$ *for all* $k\in\mathbb{N}$ *and similarly with Example 6.4, we suppose that* $(\mathbf{g}_k)_{k\in\mathbb{N}}$ *is a conditionally unbiased gradient approximation regarding canonical filtration* $(\mathcal{F}_k)_{k\in\mathbb{N}}=(\sigma(\mathbf{x}_0,\dots,\mathbf{x}_k))_{k\in\mathbb{N}}$*, i.e.* $\mathbb{E}[\mathbf{g}_k|\mathcal{F}_k]=\nabla F(\mathbf{x}_k)$ *almost surely for any* $k\in\mathbb{N}$*. Moreover, we place ourselves in the situation where the second order conditional moment verifies for all* $k\in\mathbb{N}$*,* $\mathbb{E}[\|\mathbf{g}_k-\nabla F(\mathbf{x}_k)\|^2|\mathcal{F}_k]\leq C(1+\|\mathbf{x}_k-\boldsymbol{x}^*\|^2)$ *almost surely. Following the proof of [89, Theorem 1.4.26], process* $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *satisfies:*

$$(\forall k\in\mathbb{N})\quad \mathbb{E}\left[\|\mathbf{x}_{k+1}-\boldsymbol{x}^*\|^2|\mathcal{F}_k\right]\leq(1+C\alpha_k^2)\|\mathbf{x}_k-\boldsymbol{x}^*\|^2+C\alpha_k^2+2\alpha_k\langle\nabla F(\mathbf{x}_k),\mathbf{x}_k-\boldsymbol{x}^*\rangle\ \text{ a.s.}\quad(6.16)$$

and is therefore a $(l,\mathbf{u},\mathbf{v},\mathbf{w})$*-almost-supermartingale taking* $l:(\boldsymbol{y}_k)_{k\in\mathbb{N}}\mapsto\left(\|\boldsymbol{y}_k-\boldsymbol{x}^*\|^2\right)_{k\in\mathbb{N}}$*, process* $\mathbf{u},\mathbf{v}$ *here deterministic as* $\mathbf{u}=\mathbf{v}:=\left(C\alpha_k^2\right)_{k\in\mathbb{N}}$ *and* $\mathbf{r}:=(2\alpha_k\langle\nabla F(\mathbf{x}_k),\mathbf{x}_k-\boldsymbol{x}^*\rangle)_{k\in\mathbb{N}}$*. Especially, we easily observe that condition* $\sum_{k=0}^{+\infty}\alpha_k^2<+\infty$ *is a sufficient one to legitimate the use of Theorem 6.2.*

6.3.3 Making the link between almost-sure convergence of the iterates

One of the advantage of "almost-sure" notion lies in its ability of promoting the most natural extension of deterministic behaviors. In such a context, the strategy of proof follows a common path. When one aims to demonstrate that a property $\mathbf{Q}$ is satisfied almost surely starting from another one $\mathbf{P}$ verified almost-surely, the easiest approach consists in starting from the existence of a probability-one set $\Lambda \in \mathcal{F}$ for which $\mathbf{P}(\omega)$ is true for all $\omega \in \Lambda$ and therefore to reason from a deterministic point of view fixing every $\omega \in \Lambda$. It can then be deduced that $\mathbf{P}(\omega)$ implies $\mathbf{Q}(\omega)$ for all $\omega \in \Lambda$ and finally that $\mathbf{P}$ implies $\mathbf{Q}$ almost-surely keeping in mind that $\mathbb{P}(\Lambda) = 1$. We will regularly use this strategy throughout the two next chapters and, in particular, to prove a stochastic version of Proposition 2.2.

However, this same approach also presents some disadvantage in some specific contexts. When property $\mathbf{Q}$ implies some mathematical objects, the latters always become dependent from ω-variable and are thus in turn random variables (subject to measurability, which is not always obvious). A case of particular interest to us is the one relative to the Kurdyka-Łojasiewicz theory. As seen in section 2.5 of Chapter 2, analyzing convergence of the iterates in the deterministic non-convex framework can be done through the uniform KL Property 2.6 involving the existence of three objects $\zeta, \varepsilon \in (0, +\infty)$ and $\varphi : [0, \zeta) \to \mathbb{R}_+$. If we use our previous strategy, Property 2.6 is still verified almost-surely but if so $\zeta, \varepsilon, \varphi$ fatally become ω-variables making the rest of the convergence proof highly challenging. As a response to this issue, our Chapter 8 is especially dedicated to build an alternative version of the uniform KL property so as to be better adapted to a stochastic setting and for which the ω-dependencies of $\zeta, \varepsilon, \varphi$ are alleviated.

6.4 Conclusion

As a mirror of Chapter 2, Chapter 6 was devoted to introducing the stochastic methods commonly found in the literature, as well as the theoretical background that we will use to conduct our analysis in Chapters 7 and 8. In particular, the main convergence results explained therein will each time be based on the establishment of a stochastic descent condition, more specifically with respect to the cost function itself in Chapter 7 (i.e. a simple stochastic descent condition) and on a particular Lyapunov function in Chapter 8. As mentioned, one of the aims of this thesis is to refine the properties outlined in section 6.3 so as ideally to obtain direct results on the almost-sure convergence of process $(\mathbf{x}_k)_{k\in\mathbb{N}}$ itself. Although such results are obtained under a convexity assumption in Chapter 7, this will no longer be the case in Chapter 8.

Chapter 7

Sabrina: A stochastic subspace majorization-minimization algorithm

A wide class of problems involves the minimization of a coercive and differentiable function F on $\mathbb{R}^N$ whose gradient cannot be evaluated in an exact manner. In such context, many existing convergence results from standard gradient-based optimization literature cannot be directly applied and robustness to errors in the gradient is not necessarily guaranteed. This work is dedicated to investigating the convergence of Majorization-Minimization (MM) schemes when stochastic errors affect the gradient terms. We introduce a general stochastic optimization framework, called SABRINA (StochAstic suBspace majoRIzation-miNimization Algorithm) that encompasses MM quadratic schemes possibly enhanced with a subspace acceleration strategy, as introduced in Chapter 3. New asymptotical results are built for the stochastic process generated by SABRINA. Two sets of numerical experiments in the field of machine learning and image processing are presented to support our theoretical results and illustrate the good performance of SABRINA with respect to state-of-the-art gradient-based stochastic optimization methods.

This work is based the our article: E. Chouzenoux and J-B. Fest. *SABRINA: A Stochastic Subspace Majorization-Minimization Algorithm*, published in Journal of Optimization Theory and Applications, vol. 195, pp. 919-952 2022.

Contents

7.1 Introduction

We consider the problem:

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \quad F(\boldsymbol{x}), \tag{7.1}$$

where $F : \mathbb{R}^N \to \mathbb{R}$ is a coercive and differentiable function on $\mathbb{R}^N$. We focus on the case when the gradient of F is altered by stochastic errors during the iterative optimization process. This problem has been widely studied in the optimization literature, starting from seminal works [92, 201], and has known a renewed interest in the last decade with applicative challenges arising in supervised learning on large scale datasets [13, 35]. The stability properties of gradient-based stochastic schemes are also of high interest in approximate Bayesian inference, where stochastic gradient steps are often used to improve the exploration capacities of the samplers [91, 184, 148, 239].

Probably the most relevant gradient-based stochastic optimizer is the stochastic gradient descent (SGD) algorithm, studied in [92, 20, 201]. Extension of SGD to non-differentiable case using proximal-based tools can be found in [70, 7, 4, 138]. Few convergence studies made in the deterministic case extend straightforwardly to the stochastic case. All the aforementioned works are grounded on specific probabilistic tools such as [202, 89]. SGD is rather simple but can exhibit slow convergence. Therefore, many recent works have focused on deriving accelerated variants of it. Two main families of acceleration strategies can be distinguished in the literature. The first approach, adopted for example in [105, 137, 175, 88, 157], relies on subspace (i.e., momentum) acceleration. The convergence rate is improved by using information from past iterates for the construction of new estimates. The second approach to accelerate the convergence of SGD is based on a variable metric strategy [87, 47]. The underlying metric is modified at each iteration thanks to a preconditioning matrix, which may incorporate second-order information about the function to minimize. These acceleration techniques give rise to promising practical results.

This work proposes a novel SGD-based scheme to solve Problem (7.1), by combining the two aforementioned acceleration strategies. To do so, we rely on the so-called Majorization-Minimization (MM) principle [246, 223].

At each iteration of an MM algorithm, a surrogate function majorizing the problem cost function is constructed. The next iterate is then obtained by minimizing the latter surrogate. By construction, MM method produces a sequence of iterates that decreases the cost function monotonically. MM algorithms benefit from assessed convergence properties in the convex and non-convex settings [128, 26, 59]. The extension of MM methodology to the stochastic context has been studied recently in [80, 161, 63] in restricted scenarios. The method proposed by [80] is dedicated to introducing stochastic errors into the expectation-minimization approach, a special case of MM. The MISO approach from [161] combines an MM scheme with constraining averaging rules both over surrogates and iterates to reach convergence. The work of [63] studies a scheme close to the one proposed in our chapter, but limits the analysis to the specific case of a penalized least-square criterion whose gradient is evaluated using a recursive least-squares implementation [97]. In this present work, we introduce a versatile MM scheme relying on quadratic majorization surrogates for F and allowing for subspace acceleration [58, 199]. In a nutshell, the resulting algorithm benefits from a simple structure that can be understood as an SGD method with both preconditioning and momentum-based term, and has minimal parameter tuning.

For the proposed scheme, our contributions are:

- almost sure convergence results for non necessarily convex F;
- convergence rate analysis in the strongly convex case;
- illustration of the performance and comparison with state-of-the-art on two numerical examples.

The rest of this chapter is organized as follows. Section 7.2 states notations and introduces the considered MM stochastic optimization scheme. Assumptions and some technical lemmas, essential for our theoretical study, are presented in Section 7.3. Our main contribution is concentrated in Section 7.4 presenting our convergence results and convergence rate analysis. Numerical experiments are provided in Sections 7.5 and 7.6. Finally, we conclude the chapter in Section 7.7.

7.2 Background and proposed formulation

7.2.1 Notations

We classically denote by $\|\cdot\| = \langle\cdot|\cdot\rangle$ the euclidean norm of $\mathbb{R}^N$, and $|||\cdot|||$ the spectral norm (i.e., largest singular value) of elements of $\mathbb{R}^{M\times N}$. If $\boldsymbol{M}$ is a symmetric definite positive matrix of $\mathbb{R}^{N\times N}$, $\|.\|_{\boldsymbol{M}}$ corresponds to $\langle .|\boldsymbol{M}.\rangle$. Moreover, we will use the Loewner's order $\preceq$ between two symmetric matrices $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ of $\mathbb{R}^{N\times N}$, where relation $\boldsymbol{M}_1 \preceq \boldsymbol{M}_2$ holds if and only if difference $\boldsymbol{M}_2 - \boldsymbol{M}_1$ is (symmetric) positive. $\boldsymbol{I}_N$ states for the identity matrix of $\mathbb{R}^N$, $\boldsymbol{0}_N$ the zero vector of size N, and $\boldsymbol{O}_N$ the null matrix of $\mathbb{R}^{N\times N}$. Bold symbols are used for matrix and vectors. Italic style is retained for deterministic quantities.

For every $v \in \mathbb{R}$, the level set of F relative to v will be denoted:

$$\operatorname{lev}_{=v} F := \left\{ \boldsymbol{x} \in \mathbb{R}^N \;\middle|\; F(\boldsymbol{x}) = v \right\}.$$

Subject to existence, $\widetilde{\boldsymbol{x}}$ will state for a stationary point of F. Moreover, $\operatorname{zer} \nabla F$ will denote the set of stationary points of F. We will write $\boldsymbol{x}^*$ a global minimizer for F and define $F^* := F(\boldsymbol{x}^*)$.

7.2.2 Quadratic MM (QMM) approach

MM algorithm solves Problem (7.1) iteratively by generating a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ of elements of $\mathbb{R}^N$, where the step from the iterate $\boldsymbol{x}_k$ to its successor $\boldsymbol{x}_{k+1}$ is achieved through the minimization of $h(\cdot, \boldsymbol{x}_k)$, a quadratic tangent majorization surrogate of F around $\boldsymbol{x}_k$, i.e.

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad h(\boldsymbol{x}, \boldsymbol{x}_k) \geq F(\boldsymbol{x}) \quad \text{and} \quad h(\boldsymbol{x}_k, \boldsymbol{x}_k) = F(\boldsymbol{x}_k). \tag{7.2}$$

An efficient strategy consists in resorting to a quadratic tangent majorization function, structurally analogous to a second-order Taylor's expansion of F:

$$h : (\boldsymbol{x}, \boldsymbol{y}) \mapsto F(\boldsymbol{y}) + \nabla F(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) + \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2_{\boldsymbol{A}(\boldsymbol{y})}. \tag{7.3}$$

Hereabove, for every $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{A}(\boldsymbol{y})$ is a symmetric positive definite matrix of $\mathbb{R}^{N\times N}$ chosen so as to ensure (7.2). The latter, called the majorization metric matrix, yields a complete description of $h(\cdot, \boldsymbol{y})$ and thus influences the approximation quality of F by this same surrogate. Several techniques for building suitable majorization metric matrices can be found for a wide class of problems encompassing image restoration, telecommunication or supervised learning in [246, 223, 59].

As a consequence of the invertibility of $\boldsymbol{A}(\boldsymbol{x}_k)$, for every $k \in \mathbb{N}$, we obtain the generic Quadratic MM (QMM) scheme [223]:

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} &= \operatorname{argmin}_{\boldsymbol{x}\in\mathbb{R}^N} \quad h(\boldsymbol{x}, \boldsymbol{x}_k),\\ &= \boldsymbol{x}_k - \boldsymbol{A}_k^{-1}\nabla F(\boldsymbol{x}_k)\end{aligned} \tag{7.4}$$

with $\boldsymbol{A}_k := \boldsymbol{A}(\boldsymbol{x}_k)$ and $\boldsymbol{x}_0 \in \mathbb{R}^N$. The QMM update (7.4) can be shown to map with the half-quadratic algorithm [108] when F is a penalized least-squares function. By construction, the sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ built by (7.4) guarantees a monotonic decrease of $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$. Convergence of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ to a stationary point of F can be shown under suitable technical assumptions on F and $(\boldsymbol{A}_k)_{k\in\mathbb{N}}$ [5].

7.2.3 Subspace acceleration

When using update (7.4), one needs to invert an $N \times N$ matrix. Such an operation is undesirable when N is large. The authors from [58] proposed to integrate a so-called subspace acceleration procedure [199, 242] into (7.4) leading to:

$$\begin{aligned}(\forall k \in \mathbb{N}) \qquad \boldsymbol{x}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{x}\in\operatorname{ran}(\boldsymbol{D}_k)} h(\boldsymbol{x}, \boldsymbol{x}_k)\\ &= \boldsymbol{x}_k + \boldsymbol{D}_k\boldsymbol{u}_k,\end{aligned} \tag{7.5}$$

with

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{u}_k \in \operatorname*{arg\,min}_{\boldsymbol{u}\in\mathbb{R}^{M_k}} h(\boldsymbol{x}_k + \boldsymbol{D}_k\boldsymbol{u}, \boldsymbol{x}_k), \tag{7.6}$$

and $\boldsymbol{x}_0 \in \mathbb{R}^N$. The key ingredient of the above method is the introduction of a matrix $\boldsymbol{D}_k \in \mathbb{R}^{N\times M_k}$ with $N \geq M_k \geq 1$, whose range $\operatorname{ran}(\boldsymbol{D}_k)$ (i.e., vectorial space spanned by the columns of $\boldsymbol{D}_k$) imposes a subspace to search for the new iterate $\boldsymbol{x}_{k+1}$. Taking $M_k = N$ and $\boldsymbol{D}_k = \boldsymbol{I}_N$, the identity matrix of $\mathbb{R}^N$, (7.5) goes back to scheme (7.4). In practice, only a few degrees of freedom are actually required to reach good convergence speed (see [62] for a detailed analysis of the convergence rate of scheme (7.5) as a function of $\boldsymbol{D}_k$ and $\boldsymbol{A}_k$), so M_k is typically retained as very small compared to N. Interesting choices can be found in [58, Tab.1]. Setting $\boldsymbol{D}_k = [-\nabla F(\boldsymbol{x}_k) \quad | \quad \boldsymbol{x}_k - \boldsymbol{x}_{k-1}]$ (with convention $\boldsymbol{x}_{-1} = \boldsymbol{x}_0$) brings notably to the so-called MM Memory Gradient (3MG) method whose great performances have been illustrated in [102, 58, 217]. Other choices for the subpace matrix can be found in [199, 164, 220, 249]. It is worth noting that the minimization scheme (7.5) shares strong connections with non-linear conjugate gradient algorithm [178], low-memory quasi-Newton approaches such as L-BFGS [178, 156], trust-region strategies [2], and momentum-based methods [224]. In contrast with these aforementioned works, the MM subspace scheme presents the key advantage of a simple linesearch procedure (7.6) associated with sounded convergence guarantees. Indeed, assuming, without

loss of generality that $\boldsymbol{D}_k$ has full column rank, the quadratic structure of $h(\cdot, \boldsymbol{x}_k)$ allows to obtain an analytical solution to sub-problem (7.6).

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{u}_k = -\left(\boldsymbol{D}_k^\top \boldsymbol{A}_k \boldsymbol{D}_k\right)^{-1} \boldsymbol{D}_k^\top \nabla F(\boldsymbol{x}_k). \tag{7.7}$$

The interest lies here in the fact that $\boldsymbol{D}_k^\top \boldsymbol{A}_k \boldsymbol{D}_k$ is an $M_k \times M_k$ matrix making its inversion far easier computable than the inversion of $\boldsymbol{A}_k$, as soon as M_k is small. Convergence properties of (7.5)-(7.6) have been established in the convex setting in [58], and extended to the non-convex setting in [59] using recent tools of non smooth analysis. The catalizing effect of the subspace acceleration for practical convergence speed of MM methods has been acknowledged in the survey chapter [223]. We also refer the reader to [49, 110, 43] for practical implementation of MM subspace approaches on modern high performance computing tools.

7.2.4 SABRINA, a stochastic subspace MM algorithm

We are now ready to introduce the algorithm studied in this chapter. We focus on the stability of the optimization scheme (7.5)-(7.6) when the gradient of F is affected by an additive stochastic perturbation at each iteration $k \in \mathbb{N}$, so that only the approximate value $\mathbf{g}_k$, defined below, is available:

$$(\forall k \in \mathbb{N}) \quad \mathbf{g}_k = \nabla F(\mathbf{x}_k) + \mathbf{e}_k. \tag{7.8}$$

Hereabove, $(\mathbf{e}_k)_{k\in\mathbb{N}}$ corresponds to a zero-mean stochastic process with a bounded variance in a sense that will be specified in Section 7.3.2. Formulating the stochastic counterpart of (7.5)-(7.6) requires to introduce the concept of inexact majorization function. For every $k \in \mathbb{N}$, the majorization function $h(\cdot, \mathbf{x}_k)$ will be substituted by a new function $\widehat{\mathrm{h}}_k$ with the following expression:

$$\widehat{\mathrm{h}}_k : \boldsymbol{u} \in \mathbb{R}^N \mapsto F(\mathbf{x}_k) + \mathbf{g}_k^\top (\boldsymbol{u} - \mathbf{x}_k) + \frac{1}{2}\|\boldsymbol{u} - \mathbf{x}_k\|_{\mathbf{A}_k}^2. \tag{7.9}$$

In analogy with the deterministic formulation from Section 7.2.3, the update at iteration $k \in \mathbb{N}$ will be grounded on the search of a minimizer of $\widehat{\mathrm{h}}_k$ along the directions spanned by the columns of a matrix $\mathbf{D}_k \in \mathbb{R}^{M_k \times N}$.

Let us also introduce a positive stepsize sequence $(\gamma_k)_{k\in\mathbb{N}}$ in order to promote stability of the iterates. This finally leads us to our stochastic minimization scheme called SABRINA (StochAstic suBspace majoRization mINimization Algorithm):

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \gamma_k \mathbf{D}_k \mathbf{u}_k, \tag{7.10}$$

with

$$(\forall k \in \mathbb{N}) \quad \mathbf{u}_k = -\left(\mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k\right)^{-1} \mathbf{D}_k^\top \mathbf{g}_k, \tag{7.11}$$

and $\mathbf{x}_0 \equiv \boldsymbol{x}_0 \in \mathbb{R}^N$, a deterministic quantity.

Remark 7.1. *For the sake of clarity, throughout the chapter, we distinguish deterministic and random quantities, with italic and non-italic styles, respectively. In particular, since the noise $(\mathbf{e}_k)_{k\in\mathbb{N}}$ is random, the quantities $(\mathbf{g}_k, \mathbf{x}_k, \mathbf{D}_k, \mathbf{u}_k)_{k\in\mathbb{N}}$ are too. The probabilistic notations (i.e., probability space, filtration), useful for our theoretical analysis, will be made explicit in Sec. 7.3.1.*

7.2.5 Link with stochastic preconditioned gradient algorithm

It is straightforward to rewrite SABRINA iterations (7.10)-(7.11) under the compact form:

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k \mathbf{B}_k \mathbf{g}_k, \tag{7.12}$$

with

$$(\forall k \in \mathbb{N}) \quad \mathbf{B}_k = \mathbf{D}_k \left(\mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k \right)^{-1} \mathbf{D}_k^\top. \tag{7.13}$$

The above formulation is interesting as it highlights similarities between SABRINA and the preconditioned gradient scheme with inexact gradient term, studied for instance in [42, 33]. The main distinction is that the symmetric matrix $\mathbf{B}_k \in \mathbb{R}^{N \times N}$ involved in (7.12) gathers information brought by the majorization matrix $\mathbf{A}_k$ and by the retained subspace $\mathbf{D}_k$, as described in (7.13). The formulation above suggests that controlling the behaviour of $(\mathbf{x}_k)_{k \in \mathbb{N}}$ requires studying the properties of $(\mathbf{B}_k)_{k \in \mathbb{N}}$, which raises two main theoretical challenges that we plan to tackle in this chapter: (i) $\mathbf{B}_k$ is a random matrix with non necessarily full rank, (ii) F is not assumed to be a convex function. Up to our knowledge, the general scheme (7.12) has never been analysed under these two restrictions.

7.3 Preliminary Lemmas

In this section, we introduce our probabilistic notations. We present and discuss our assumptions. Finally, we prove three technical lemmas that appear essential for establishing our main convergence results presented in Section 7.4.

7.3.1 Probabilist framework

In the remainder of the chapter, we consider $(\Omega, \mathcal{F}, P)$ a probability space to which we associate the filtration $(\mathcal{F}_k)_{k \in \mathbb{N}}$ where $\mathcal{F}_0 = \{\Omega, \emptyset\}$ and for all $k \geq 1$, $\mathcal{F}_k = \sigma(\mathbf{e}_0, \mathbf{x}_1, ..., \mathbf{e}_{k-1}, \mathbf{x}_k)$ corresponds to the sub-sigma algebra generated by the family $\{\mathbf{e}_0, \mathbf{x}_1, ..., \mathbf{e}_{k-1}, \mathbf{x}_k\}$ of random variables. For each $k \in \mathbb{N}$, $\mathcal{F}_k$ gathers all the information available from the origin of the process to iteration k. A mathematical property will be said to be verified *almost surely* or *a.s.* if it holds on a probability-one set belonging to $\mathcal{F}$. We also remind that an element of $\mathcal{F}$ is negligible if it is a probability-zero one. For a given $k \in \mathbb{N}$ and subject to existence, we will denote $\mathbb{E}(.|\mathcal{F}_k)$, the conditional expectancy operator associated to $\mathcal{F}_k$.

7.3.2 Assumptions

The following assumptions will guide us throughout the rest of the study.

Assumption 7.1. *F is coercive and β-Lipschitz differentiable on $\mathbb{R}^N$, i.e. there exists $\beta > 0$ such that:*

$$(\forall (\boldsymbol{x}, \boldsymbol{y}) \in (\mathbb{R}^N)^2) \quad \|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})| \leq \beta \, \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{7.14}$$

Assumption 7.2. *There exists* $(\eta, \nu) > 0$ *such that:*

$$(\forall k \in \mathbb{N}) \quad \eta \boldsymbol{I}_N \preceq \mathbf{A}_k \preceq \nu \boldsymbol{I}_N \quad \text{a.s..} \tag{7.15}$$

Assumption 7.3. *For every iteration* $k \in \mathbb{N}$,

(i) $\operatorname{rank}(\mathbf{D}_k) = M_k \quad$ a.s.,

(ii) $\mathbf{g}_k \in \ker(\mathbf{D}_k^\top)^\perp \quad$ a.s..

Assumption 7.4. *The stochastic noise process* $(\mathbf{e}_k)_{k\in\mathbb{N}}$ *fulfills:*

(i) $(\forall k \in \mathbb{N}) \;\; \mathbb{E}\left(\mathbf{e}_k | \mathcal{F}_k\right) = 0 \quad$ a.s.,

(ii) There exists $C \in (0, C_{\max})$ *with* $C_{\max} = \frac{1}{2}\left((1 + \frac{4\eta}{\nu})^{\frac{1}{2}} - 1\right)$ *such that:*

$$(\forall k \in \mathbb{N}) \;\; \mathbb{E}\left(\|\mathbf{e}_k\|^2 | \mathcal{F}_k\right) \leq C^2 \|\nabla F(\mathbf{x}_k)\|^2 \quad \text{a.s..} \tag{7.16}$$

Assumption 7.5. $(\gamma_k)_{k\in\mathbb{N}}$ *is a sequence of strictly positive scalars satisfying:*

$$\gamma_k \underset{k\to+\infty}{\longrightarrow} 0 \quad \textit{and} \quad \sum_{k=0}^{+\infty} \gamma_k = +\infty.$$

7.3.3 Discussion on the assumptions

Assumption 7.1 is rather standard in the analysis of stochastic gradient-based methods [111, 161]. It is worth noting that the knowledge of the Lipschitz constant of ∇F is not necessary for the practical implementation of the method.

Assumption 7.2 is essential for ensuring convergence of MM methods involving quadratic majorization functions, as it ensures that the majorization metric matrices remain well-conditioned. Let us remark that the existence of such matrices is guaranteed by the descent lemma, since one can set $\mathbf{A}_k \equiv \beta \boldsymbol{I}_N$, with β the Lipschitz constant of ∇F (see Assumption 7.1). For such choice, SABRINA identifies with SGD with specific MM-based closed-form formulas for the stepsize and the momentum weight. As we will show in our experimental tests, it is however usually worthy to search for more sophisticated choices for $(\mathbf{A}_k)_{k\in\mathbb{N}}$, leading usually to faster practical convergence (See also [5, Sec.IV],[62] for the role of majorization mappings in the convergence speed of quadratic MM methods).

Assumptions 7.3(i) and 7.3(ii) work as a peer, and control the validity of the subspace construction. These requirements are standard in subspace-based optimization methods [242, 58, 199]. Assumption 7.3(i) ensures the non-redundancy of the information within the subspace. Assumption 7.3(ii) enhances some descent properties of the algorithm. Note that the latter Assumption is verified as soon as one of the columns of $\mathbf{D}_k$ identifies with $-\mathbf{g}_k$ (i.e. the SGD direction). Interestingly, for $\mathbf{D}_k \equiv -\mathbf{g}_k$, SABRINA reads as a preconditioned SGD algorithm, with MM-based preconditioner.

Assumption 7.4(i) is often required for studying the stability of gradient-based optimization schemes in the presence of stochastic errors [85, 111]. Assumption 7.4(ii) corresponds to a second order moment property and can be seen as a particular case of [35, Assumption 4.3.c]. It states that uncertainty $\mathbf{e}_k$ should remain reasonable with respect to the norm of the (true) gradient of F at $\mathbf{x}_k$. The larger condition number η/ν of the majorization metrics, the more permissive upper bound $C_{\max}$ is. The maximum theoretical bound $\frac{\sqrt{5}-1}{2} \simeq 6.18 \times 10^{-1}$ is reached if and only if $\eta \equiv \nu$. Such a situation occurs for instance when $\mathbf{A}_k$ equals to a positive constant times identity. Typical choice would be $\mathbf{A}_k \equiv \beta \boldsymbol{I}_N$, but, as already mentioned, this choice might be detrimental to the convergence speed. In contrast, one can easily show that $C_{\max} \sim \eta/\nu$ for $\eta/\nu \to 0^+$, which means that poorly conditioned majorization mappings would demand a high level of tolerance on the gradient's uncertainty. This suggests that a compromise must be achieved between the convergence speed and the requirements in terms of stability to noise.

Assumption 7.5 is a relaxed version of the classical σ-sequence hypothesis [104]. In particular, a main feature of our study is that it is not necessary to impose the usual condition $\sum_{k=0}^{+\infty} \gamma_k^2 < +\infty$. Assumption 7.5 allows to choose a stepsize $(\gamma_k)_{k\in\mathbb{N}}$ with a slow convergence to 0 (e.g., an inverse logarithmic one).

7.3.4 Properties of the preconditioning matrices

As mentioned in Section 7.2.5, the behaviour of SABRINA iterates depends on the properties of $(\mathbf{B}_k)_{k\in\mathbb{N}}$ expressed in (7.13). We derive some useful technical properties for these matrices, gathered in the lemma below.

Lemma 7.1. *Under Assumptions 7.2 and 7.3(i), for all $k \in \mathbb{N}$, $\mathbf{B}_k$ is almost surely well-defined and satisfies:*

$$\mathbf{D}_k \mathbf{u}_k = -\mathbf{B}_k \mathbf{g}_k, \tag{7.17a}$$

$$\boldsymbol{O}_N \preceq \mathbf{B}_k \preceq \frac{1}{\eta} \boldsymbol{I}_N, \tag{7.17b}$$

$$(\forall \boldsymbol{x} \in \ker(\mathbf{D}_k^\top)^\perp) \quad \boldsymbol{x}^\top \mathbf{B}_k \boldsymbol{x} \geq \frac{1}{\nu} \|\boldsymbol{x}\|^2. \tag{7.17c}$$

Proof. Let $k \in \mathbb{N}$.

Matrix $\mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k$ is symmetric. Using Loewner's order property and Assumption 7.2, we almost surely have

$$\eta \mathbf{D}_k^\top \mathbf{D}_k \preceq \mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k \preceq \nu \mathbf{D}_k^\top \mathbf{D}_k. \tag{7.18}$$

Assumption 7.3(i) ensures that $\mathbf{D}_k$ is an injective operator. It follows that $\mathbf{D}_k^\top \mathbf{D}_k$ is a symmetric definite positive matrix and according to (7.18) and $\eta > 0$, so is $\mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k$. This ensures that $\mathbf{B}_k$, as defined in (7.13), exists. Then, (7.17a) directly comes from (7.10) and (7.12).

Moreover, since the three terms in (7.18) are invertible matrices, we have:

$$\frac{1}{\nu} (\mathbf{D}_k^\top \mathbf{D}_k)^{-1} \preceq (\mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k)^{-1} \preceq \frac{1}{\eta} (\mathbf{D}_k^\top \mathbf{D}_k)^{-1}, \tag{7.19}$$

so that

$$\frac{1}{\nu}\mathbf{D}_k(\mathbf{D}_k^\top\mathbf{D}_k)^{-1}\mathbf{D}_k^\top \preceq \mathbf{B}_k \preceq \frac{1}{\eta}\mathbf{D}_k(\mathbf{D}_k^\top\mathbf{D}_k)^{-1}\mathbf{D}_k^\top. \tag{7.20}$$

Let us denote:

$$\mathbf{P}_k = \mathbf{D}_k(\mathbf{D}_k^\top\mathbf{D}_k)^{-1}\mathbf{D}_k^\top. \tag{7.21}$$

$\mathbf{P}_k \in \mathbb{R}^{N\times N}$ is an orthogonal projection operator since it is symmetric and verifies $\mathbf{P}_k^2 = \mathbf{P}_k$. It follows that:

$$\boldsymbol{O}_N \preceq \mathbf{P}_k \preceq \boldsymbol{I}_N. \tag{7.22}$$

(7.17b) is then directly obtained by replacing (7.22) in (7.20).

As an orthogonal projection matrix, $\mathbf{P}_k$ satisfies

$$(\forall \boldsymbol{x} \in \ker(\mathbf{P}_k)^\perp) \quad \mathbf{P}_k\boldsymbol{x} = \boldsymbol{x}. \tag{7.23}$$

Combining (7.23) with the left inequality of (7.20) yields:

$$(\forall \boldsymbol{x} \in \ker(\mathbf{P}_k)^\perp) \quad \boldsymbol{x}^\top\mathbf{B}_k\boldsymbol{x} \geq \frac{1}{\nu}\boldsymbol{x}^\top\mathbf{P}_k\boldsymbol{x} = \frac{1}{\nu}\|\boldsymbol{x}\|^2. \tag{7.24}$$

There remains to prove the relation $\ker(\mathbf{D}_k^\top) = \ker(\mathbf{P}_k)$. Inclusion $\ker(\mathbf{D}_k^\top) \subset \ker(\mathbf{P}_k)$ is straightforward. Since $\boldsymbol{x} \in \ker(\mathbf{P}_k)$, from the expression of $\mathbf{P}_k$ and left multiplication by $\boldsymbol{x}^\top$, we have

$$\boldsymbol{x}^\top\mathbf{D}_k(\mathbf{D}_k^\top\mathbf{D}_k)^{-1}\mathbf{D}_k^\top\boldsymbol{x} = 0. \tag{7.25}$$

Since $\mathbf{D}_k^\top\mathbf{D}_k$ is definite positive matrix, its inverse is too, so that $\mathbf{D}_k^\top\boldsymbol{x} = \mathbf{0}$, i.e. $\boldsymbol{x} \in \ker(\mathbf{D}_k^\top)$ which concludes the proof of (7.17c).

□

Relation (7.17c) brings light into our interpretation of Assumption 7.3(ii) as a descent condition. Indeed, taking $\boldsymbol{x} = -\mathbf{g}_k$ in (7.17c) leads to the gradient-related inequality [19] considered for instance in the analysis of [58, 59]. Relation (7.17c) will actually play a key role in the asymptotical analysis of Section 7.4.

7.3.5 Two additional technical lemmas

The next lemma is essential as it guarantees the integrability of all the probabilistic quantities we will manipulate in our convergence analysis. It especially validates the use of the conditional expectation operator and of its associate properties in every situation encountered in our proofs.

Lemma 7.2. *Under Assumptions 7.1, 7.2, 7.3(i) and 7.4(ii), for every $k \in \mathbb{N}$, $\mathbf{x}_k$, $\nabla F(\mathbf{x}_k)$, $\mathbf{e}_k$ and $\mathbf{g}_k$ are square-integrable random vectors of $\mathbb{R}^N$. Moreover, $F(\mathbf{x}_k)$ is an integrable random variable of $\mathbb{R}$.*

Proof. First, according to Assumption 7.1, F is a differentiable and coercive function on $\mathbb{R}^N$, which ensures the existence of a global minimizer $\boldsymbol{x}^*$ satisfying $\nabla F(\boldsymbol{x}^*) = \mathbf{0}_N$. Let us denote by F^* the minimal value of F on $\mathbb{R}^N$, i.e. $F^* = F(\boldsymbol{x}^*)$.

We start by proving the desired property for sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$. We here proceed by induction.

The case $k = 0$ is straightforward as $\mathbf{x}_0$ is a deterministic variable.

Assume that $\mathbf{x}_k$ is square-integrable for a given $k \in \mathbb{N}$. Then almost surely, and using Lemma 7.1,

$$\begin{aligned}\|\mathbf{x}_{k+1}\|^2 &= \|\mathbf{x}_k - \gamma_k \mathbf{B}_k \mathbf{g}_k\|^2, \\ &\leq 2\|\mathbf{x}_k\|^2 + 2\gamma_k^2 \|\mathbf{B}_k \mathbf{g}_k\|^2, \\ &\leq 2\|\mathbf{x}_k\|^2 + 2\frac{\gamma_k^2}{\eta^2}\|\mathbf{g}_k\|^2. \end{aligned} \tag{7.26}$$

with

$$\begin{aligned}\|\mathbf{g}_k\|^2 &= \|\nabla F(\mathbf{x}_k) + \mathbf{e}_k\|^2, \\ &\leq 2\|\nabla F(\mathbf{x}_k)\|^2 + 2\|\mathbf{e}_k\|^2. \end{aligned} \tag{7.27}$$

Hereabove, the positivity of all the manipulated random variables makes possible to take the conditional expectations. Since $\nabla F(\mathbf{x}_k)$ is $\mathcal{F}_k$-measurable, the next inequalities follow by using Assumptions 7.1 and 7.3(i), almost surely

$$\begin{aligned}\mathbb{E}(\|\mathbf{g}_k\|^2|\mathcal{F}_k) &= \mathbb{E}(\|\nabla F(\mathbf{x}_k) + \mathbf{e}_k\|^2|\mathcal{F}_k), \\ &\leq 2\ \mathbb{E}(\|\nabla F(\mathbf{x}_k)\|^2|\mathcal{F}_k) + 2\ \mathbb{E}(\|\mathbf{e}_k\|^2|\mathcal{F}_k), \\ &= 2\|\nabla F(\mathbf{x}_k)\|^2 + 2\ \mathbb{E}(\|\mathbf{e}_k\|^2|\mathcal{F}_k), \\ &\leq 2(1 + C^2)\|\nabla F(\mathbf{x}_k)\|^2, \\ &\leq 2\beta^2(1 + C^2)\|\mathbf{x}_k - \boldsymbol{x}^*\|^2, \\ &\leq 4\beta^2(1 + C^2)(\|\mathbf{x}_k\|^2 + \|\boldsymbol{x}^*\|^2). \end{aligned} \tag{7.28}$$

Taking the expectations yields

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}_k\|^2\right] &= \mathbb{E}\left[\mathbb{E}(\|\mathbf{g}_k\|^2|\mathcal{F}_k)\right], \\ &\leq 4(1 + C^2)\beta^2(\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + \|\boldsymbol{x}^*\|^2). \end{aligned} \tag{7.29}$$

By the induction hypothesis, we have $\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] < +\infty$, so that using (7.26)-(7.29)

$$\mathbb{E}[\|\mathbf{x}_{k+1}\|^2] \leq 2\ \mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + 8\beta^2\frac{\gamma_k^2}{\eta^2}(1 + C^2)\left(\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + \|\boldsymbol{x}^*\|^2\right) < +\infty, \tag{7.30}$$

which concludes this part of the proof.

We now focus on $\mathbf{g}_k$. The developments above shown that $\mathbb{E}\left[\|\mathbf{g}_k\|^2\right]$ is upper-bounded by a positive affine function of $\mathbb{E}\left[\|\mathbf{x}_k\|^2\right]$, itself being strictly lower than $+\infty$. Consequently, $\mathbb{E}\left[\|\mathbf{g}_k\|^2\right] < +\infty$.

Regarding $\nabla F(\mathbf{x}_k)$, we almost surely have

$$\begin{aligned}\|\nabla F(\mathbf{x}_k)\|^2 &\leq \beta^2\|\mathbf{x}_k - \boldsymbol{x}^*\|^2, \\ &\leq 2\beta^2(\|\mathbf{x}_k\|^2 + \|\boldsymbol{x}^*\|^2). \end{aligned} \tag{7.31}$$

The right member in the above equation is integrable, and so is the same for $\|\nabla F(\mathbf{x}_k)\|^2$.

The integrability of $\|\mathbf{e}_k\|^2$ arises directly from Assumption 7.4(ii), passing directly to the expectation.

The descent lemma applied to F, which is a β-Lipschitz differentiable function of $\mathbb{R}^N$ according to Assumption 7.1, leads to

$$\begin{aligned} F(\mathbf{x}_k) - F^* &\leq \frac{\beta}{2}\|\mathbf{x}_k - \boldsymbol{x}^*\|^2, \\ &\leq \frac{\beta}{2}(\|\mathbf{x}_k\|^2 + \|\boldsymbol{x}^*\|^2). \end{aligned} \tag{7.32}$$

The integrability of the right member of the above inequality yields the integrability of $F(\mathbf{x}_k)$.

□

We end this section with one last technical result which provides a rational for the expression of the bound $C_{\max}$ introduced in Assumption 7.4.

Lemma 7.3. *For every $C \in (0, C_{\max})$, there exists $\rho_0 > 0$ such that P_{ρ_0} is strictly negative on $[0, C]$ where for all $\rho > 0$, P_ρ refers to the polynomial*

$$P_\rho(X) = \left(1 + \frac{\nu\rho}{2\eta}\right)\frac{X^2}{\eta} + \frac{X}{\eta} + \left(\frac{\nu\rho}{2\eta^2} - \frac{1}{\nu}\right). \tag{7.33}$$

Proof. For all $\rho > 0$, P_ρ is a second order polynomial whose discriminant Δ_ρ is

$$\Delta_\rho = \frac{1}{\eta^2} + \frac{4}{\eta}\left(1 + \frac{\nu\rho}{2\eta}\right)\left(\frac{1}{\nu} - \frac{\nu\rho}{2\eta^2}\right). \tag{7.34}$$

Taking $\rho \in \left(0, 2\left(\eta/\nu\right)^2\right)$, it follows that Δ_ρ is strictly positive. Thus, P_ρ admits two distinct roots

$$w_{\rho,1} = -\frac{\eta^2\sqrt{\Delta_\rho} + \eta}{\nu\rho + 2\eta} < 0, \quad \text{and} \quad w_{\rho,2} = \frac{\eta^2\sqrt{\Delta_\rho} - \eta}{\nu\rho + 2\eta}. \tag{7.35}$$

Taking the limit for vanishing ρ yields:

$$\begin{aligned} \lim_{\rho \to 0^+} w_{\rho,2} &= \frac{\eta\sqrt{\frac{1}{\eta^2} + \frac{4}{\eta\nu}} - 1}{2}, \\ &= \frac{1}{2}\left(\sqrt{1 + \frac{4\eta}{\nu}} - 1\right), \\ &= C_{\max}. \end{aligned} \tag{7.36}$$

Using $C < C_{\max}$ and (7.36) ensures the existence of $\rho_0 \in \left(0, 2\left(\eta/\nu\right)^2\right)$ such that $w_{\rho_0,2} > C$. Moreover, the second degree coefficient of P_{ρ_0} is strictly positive, so that P_{ρ_0} is strictly negative on $(w_{\rho_0,1}, w_{\rho_0,2}) \supset [0, C]$ which completes the proof.

□

7.4 Asymptotical Analysis of SABRINA

7.4.1 Stochastic majoration of $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$

Proposition 7.1. *Under Assumptions 7.1-7.4, the following majoration holds almost surely:*

$$(\forall k \in \mathbb{N}) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1})|\mathcal{F}_k\right] \leq F(\mathbf{x}_k) + \gamma_k \|\nabla F(\mathbf{x}_k)\|^2 P_{\gamma_k}(C), \tag{7.37}$$

where P_{γ_k} is the polynomial quantity defined in Lemma 7.3.

Proof. Let $k \in \mathbb{N}$. We start by using the majoration property (7.2)-(7.3) of $h(., \mathbf{x}_k)$ on F at $\mathbf{x}_{k+1}$

$$\begin{aligned} F(\mathbf{x}_{k+1}) &\leq F(\mathbf{x}_k) + \nabla F(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{1}{2}\|\boldsymbol{x}_{k+1} - \mathbf{x}_k\|^2_{\mathbf{A}_k}, \\ &\leq F(\mathbf{x}_k) + \nabla F(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{\nu}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \quad \text{a.s.}, \end{aligned} \tag{7.38}$$

where (7.38) is a direct consequence of Assumption 7.2.

Using scheme (7.12) and the definition (7.8), inequality (7.38) can be written:

$$\begin{aligned} F(\mathbf{x}_{k+1}) &\leq F(\mathbf{x}_k) - \gamma_k \nabla F(\mathbf{x}_k)^\top \mathbf{B}_k \mathbf{g}_k + \frac{\nu}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2, \\ &= F(\mathbf{x}_k) - \gamma_k\ \mathbf{g}_k^\top \mathbf{B}_k \mathbf{g}_k + \gamma_k\ \mathbf{e}_k^\top \mathbf{B}_k \mathbf{g}_k + \frac{\nu}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2, \\ &= F(\mathbf{x}_k) - \gamma_k\ \mathbf{g}_k^\top \mathbf{B}_k \mathbf{g}_k + \gamma_k\ \mathbf{e}_k^\top \mathbf{B}_k \nabla F(\mathbf{x}_k) + \gamma_k\ \mathbf{e}_k^\top \mathbf{B}_k \mathbf{e}_k \\ &\quad + \frac{\nu}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \quad \text{a.s.}. \end{aligned} \tag{7.39}$$

On the one hand, Assumption 7.3(ii) guarantees that $\mathbf{g}_k \in \ker(\mathbf{D}_k^\top)^\perp$ almost surely. Hence, the left inequality (7.17c) of Lemma 7.1 yields

$$\mathbf{g}_k^\top \mathbf{B}_k \mathbf{g}_k \geq \frac{1}{\nu}\|\mathbf{g}_k\|^2 \quad \text{a.s.}. \tag{7.40}$$

On the other hand, the use of Cauchy-Schwarz inequality and relation (7.17b) from Lemma 7.1 gives

$$\mathbf{e}_k^\top \mathbf{B}_k \nabla F(\mathbf{x}_k) \leq \frac{1}{\eta}\|\nabla F(\mathbf{x}_k)\|\ \|\mathbf{e}_k\| \quad \text{a.s.}. \tag{7.41}$$

Moreover, (7.17b) also leads to:

$$\mathbf{e}_k^\top \mathbf{B}_k \mathbf{e}_k \leq \frac{1}{\eta}\|\mathbf{e}_k\|^2 \quad \text{a.s.}. \tag{7.42}$$

And, again as a consequence of (7.17b),

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 &= \gamma_k^2 \|\mathbf{B}_k \mathbf{g}_k\|^2, \\ &\leq \frac{\gamma_k^2}{\eta^2}\|\mathbf{g}_k\|^2 \quad \text{a.s.}. \end{aligned} \tag{7.43}$$

Plugging (7.40)-(7.43) into (7.39) leads to:

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \frac{\gamma_k}{\nu}\|\mathbf{g}_k\|^2 + \frac{\gamma_k}{\eta}\|\nabla F(\mathbf{x}_k)\|\ \|\mathbf{e}_k\| + \frac{\gamma_k}{\eta}\|\mathbf{e}_k\|^2 + \frac{\nu\gamma_k^2}{2\eta^2}\|\mathbf{g}_k\|^2 \quad \text{a.s.}. \tag{7.44}$$

Thanks to Lemma 7.2, we can take the conditional expectation in (7.44) and use the fact that it is a linear operator. Moreover, accounting for $\mathcal{F}_k$-measurability of $F(\mathbf{x}_k)$ and $\nabla F(\mathbf{x}_k)$, we obtain

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{x}_{k+1})|\mathcal{F}_k\right] \leq F(\mathbf{x}_k) \; &- \frac{\gamma_k}{\nu}\,\mathbb{E}\left[\|\mathbf{g}_k\|^2|\mathcal{F}_k\right] + \frac{\gamma_k}{\eta}\,\|\nabla F(\mathbf{x}_k)\|\,\mathbb{E}\left[\|\mathbf{e}_k\|\;|\mathcal{F}_k\right] \\ &+ \frac{\gamma_k}{\eta}\,\mathbb{E}\left[\|\mathbf{e}_k\|^2|\mathcal{F}_k\right] + \frac{\nu\gamma_k^2}{2\eta^2}\,\mathbb{E}\left[\|\mathbf{g}_k\|^2|\mathcal{F}_k\right] \quad \text{a.s..}\end{aligned} \tag{7.45}$$

The end of the proof aims at finding an upper bound of the last four terms in (7.45), depending only on $\nabla F(\mathbf{x}_k)$.

First, Definition (7.8) and the parallelogram identity give

$$\mathbb{E}\left[\|\mathbf{g}_k\|^2|\mathcal{F}_k\right] = \|\nabla F(\mathbf{x}_k)\|^2 + 2\,\mathbb{E}\left[\nabla F(\mathbf{x}_k)^\top \mathbf{e}_k|\mathcal{F}_k\right] + \mathbb{E}\left[\|\mathbf{e}_k\|^2|\mathcal{F}_k\right] \quad \text{a.s..} \tag{7.46}$$

Since $\nabla F(\boldsymbol{x}_k)$ is $\mathcal{F}_k$-measurable, and using Assumption 7.4(ii), we have

$$\begin{aligned}\mathbb{E}\left[\nabla F(\mathbf{x}_k)^\top \mathbf{e}_k|\mathcal{F}_k\right] &= \nabla F(\mathbf{x}_k)^\top \mathbb{E}\left[\mathbf{e}_k|\mathcal{F}_k\right], \\ &= 0 \quad \text{a.s.},\end{aligned} \tag{7.47}$$

which leads to the conditional equality

$$\mathbb{E}\left[\|\mathbf{g}_k\|^2|\mathcal{F}_k\right] = \|\nabla F(\mathbf{x}_k)\|^2 + \mathbb{E}\left[\|\mathbf{e}_k\|^2|\mathcal{F}_k\right] \quad \text{a.s..} \tag{7.48}$$

Using Assumption 7.4(ii) we then deduce the following bounds for $\|\mathbf{g}_k\|^2$

$$\|\nabla F(\mathbf{x}_k)\|^2 \leq \; \mathbb{E}\left[\|\mathbf{g}_k\|^2|\mathcal{F}_k\right] \leq \left(1 + C^2\right)\|\nabla F(\mathbf{x}_k)\|^2 \quad \text{a.s..} \tag{7.49}$$

Second, the following stochastic majoration of $\mathbb{E}\left[\|\boldsymbol{\epsilon}_k\|\;|\mathcal{F}_k\right]$ is obtained by Jensen's inequality and Equation (7.16)

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{e}_k\|\;|\mathcal{F}_k\right] &\leq \sqrt{\mathbb{E}\left[\|\mathbf{e}_k\|^2|\mathcal{F}_k\right]}, \\ &\leq C\|\nabla F(\mathbf{x}_k)\| \quad \text{a.s.},\end{aligned} \tag{7.50}$$

where (7.50) arises from Assumption 7.4(ii).

Finally, Inequalities (7.16), (7.49), (7.50) combined with (7.45) give the desired result

$$\mathbb{E}\left[F(\mathbf{x}_{k+1})|\mathcal{F}_k\right] \leq F(\mathbf{x}_k) + \gamma_k\|\nabla F(\mathbf{x}_k)\|^2\left[\left(1 + \frac{\nu\gamma_k}{2\eta}\right)\frac{C^2}{\eta} + \frac{C}{\eta} + \left(\frac{\nu\gamma_k}{2\eta^2} - \frac{1}{\nu}\right)\right] \quad \text{a.s..} \tag{7.51}$$

□

Proposition 7.2. *Under Assumptions 7.1-7.5, for every $\rho > 0$, there exists k_ρ such that*

$$(\forall k \geq k_\rho) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1})|\mathcal{F}_k\right] \leq F(\mathbf{x}_k) + \gamma_k\|\nabla F(\mathbf{x}_k)\|^2 P_\rho(C) \quad \text{a.s..} \tag{7.52}$$

Proof. By Assumption 7.5, $\gamma_k \underset{k\to+\infty}{\longrightarrow} 0$, which ensures the existence of k_ρ such that $\gamma_k \leq \rho$ for all $k \geq k_\rho$. Thus,

$$\begin{aligned} P_{\gamma_k}(C) &= \left[\left(1+\frac{\nu\gamma_k}{2\eta}\right)\frac{C^2}{\eta}+\frac{C}{\eta}+\left(\frac{\nu\gamma_k}{2\eta^2}-\frac{1}{\nu}\right)\right], \\ &\leq \left[\left(1+\frac{\nu\rho}{2\eta}\right)\frac{C^2}{\eta}+\frac{C}{\eta}+\left(\frac{\nu\rho}{2\eta^2}-\frac{1}{\nu}\right)\right] = P_\rho(C). \end{aligned} \tag{7.53}$$

Inequality (7.52) directly follows from (7.37) of Proposition 7.1.

□

7.4.2 General convergence theorem

We start with the following theorem which gives a general result for SABRINA without any convexity hypothesis:

Theorem 7.1. *Under Assumptions 7.1-7.5, sequence $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ converges a.s. to an almost surely finite random variable. Moreover, $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is such that*

$$\sum_{k=0}^{+\infty} \|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 < +\infty \quad \text{a.s.}, \tag{7.54a}$$

$$\liminf_{k\to+\infty} \|\nabla F(\mathbf{x}_k)\| = 0 \quad \text{a.s.}. \tag{7.54b}$$

Proof. From Lemma 7.3, there exists $\rho_0 > 0$ for which P_{ρ_0} is strictly negative on $[0, C]$. Applying Proposition 7.2 with $\rho = \rho_0$, yields the existence of k_{ρ_0} such that

$$(\forall k \geq k_{\rho_0}) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1})|\mathcal{F}_k\right] \leq F(\boldsymbol{x}_k) + \gamma_k\|\nabla F(\mathbf{x}_k)\|^2 P_{\rho_0}(C) \quad \text{a.s.}. \tag{7.55}$$

Subtracting F^*, the minimal value of F on each side of (7.55) yields

$$(\forall k \geq k_{\rho_0}) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1})-F^*|\mathcal{F}_k\right] \leq \left[F(\mathbf{x}_k)-F^*\right] + \gamma_k\|\nabla F(\mathbf{x}_k)\|^2 P_{\rho_0}(C) \quad \text{a.s.}. \tag{7.56}$$

All random variables involved in (7.56) are positive and integrable. Moreover, we have $P_{\rho_0}(C) < 0$ (since P_{ρ_0} is strictly negative on $[0, C]$). Thus, we can invoke Robbins-Siegmund's lemma [202]. The *a.s.* convergence of $(F(\mathbf{x}_k) - F^*)_{k\in\mathbb{N}}$ to an *a.s.* finite random variable is guaranteed, and so it is for $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$. Moreover again from Robbins-Siegmund's lemma, we have the following property

$$\sum_{k=0}^{+\infty} \gamma_k\|\nabla F(\mathbf{x}_k)\|^2 < +\infty \quad \text{a.s.}. \tag{7.57}$$

First, using (7.12), (7.17b) and then (7.49), yields

$$\begin{aligned} \sum_{k=0}^{+\infty} \mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2|\mathcal{F}_k\right] &\leq \sum_{k=0}^{+\infty} \frac{\gamma_k^2}{\eta^2}\, E[\|\mathbf{g}_k\|^2|\mathcal{F}_k], \\ &\leq \frac{1+C^2}{\eta^2}\sum_{k=0}^{+\infty}\gamma_k^2\|\nabla F(\mathbf{x}_k)\|^2 \quad \text{a.s.}. \end{aligned} \tag{7.58}$$

By Assumption 7.5, $(\gamma_k)_{k\in\mathbb{N}}$ is positive and converges to 0. Thus, $\gamma_k^2\|\nabla F(\mathbf{x}_k)\|^2 \leq \gamma_k\|\nabla F(\mathbf{x}_k)\|^2$ from a certain range k. It follows that the right term in (7.58) is a finite random variable and, as a consequence, $\sum_{k=0}^{+\infty}\mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2|\mathcal{F}_k\right] < +\infty$. Positivity of sequence $\left(\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2\right)_{k\in\mathbb{N}}$ finally allows us to apply [111, Prop.1] which gives (7.54a).

Our proof of (7.54b) is similar to the one of Zoutendijk condition for gradient-based optimization methods [32], adapted to a stochastic framework. To do so, we stand on complementary set $\left\{\omega\in\Omega\,|\,\liminf_{k\to+\infty}\|\nabla F(\mathbf{x}_k(\omega))\|>0\right\}$ and prove that it is of zero probability.

For all $\omega\in\Omega$ such that $\liminf_{k\to+\infty}\|\nabla F(\mathbf{x}_k(\omega))\|>0$, following the definition of lim inf, there exists $\varepsilon(\omega)>0$ and a range $\mathbf{k}_0(\omega)\in\mathbb{N}$ for which for all $k\geq\mathbf{k}_0(\omega)$, $\|\nabla F(\mathbf{x}_k(\omega))\|\geq\varepsilon(\omega)$. Thus

$$(\forall k\geq\mathbf{k}_0(\omega))\quad \gamma_k\|\nabla F(\mathbf{x}_k(\omega))\|^2\geq\gamma_k\varepsilon(\omega)^2. \tag{7.59}$$

Summing (7.59) from $\mathbf{k}_0(\omega)$ to $+\infty$, and using Assumption 7.5, we deduce

$$\begin{aligned}\sum_{k=\mathbf{k}_0(\omega)}^{+\infty}\gamma_k\|\nabla F(\mathbf{x}_k(\omega))\|^2 &\geq \varepsilon(\omega)^2\sum_{k=\mathbf{k}_0(\omega)}^{+\infty}\gamma_k,\\ &= +\infty.\end{aligned} \tag{7.60}$$

This leads to inclusion

$$\left\{\omega\in\Omega\,|\,\liminf_{k\to+\infty}\|\nabla F(\mathbf{x}_k(\omega))\|>0\right\}\subset\left\{\omega\in\Omega\,|\,\sum_{k=0}^{+\infty}\gamma_k\|\nabla F(\mathbf{x}_k(\omega))\|^2=+\infty\right\}. \tag{7.61}$$

The term in the right side of (7.61) is a negligible set according to (7.57). As a consequence, the left side of (7.61) is also a negligible set and (7.54b) holds by taking the complement.

□

Result (7.54a) ensures that sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ has a finite length [59]. Although some recent works consider (7.54b) as a sufficient convergence criterion [111], its scope remains limited since it only holds for a given subsequence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$. In the following, we make use of topological arguments to derive useful corollaries of Theorem 7.1.

Corollary 7.1. *Under Assumptions 7.1-7.5, the set χ^∞ of accumulation points of sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is almost surely non empty, compact and connex. Moreover, it contains at least one element of* $\operatorname{zer}\nabla F$.

Proof. Since Theorem 7.1 holds *a.s.*, there exists a set $\Lambda\subset\Omega$ of probability one set where, for all $\omega\in\Lambda$,

$$\lim_{k\to+\infty}F(\mathbf{x}_k(\omega))<+\infty, \tag{7.62a}$$

$$\sum_{k=0}^{+\infty}\|\mathbf{x}_{k+1}(\omega)-\mathbf{x}_k(\omega)\|^2<+\infty \tag{7.62b}$$

$$\liminf_{k\to+\infty}\|\nabla F(\mathbf{x}_k(\omega))\|=0. \tag{7.62c}$$

Inequality (7.62a) implies that $(F(\mathbf{x}_k(\omega)))_{k\in\mathbb{N}}$ is a bounded sequence. The coercivity of F, in Assumption 7.1, ensures this same property for $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$. It follows that the set of cluster points $\chi^\infty(\omega)$ is non empty and bounded. Moreover, it is compact due to its closure (in finite dimension).

Moreover, (7.62b) leads to:

$$\mathbf{x}_{k+1}(\omega) - \mathbf{x}_k(\omega) \underset{k\to+\infty}{\longrightarrow} \mathbf{0}_N. \tag{7.63}$$

Equation (7.63), and the boundedness of $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ enables the use of Ostrowski's theorem [182, 26.1]) which directly gives the connexity of $\chi^\infty(w)$.

From the boundedness of $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$, and (7.62c), we deduce that there exists a convergent subsequence $\left(\mathbf{x}_{\varphi(k)}(\omega)\right)_{k\in\mathbb{N}}$ such that

$$\nabla F(\mathbf{x}_{\varphi(k)}(\omega)) \underset{k\to+\infty}{\longrightarrow} \mathbf{0}_N. \tag{7.64}$$

Let us denote by $\mathbf{x}^\infty(w)$ the limit point of $\left(\mathbf{x}_{\varphi(k)}(\omega)\right)_k$. By construction, $\mathbf{x}^\infty(\omega)$ belongs to $\chi^\infty(\omega)$. Since F is gradient-Lipschitz, by Assumption 7.1, its gradient is continuous and we finally obtain:

$$\nabla F(\mathbf{x}^\infty(\omega)) = \mathbf{0}_N. \tag{7.65}$$

□

Corollary 7.1 provides us an overview of the distribution formed by the accumulation points of sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$. In order to refine the convergence theorem, we must introduce extra assumptions on the level sets of function F (see 7.2.1 for useful notations). We then propose an result, when F is convex with isolated stationary points.

Corollary 7.2. *Under Assumptions 7.1-7.5, if F is convex with isolated stationary points i.e.,*

$$(\forall v \in \mathbb{R}) \quad \mathrm{Card}(\mathrm{lev}_{=v} F \cap \mathrm{zer}\, \nabla F) < +\infty, \tag{7.66}$$

then $(\mathbf{x}_k)_{k\in\mathbb{N}}$ converges almost surely to a (global) minimizer of F.

Proof. On the one hand, by Theorem 7.1 and Corollary 7.1, there exists a probability-one set $\Lambda \subset \Omega$ for which $\omega \in \Lambda$, $(F(\mathbf{x}_k))_k$ possesses a finite limit F_w^∞. Moreover, $\chi^\infty(\omega)$ is connex and possess a point $\widetilde{\mathbf{x}}(\omega)$ which also belongs to $\mathrm{zer}\, \nabla F$. On the other hand, the convexity of F induces that the set of its (global) minimizer maps with $\mathrm{zer}\, \nabla F$ (by Fermat's rule). It follows that $\widetilde{\mathbf{x}}(\omega)$ is a global minimizer of F, that we will denote $\mathbf{x}^*(\omega)$ (following notations from Sec. 7.2.1). By continuity of F,

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad F(\boldsymbol{x}) \geq F(\mathbf{x}^*(\omega)) = F_\omega^\infty. \tag{7.67}$$

Moreover, again by continuity of F, for all $\boldsymbol{x} \in \chi^\infty(\omega)$, $F(\boldsymbol{x}) = F_\omega^\infty = F(\mathbf{x}^*(\omega))$. Thanks to (7.67), we deduce that every $\boldsymbol{x} \in \chi^\infty(\omega)$ is a (global) minimizer of F and then one of its stationary point. This yields the inclusion:

$$\chi^\infty(\omega) \subset \left(\mathrm{lev}_{=F(\boldsymbol{x}^*(\omega))} F \cap \mathrm{zer}\, \nabla F\right). \tag{7.68}$$

Since all stationary points of F are assumed to be isolated, set $\mathrm{lev}_{=F(\mathbf{x}^*(\omega))} F \cap \mathrm{zer}\, \nabla F$ is finite, and so is set $\chi^\infty(\omega)$. Moreover, set $\chi^\infty(\omega)$ is non empty and connex, so we obtain $\chi^\infty(\omega) = \{\mathbf{x}^*(\omega)\}$ which completes the proof.

□

7.4.3 Convergence rate analysis

We provide here our second main theoretical result, regarding the convergence rate of SABRINA, in the case when F satisfies a strong convexity property.

Theorem 7.2. *If F is α-strongly convex (i.e., $F - \frac{\alpha}{2}\|\cdot\|^2$ is convex) and Lipschitz differentiable function on $\mathbb{R}^N$ then, under Assumptions 7.2-7.5, there exist $D_\alpha > 0$ and a sequence $(r_k)_{k\in\mathbb{N}}$ such that, for k sufficiently large,*

$$\ln\left(\mathbb{E}\left[F(\mathbf{x}_{k+1}) - F^*\right]\right) \leq r_k. \tag{7.69}$$

Moreover,

$$r_k \underset{k\to+\infty}{\sim} D_\alpha \times \left(-\sum_{i=0}^{k} \gamma_i\right). \tag{7.70}$$

Proof. First, let us notice that Assumption 7.1 holds, since F is supposed to be α-convex and Lipschitz differentiable. Since Assumptions 7.2-7.5 also hold, we can thus come back to (7.56) (from the proof of Theorem 7.1, and using the same notations), and take the expectation to obtain:

$$(\forall k \geq k_{\rho_0}) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1}) - F^*\right] \leq \mathbb{E}\left[F(\mathbf{x}_k) - F^*\right] + \gamma_k P_{\rho_0}(C)\, \mathbb{E}\left[\|\nabla F(\mathbf{x}_k)\|^2\right] \quad \text{a.s..} \tag{7.71}$$

Let us make use of [35, Eq. (4.12)] related to strongly convex functions, which reads:

$$(\forall k \in \mathbb{N}) \quad \|\nabla F(\mathbf{x}_k)\|^2 \geq 2\alpha(F(\mathbf{x}_k) - F^*). \tag{7.72}$$

Substituting (7.72) in (7.71) then leads to:

$$(\forall k \geq k_{\rho_0}) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1}) - F^*\right] \leq (1 + \widehat{\gamma}_k)\, \mathbb{E}\left[F(\mathbf{x}_k) - F^*\right], \tag{7.73}$$

with

$$\widehat{\gamma}_k = 2\alpha P_{\rho_0}(C)\gamma_k < 0 \quad (\text{since} \quad P_{\rho_0}(C) < 0). \tag{7.74}$$

Moreover, by Assumption 7.5, $(\gamma_k)_{k\in\mathbb{N}}$ converges to 0 so that there exists $k_1 > k_{\rho_0}$ such that:

$$(\forall k \geq k_1) \quad 1 + \widehat{\gamma}_k \in (0, 1). \tag{7.75}$$

Then, by induction, it follows that for all $k \geq k_1 + 1$,

$$\mathbb{E}\left[F(\mathbf{x}_k) - F^*\right] \leq \mathbb{E}\left[F\left(\mathbf{x}_{k_1}\right) - F^*\right] \prod_{i=k_1}^{k-1} \left(1 + \widehat{\gamma}_i\right). \tag{7.76}$$

Taking the logarithm in (7.76), by virtue of Condition (7.75), then yields:

$$\ln\left(\mathbb{E}\left[F(\mathbf{x}_k) - F^*\right]\right) \leq \sum_{i=k_1}^{k-1} \ln\left(1 + \widehat{\gamma}_i\right) + \ln\left(\mathbb{E}\left[F\left(\mathbf{x}_{k_1}\right) - F^*\right]\right). \tag{7.77}$$

The end of the proof consists in searching for an asymptotic equivalent of the right member of (7.77). Convergence of $(\gamma_k)_{k\in\mathbb{N}}$ to 0 (by Assumption 7.5) ensures:

$$\ln(1 + \widehat{\gamma}_k) \underset{k\to+\infty}{\sim} \widehat{\gamma}_k. \tag{7.78}$$

Sequences $(\ln(1+\widehat{\gamma}_k))_{k\geq k_1}, (\widehat{\gamma}_k)_{k\geq k_1}$ are both negative. Moreover, Assumption 7.5 yields:

$$\sum_{i=k_1}^{+\infty} \widehat{\gamma}_i = -\infty. \tag{7.79}$$

We can thus deduce:

$$\begin{aligned}\sum_{i=k_1}^{k-1} \ln(1+\widehat{\gamma}_i) &\underset{k\to+\infty}{\sim} \sum_{i=k_1}^{k-1} \widehat{\gamma}_i, \\ &= 2\alpha P_{\rho_0}(C) \sum_{i=k_1}^{k-1} \gamma_i. \end{aligned} \tag{7.80}$$

Since the series $\sum_{i=k_1}^{k-1} \ln(1+\widehat{\gamma}_i)$ diverges to $-\infty$, it follows that

$$\begin{aligned}\sum_{i=k_1}^{k-1} \ln\left(1+\widehat{\gamma}_i\right) + \ln\left(\mathbb{E}\left[F\left(\boldsymbol{x}_{k_1}\right) - F^*\right]\right) &\underset{k\to+\infty}{\sim} 2\alpha P_{\rho_0}(C) \sum_{i=k_1}^{k-1} \gamma_i, \\ &\underset{k\to+\infty}{\sim} 2\alpha P_{\rho_0}(C) \sum_{i=0}^{k} \gamma_i. \end{aligned} \tag{7.81}$$

Hereabove, (7.81) arises from the convergence of $(\gamma_k)_{k\in\mathbb{N}}$ to 0 (Assumption 7.5). The desired conclusion is reached by setting $D_\alpha = -2\alpha P_{\rho_0}(C)$.

□

Theorem 7.2 guarantees an ℓ_1 convergence to F^* for sequence $F(\mathbf{x}_k)_{k\in\mathbb{N}}$ generated by SABRINA. Although sequence $(r_k)_{k\in\mathbb{N}}$ share the same asymptotical behaviour than the stepsize sequence, it does not necessarily ensure the same situation for $(F(\mathbf{x}_k)_{k\in\mathbb{N}} - F^*)$ since passing to exponential does not preserve the speed of convergence. It should be emphasized that the choice of $(\gamma_k)_{k\in\mathbb{N}}$ that ensures convergence is rather permissive due to Assumption 7.5. One interesting practical choice thus consists in setting $(\gamma_k)_{k\in\mathbb{N}}$ as a sequence converging to zero as slow as allowed by Assumption 7.5. As a result, relation (7.70) ensures that the logarithmic expectation in (7.69) tends quickly to $-\infty$.

7.4.4 Link to existing works

Our "liminf" convergence criterion (7.54b) is probably the most encountered one in optimization [32, chapter 1.4] among those introduced in Theorem 7.1. Similar result is also obtained in [111, 105] considering a stochastic context. The aforementioned works focused on an method closed to ADAM [137], that has been quite notorious in the field of deep learning this last decade. To a certain extent, the scheme in [111, 105] can be interpreted as a specific case of ours without using MM metric (i.e., $\mathbf{A}_k \equiv \boldsymbol{I}_N$) and where subspace acceleration is replaced by momentum weights combining with a manually tuned stepsize. By including an MM approach in a non-convex situation, the MISO algorithm from [161] shares common features with the one we develop here. The asymptotical result from [161,

Prop.3.3] is also expressed as a "liminf" condition but, up to our knowledge, might appear harder to interpret than (7.54b). Our result (7.54a) is not as common in the litterature of stochastic optimization, as its counterpart (7.54b), probably since it is slightly less tractable. It shares structural similarities with the finite length condition stated in [59, Theo.3] studying the MM subspace algorithm without noisy gradient. When considering noisy gradient, we manage here to show (7.54a), which is weaker in the sense that the square norm summation does not ensure necessarily that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is a Cauchy sequence, and thus does not allow to easily conclude on its almost sure convergence.

More generally, Robbins-Siegmund's lemma [202] is a widely used tool to deduce asymptotical properties of stochastic approximation schemes [89, 105]. Our use of Ostrowski's theorem [182, Theo.26.1] and the connexity argument to obtain Corollary 7.1 is reminiscent from [63]. However, in contrast with the aforementioned work, we use the convexity hypothesis only at the very end in Corollary 7.2. The idea of using level-set as an alternative arises from [104].

The supervised learning context has highly promoted studies relative to speed estimation of stochastic algorithms and especially for gradient methods both in a convex [175, 104, 35] and more recently in a non-convex setting [96]. [42, 33] also focus on quasi-Newton approximation approaches and obtain an ℓ_1 behaviour. Their approach can actually be seen to a particular subspace choice within our method, although no MM metric/stepsize are taken into consideration in [42, 33].

7.5 Application to Binary Classification

As a first illustrative example, we focus on a supervised binary classification problem. We consider M feature vectors $(\boldsymbol{v}_m)_{1\leq m\leq M} \in \mathbb{R}^N$, with their associated labels $(y_m)_{1\leq m\leq M} \in \{-1,1\}$ as a training dataset. In a linear classification context, one possibility to estimate parameter model $\boldsymbol{x}^* \in \mathbb{R}^N$ consists in searching the best linear classifier through the minimization of the log-loss penalized empirical risk [39]:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad F(\boldsymbol{x}) = \frac{1}{M}\sum_{m=1}^{M} \log\left(1+\exp\left(-[\boldsymbol{H}\boldsymbol{x}]_m\right)\right) + \mu \sum_{n=1}^{N} \log\left(1+\frac{x_n^2}{\delta^2}\right). \tag{7.82}$$

Matrix $\boldsymbol{H} = \mathrm{Diag}\{(y_m)_{1\leq i\leq M}\}[\boldsymbol{v}_1,...,\boldsymbol{v}_M]^\top \in \mathbb{R}^{M\times N}$ involved in the so-called data-fidelity term, gathers the information brought by the training dataset. The second term in (7.82) is a regularization term weighted by $\mu > 0$, which aims at promoting the sparsity of the estimated model so as to limit overfitting issues. The retained regularization is a coercive, continuous but non-convex approximation of the ℓ_0 norm, which is at the core of re-weighted ℓ_1 schemes [44, 197]. Function (7.82) is Lipschitz differentiable on $\mathbb{R}^N$ and coercive. However, it is non convex due to the regularization term.

7.5.1 Majorization mapping and convergence guarantees

First, using the properties from [37, 63], we are able to exhibit the following majorization mapping $\boldsymbol{A}(\cdot)$ for the objective function (7.82) :

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad \boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{H}^\top \mathrm{Diag}\left\{(\vartheta([\boldsymbol{L}\boldsymbol{x}]_m)_{1\leq m\leq M}\right\}\boldsymbol{H} + \mu \mathrm{Diag}\left\{\left(\frac{2}{x_n^2+\delta^2}\right)_{1\leq n\leq N}\right\} + \tau \boldsymbol{I}_N, \quad (7.83)$$

with $\vartheta : u \mapsto \frac{1}{u}\left(\frac{1}{1+\exp(-u)} - \frac{1}{2}\right)$ extended by continuity in 0. Moreover, τ is a strictly positive constant ensuring the fulfilment of Assumption 7.2. For such choice of mapping, Assumption 7.2 holds with:

$$\eta = \tau, \quad \nu = \tau + \frac{1}{4M}|||\boldsymbol{H}|||^2 + 2\frac{\mu}{\delta^2}. \quad (7.84)$$

We propose to implement SABRINA by considering two choices for the subspace, namely $\boldsymbol{D}_k = \boldsymbol{I}_N$, and $\boldsymbol{D}_k = [-\boldsymbol{g}_k \mid \boldsymbol{x}_k - \boldsymbol{x}_{k-1}]$. Both satisfy Assumption 7.3 and respectively yield the so-called SABRINA-I and SABRINA-MG algorithms. If Assumptions 7.4 and 7.5 hold, sequences generated by these two algorithms verify Theorem 7.1 and Corollary 7.1. Otherwise stated, for suitable stepsize and noise perturbation settings, our theoretical analysis ensures an almost sure convergence to a stationary point of F for a subsequence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$. Function F is non convex and does not have finite level sets, so that the stronger convergence results established in our study cannot be applied.

7.5.2 Numerical settings

When using the SABRINA-I scheme, the majorization function minimization requires to invert an $N \times N$ system, which is performed using the linear solver from [221]. The gradient perturbation is simulated by applying a multiplicative noise following a uniform law centered in 0 on every component of the gradient at each iteration that is, for every $k \in \mathbb{N}$,

$$\mathbf{e}_k = C \times \mathrm{Diag}\{(\mathrm{u}_{n,k})_{1\leq n\leq N}\}\nabla F(\mathbf{x}_k), \quad (7.85)$$

where each entry of $\mathbf{u}_k = (\mathrm{u}_{n,k})_{1\leq n\leq N} \in \mathbb{R}^N$ is an independant realization of a uniform law between $[-1, 1]$. By construction, Condition (7.16) holds since, for every $k \in \mathbb{N}$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{e}_k\|^2|\mathcal{F}_k\right] &\leq C^2\mathbb{E}\left[\|\mathrm{Diag}\{(\mathrm{u}_{n,k})_{1\leq n\leq N}\}\mathbf{u}_k\|_\infty|\mathcal{F}_k\right]\|\nabla F(\mathbf{x}_k)\|^2, \\ &= C^2\|\nabla F(\mathbf{x}_k)\|^2. \end{aligned} \quad (7.86)$$

Equation (7.86) also guarantees the integrability of $\mathbf{e}_k$. Moreover, $\mathbf{u}_k$ is zero-mean so that Assumption 7.4(i) also holds. We set the decreasing step-size $\gamma_k = 1/(k+1)^{0.01}$, for $k \in \mathbb{N}$, thus satisfying Assumption 7.5. Performance of SABRINA are evaluated against those of state-of-the-art stochastic gradient-based schemes from the machine learning field, namely SGD [20], ADAGRAD [87] and RMSprop [120]. The parameter tuning for these methods (e.g., learning rate, momentum weight) was made empirically, following recommendations from [207], to obtain best possible practical convergence behaviours.

Datasets `rcv1` and `a8a` are extracted from LIBSVM library [52]. Table 7.1 lists properties of these datasets and the retained hyperparameters μ, δ and τ. The latter has been manually chosen to ensure

a satisfying compromise between a good conditioning of the majorization mapping (and then, a wide range of values for C, see Sec. 7.3.3) and a fast convergence rate.

Dataset	Train M	Test	Features N	$\|\|\|\boldsymbol{H}\|\|\|^2/(4M)$	μ	δ	τ	$C_{\max}$
`rcv1`	20242	677399	47236	5.5×10^{-3}	10^{-1}	1	1	0.54
`a8a`	9865	22696	122	1.6	10^{-2}	1	0.5	0.2

Table 7.1: Dataset properties and hyperparameter settings

7.5.3 Experimental results

In Fig. 7.1, we illustrate the efficiency of every competitor through the evolution of the gradient norm of their iterates along time for a Matlab 2020a code ran on a desktop computer equipped with an Intel Core i7 3.2 GHz pro and 16 GB RAM. For this figure, we set $C = 0.95 \times C_{\max}$, so as to meet the conditions imposed by Assumption 7.4(ii) and then convergence of SABRINA is ensured in the sense of Theorem 7.1 and Corollary 7.1. It is noticeable that both SABRINA variants reach the best performance when compared to their competitors. Moreover, for both datasets, the interest of subspace acceleration is visible, as SABRINA-MG reaches faster convergence than SABRINA-I. Finally, let us emphasize that SABRINA implementation does not impose any tedious manual learning/momentum rate tuning, as it was the case for the other methods. Table 7.2 lists classification scores obtain by SABRINA-MG at convergence for both datasets.

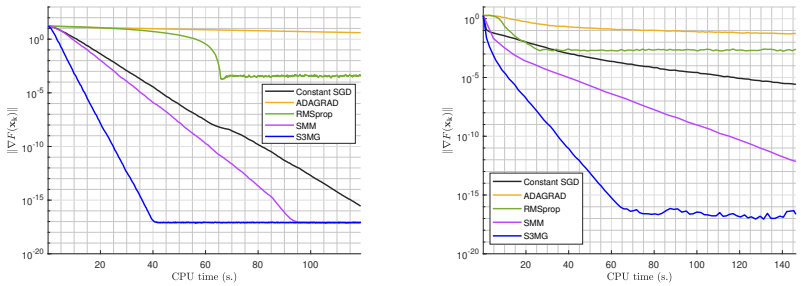


Figure 7.1: Evolution of the gradient norm along time for various algorithms, on dataset `rcv1` (left) and `a8a` (right). Noise amplitude $C = 0.95 \times C_{\max}$.

Dataset	Accuracy	AUC	Precision	Recall
`rcv1`	$9,2 \times 10^{-1}$	$9,7 \times 10^{-1}$	$9,3 \times 10^{-1}$	$9,1 \times 10^{-1}$
`a8a`	$8,4 \times 10^{-1}$	$8,9 \times 10^{-1}$	$7,5 \times 10^{-1}$	$5,2 \times 10^{-1}$

Table 7.2: Classification scores after running SABRINA-MG for 60 s.

Fig. 7.2 illustrates the evolution of the gradient norm along SABRINA-MG iterations for various levels of noise on the gradient term, when considering the `rcv1` example. Increasing the noise level

obviously slows down the convergence of the method. Moreover, one can see that SABRINA-MG starts showing some oscillating behaviour when $C \geq C_{\max}$. Considering an order of magnitude ten times higher than $C_{\max}$, one can observe a change of regime where the convergence of the algorithm seems compromised. Such phenomenons suggest that the bound $C_{\max}$ involved in Assumption 7.4 is consistent and not over pessimistic in this example for ensuring practical stability of the algorithm.

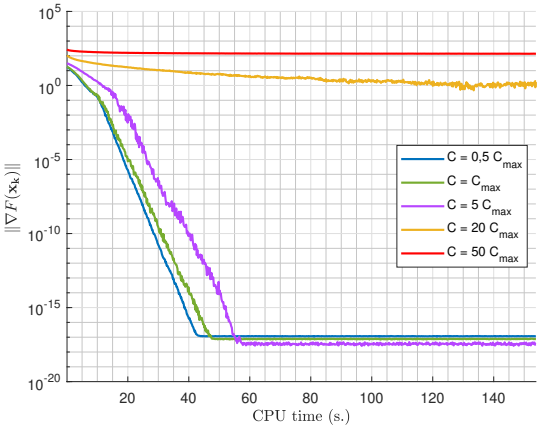


Figure 7.2: `rcv1`: Evolution of the gradient norm along time for various noise amplitudes affecting the gradient term in SABRINA-MG.

7.6 Application to Robust Blur Kernel Identification

We now consider a problem of robust blur kernel identification. The problem is reminiscent from the one studied in [63]. We seek for solving the inverse problem

$$\boldsymbol{y} = \boldsymbol{H}\overline{\boldsymbol{x}} + \boldsymbol{n}. \tag{7.87}$$

$\boldsymbol{y} \in \mathbb{R}^M$ corresponds to the vectorized version (in lexicographic order) of an original image $\boldsymbol{z} \in \mathbb{R}^M$, degraded by an unknown blur kernel $\overline{\boldsymbol{x}} \in \mathbb{R}^N$ to be determined, and an additive noise $\boldsymbol{n} \in \mathbb{R}^M$. The blur operation corresponds to a 2D discrete convolution (with circulant-padding assumption) between $\boldsymbol{z}$ and $\overline{\boldsymbol{x}}$, that is rewritten equivalently as the application of the linear Hankel-block operator $\boldsymbol{H} \in \mathbb{R}^{M\times N}$ (related to $\boldsymbol{z}$) on the kernel $\overline{\boldsymbol{x}}$. We consider a more challenging noise scenario than in [63], where outliers can arise in the observed data. Specifically, $\boldsymbol{n} \in \mathbb{R}^N$ is the realization of a Gaussian mixture noise with standard deviations $(\sigma_1, \sigma_2) > 0$ and mixing rate $\varrho \in]0, 1[$, where typically $\sigma_1 \ll \sigma_2$. An efficient strategy for solving (7.87) consists in minimizing a penalized criterion:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad F(\boldsymbol{x}) = L(\boldsymbol{x}) + R(\boldsymbol{x}), \tag{7.88}$$

where L plays the role of the data fidelity term, accounting for the mixture noise model, and R is a regularization function promoting desirable prior assumption on the sought $\boldsymbol{x}$.

Due to the presence of outliers in the noise, we opt for the Huber data fidelity term, well suited for robust inverse problem resolution

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad L(\boldsymbol{x}) = \sum_{m=1}^{M} \ell_m([\boldsymbol{H}\boldsymbol{x}]_m]), \tag{7.89}$$

with

$$(\forall m = 1, \ldots, M)(\forall t \in \mathbb{R}) \quad \ell_m(t) = \begin{cases} \frac{1}{2}(t - y_m)^2 & \text{if } |t - y_m| \leq p \\ p|t - y_m| - \frac{1}{2}p^2 & \text{otherwise.} \end{cases} \tag{7.90}$$

Moreover, we choose to promote smoothness of the restored kernel, by setting:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad R(\boldsymbol{x}) = \sum_{n=1}^{N} \psi(\|\boldsymbol{\Delta}_n \boldsymbol{x}\|). \tag{7.91}$$

Hereabove, for every $n \in \{1, \ldots, N\}$, $\boldsymbol{\Delta}_n \in \mathbb{R}^{2\times N}$ corresponds to the discrete vertical and horizontal gradient operators applied to the n-th pixel of the 2D reshaped kernel $\boldsymbol{x}$. Moreover, $\psi : u \mapsto \lambda\sqrt{1 + u^2/\kappa^2}$ is the hyperpolic penalty with smoothness parameter $\kappa > 0$. Function (7.91) can thus be viewed as a smoothed version of the classical total-variation norm widely used in image processing. Parameter $\lambda > 0$ is a regularization parameter.

The resulting function (7.88) is convex and Lipschitz differentiable on $\mathbb{R}^N$. Moreover, according to [203, Proposition 2.5], F is coercive if and only if

$$\ker(\boldsymbol{H}) \cap \ker(\boldsymbol{\Delta}_1) \cap \ldots \cap \ker(\boldsymbol{\Delta}_N) = \{\boldsymbol{0}_N\}, \tag{7.92}$$

which actually holds for our practical choices for $\boldsymbol{H}$ and $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_N$ [63].

7.6.1 Majorization mappings and convergence guarantees

The Huber potential terms $(\ell_m)_{1\leq m\leq M}$ satisfy the assumptions from [58, Sec.III] so that we can build the following majorization mapping:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad \boldsymbol{A}_L(\boldsymbol{x}) = \boldsymbol{H}^\top \mathrm{Diag}\left(\zeta_m([\boldsymbol{H}\boldsymbol{x}]_m])\right) \boldsymbol{H}, \tag{7.93}$$

with

$$(\forall m = 1, \ldots, M)(\forall t \in \mathbb{R}) \quad \zeta_m(t) = \begin{cases} 1 & \text{if } |t - y_m| \leq p \\ \frac{p}{|t-y_m|} & \text{otherwise.} \end{cases} \tag{7.94}$$

Function ψ satisfies the properties of [58, Sec.III], allowing us to build a majorization matrix for penalization (7.91):

$$(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \boldsymbol{A}_R(\boldsymbol{x}) = \boldsymbol{\Delta}^\top \mathrm{Diag}(\boldsymbol{\rho}(\boldsymbol{x}))\boldsymbol{\Delta}, \tag{7.95}$$

with $\boldsymbol{\Delta} = [\boldsymbol{\Delta}_1^\top \mid \ldots \mid \boldsymbol{\Delta}_N^\top]^\top \in \mathbb{R}^{2N\times N}$. Moreover,

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad \boldsymbol{\rho}(\boldsymbol{x}) = \begin{bmatrix} \omega(\|\boldsymbol{\Delta}_1 \boldsymbol{x}\|) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ \vdots \\ \omega(\|\boldsymbol{\Delta}_N \boldsymbol{x}\|) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{bmatrix} \in \mathbb{R}^{2N}, \tag{7.96}$$

with $\omega : u \mapsto (1 + u^2/\kappa^2)^{-1/2}$. Studying the variations of function ω allows to deduce:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad \boldsymbol{O}_N \preceq \boldsymbol{A}_R(\boldsymbol{x}) \preceq \lambda \frac{|||\boldsymbol{\Delta}|||^2}{\kappa^2} \boldsymbol{I}_N. \tag{7.97}$$

In a nutshell, $\boldsymbol{A}_R + \boldsymbol{A}_L$ would constitute a valid majorization mapping for function F. However, it does not necessarily satisfy Assumption 7.2 since no strictly positive lower-bound is guaranteed for such mapping. We thus hereagain use the corrected mapping:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad \boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{A}_L(\boldsymbol{x}) + \boldsymbol{A}_R(\boldsymbol{x}) + \tau \boldsymbol{I}_N, \tag{7.98}$$

with $\tau > 0$. We can thus deduce from (7.97) and (7.93) that the mapping (7.98) satisfies Assumption 7.2 with:

$$\eta = \tau, \quad \nu = \tau + |||\boldsymbol{H}|||^2 + \lambda \frac{|||\boldsymbol{\Delta}|||^2}{\kappa^2}. \tag{7.99}$$

We focus on the minimization of (7.88) using the proposed SABRINA scheme for various choices of subspace matrices. We discard the choice $\mathbf{D}_k \equiv \boldsymbol{I}_N$, that appears to be not well suited with such large dimension problem. Instead, we focus on the so-called super-memory gradient subspace family [220], where:

$$(\forall k \in \mathbb{N}) \quad \mathbf{D}_k = [-\mathbf{g}_k \,|\, \mathbf{x}_k - \mathbf{x}_{k-1} \,|\, \ldots \,|\, \mathbf{x}_{k-M_k+1} - \mathbf{x}_{k-M_k}] \in \mathbb{R}^{N \times M_k}, \tag{7.100}$$

with the convention $\mathbf{x}_i = \mathbf{0}_N$ for $i < 0$, and $M_k \geq 1$ a memory size parameter. The resulting algorithms are denoted SABRINA-SMG-M_k. When $M_k = 1$, we retrieve the gradient direction $\mathbf{D}_k = -\mathbf{g}_k$, while for $M_k = 2$ we obtain the memory gradient subspace $\mathbf{D}_k = [-\mathbf{g}_k \,|\, \mathbf{x}_k - \mathbf{x}_{k-1}]$, so that SABRINA-SMG-2 identifies with SABRINA-MG considered in our previous experimental example. Subspace (7.100) satisfies Assumption 7.3 for any $M_k \geq 1$[1]. Thus, Theorem 7.1 and Corollary 7.1 hold under a moderate gradient noise (see Assumption 7.4). Assuming than F has isolated stationary points would yield the applicability of Corollary 7.2, which would guarantee the almost sure convergence of $(\mathbf{x}_k)_{k \in \mathbb{N}}$ to a global minimizer of F. Although it is not possible to show the fulfilment of this technical condition, we did not observe any convergence instability on the sequence $(\mathbf{x}_k)_{k \in \mathbb{N}}$.

7.6.2 Presentation of the data and settings

The original image $\boldsymbol{z}$ is the satellite image `SanDiego` of size $M = 1024 \times 1024$ pixels. The blur kernel is a non-uniform motion blur with size $N = 21 \times 21$. The noise parameters are $\sigma_1 = 5 \times 10^{-4}$, $\sigma_2 = 200\,\sigma_1$ and $\varrho = 0.1$, so that the signal to noise ratio of the observed image is 13.3 dB. The original image, its degraded version $\mathbf{y}$ and the blur kernel to reconstruct are displayed in Fig. 7.3.

[1]Here, it is assumed that, for every $k \in \mathbb{N}$, columns of $\mathbf{D}_k$ that would be co-linear to the first column $-\mathbf{g}_k$ are removed and M_k adjusted, so as to satisfy Assumption 7.3(i).

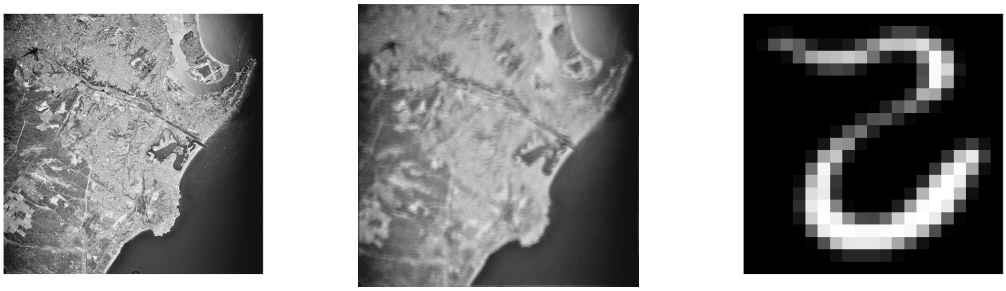


Figure 7.3: (Left) Original image $\boldsymbol{z}$; (Middle) Blurred and noisy image $\boldsymbol{y}$; (Right) Original blur kernel $\overline{\boldsymbol{x}}$.

The numerical experiments are performed on the same computer with the same software details as for the example of Section 7.5. We use the same uniform multiplicative noise (see Sec. 7.5.2) for the gradient perturbations in our proposed method, so as to satisfy Assumption 7.4. Once again we set $k = 1/(k+1)^{0.01}$ as the step-size for every $k \in \mathbb{N}$. Finally, the hyperparameters are tuned through gridsearch so as to minimize the relative mean square error (RMSE) on the kernel estimation, to $(p, \lambda, \kappa) = (1, 10, 10)$.

7.6.3 Calculation of $C_{\max}$

The ratio between bounds (η, ν) involved in (7.99) allows to compute the allowed tolerance on the gradient uncertainty, following Assumption 7.4(ii). However, in the particular problem of blur identification, $|||\boldsymbol{H}|||^2$ may be very large so that $\eta/\nu \ll 1$ and thus $C_{\max} \ll 1$. Typically, in our example, we obtain a theoretical $C_{\max}$ close to 8×10^{-9} which is very constraining in term of gradient noise. Actually, the difficulty lies in the over pessimistic lower bound $\eta = \tau$ in (7.99). Let us first point out that, according to (7.93),

$$(\forall k \in \mathbb{N}) \quad (\tau + \min_{1 \leq m \leq M} \zeta_m([\boldsymbol{H}\mathbf{x}_k])\, |||\boldsymbol{H}|||^2)\boldsymbol{I}_N \preceq \boldsymbol{A}_k. \tag{7.101}$$

We computed the actual values for $\min_{1 \leq m \leq M} \zeta_m([\boldsymbol{H}\mathbf{x}_k])$ along iterations, in our practical experiment, and observed that this quantity actually goes rapidly to 1 after few iterations. This leads us to consider

$$\widetilde{\eta} = \tau + |||\boldsymbol{H}|||^2 \tag{7.102}$$

as an empirical lower bound. Note that such bound gets valid as long as p is large enough, with respect to the absolute value of the entries of vector $\boldsymbol{H}\boldsymbol{x} - \boldsymbol{y}$. We denote:

$$\widetilde{C}_{\max} = \frac{1}{2}\left((1 + \frac{4\widetilde{\eta}}{\nu})^{\frac{1}{2}} - 1\right), \tag{7.103}$$

and express the gradient perturbation level C used in the experiment, as a function of $\widetilde{C}_{\max}$. Note that, in the present experiment, $\widetilde{C}_{\max} = 6.18 \times 10^{-1}$, which is closer to the best case bound mentioned in Section 7.3.3.

7.6.4 *Numerical results*

We first compare the performance of SABRINA with classical stochastic algorithms. We also include ADAM method [137], as it shows rather good performance in that example. The gradient perturbation is set to $C = 0.25 \times \widetilde{C}_{\max}$. The methods are compared in terms of RMSE between the current iterate and the sought kernel $\overline{\boldsymbol{x}}$.

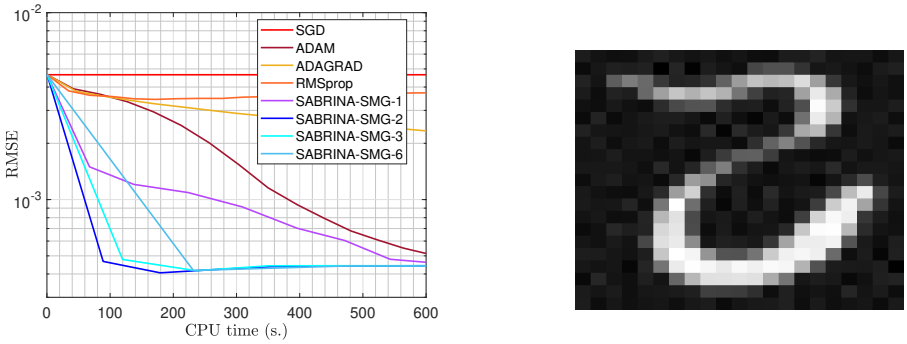


Figure 7.4: (Left) Evolution of the RMSE along time for various algorithms ; (Right) Estimated kernel using SABRINA-SMG-2, RMSE $= 4.4 \times 10^{-4}$. Noise amplitude $C = 0.25 \times \widetilde{C}_{\max}$, and starting point $\mathbf{x}_0 = \mathbf{0}_N$.

Fig. 7.4(left) shows that SABRINA-SMG-2 is the fastest of the algorithms to reach convergence. The other choices of memory size, for the super-memory gradient subspace, appear less competitive, which is in accordance with the observations from [58, 63]. The RMSE of the reconstructed kernel, displayed in Fig. 7.4(right), is equal to 4.4×10^{-4} for an estimated computational time of 600 s.

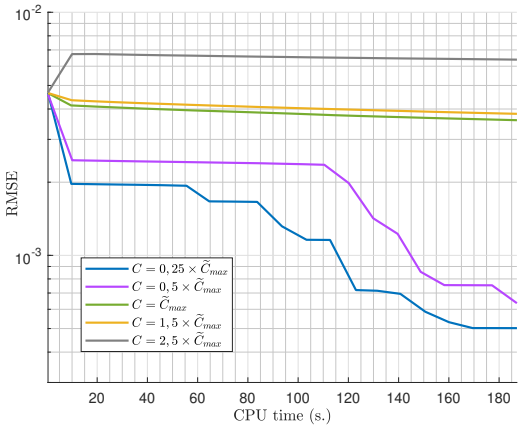


Figure 7.5: Evolution of the RMSE along time for various noise amplitudes affecting the gradient term in SABRINA-SMG-2.

For the setting of Fig. 7.4(left), the value of C actually exceeds the maximal numerical tolerance $C_{\max}$, but is chosen lower than $\widetilde{C}_{\max}$ which appears sufficient in practice to reach convergence for all

SABRINA variants tested here. In Fig. 7.5, we now present the evolution of the RMSE along time for SABRINA-SMG-2, when its gradient term is affected by various levels of noise C. One can notice that our corrected bound $\widetilde{C}_{\max}$ clearly maps with the delineation of two regimes for the convergence of SABRINA-SMG-2. As soon as C is sufficiently low compared to $\widetilde{C}_{\max}$, convergence is fast. On the contrary, a too high C seems to compromise the behaviour of the method, as expected and divergence can even be observed for large C.

7.7 Conclusion

The work in this chapter provides new insights into the stability of MM schemes suffering from stochastic noise perturbations on their gradient evaluation. New asymptotical results and convergence rate analysis are demonstrated under reasonably mild assumptions, and in the challenging scenario of a non necessarily convex cost function. Two numerical experiments in the fields of machine learning and image processing illustrate the high relevancy of the considered MM schemes compared to several classical competitors both regarding their speed of convergence and their robustness to noise. In particular, the experimental results emphasize the impressive performance of MM algorithm associated to a memory gradient subspace, already assessed in some previous works [58, 63]. It is remarkable to notice that, for such subspace choice, our contribution can be understood as providing novel theoretical guarantees on a stochastic non-linear conjugate gradient method with MM-based formula for stepsize and conjugacy parameters. One avenue for future work would be to extend our convergence rate analysis to a larger class of function by alleviating the strong convexity condition.

Chapter 8

A Kurdyka-Łojasiewicz property for stochastic optimization algorithms in a non-convex setting

Stochastic differentiable approximation schemes are widely used for solving high dimensional problems. Most of existing methods satisfy some desirable properties, including conditional descent inequalities [202, 112], and almost sure (a.s.) convergence guarantees on the objective function, or on the involved gradient [209, 111]. However, for non-convex objective functions, a.s. convergence of the iterates, i.e., the stochastic process, to a critical point is usually not guaranteed, and remains an important challenge. In this chapter, we develop a framework to bridge the gap between descent-type inequalities and a.s. convergence of the associated stochastic process. Leveraging a novel Kurdyka-Łojasiewicz property, we show convergence guarantees of stochastic processes under mild assumptions on the objective function. We also provide examples of stochastic algorithms benefiting from the proposed framework and derive a.s. convergence guarantees on the iterates.

This work relies on the article: E. Chouzenoux, J-B. Fest and A. Repetti. *A Kurdyka-Lojasiewicz property for stochastic optimization algorithms in a non-convex setting*, available at the following adress https://arxiv.org/abs/2302.06447.

This research work has been done in collaboration with Audrey Repetti (Heriot-Watt University, Edinburgh, UK).

Contents

8.1 Introduction

In this chapter, we aim at solving problem

$$\underset{\boldsymbol{x}\in\mathcal{H}}{\text{minimize}} \quad F(\boldsymbol{x}), \tag{8.1}$$

where $F : \mathcal{H} \to \mathbb{R}$ is a continuously differentiable function defined on a finite-dimensional real Hilbert space $\mathcal{H}$. We consider a generic stochastic process $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to solve (8.1) on a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$. Generating $(\mathbf{x}_k)_{k\in\mathbb{N}}$ typically results from a stochastic approximation scheme [201, 89]. Celebrated examples are often based on stochastic gradient descent (SGD) algorithms, where $(\mathbf{x}_k)_{k\in\mathbb{N}}$ arises from a gradient descent scheme with the gradient of F being replaced by stochastic approximations.

The objective of the work of this chapter is to develop a new framework, based on Kurdyka-Łojasiewicz (KL) theory [141, 25], to derive almost sure (a.s.) convergence guarantees of the stochastic process $(\mathbf{x}_k)_{k\in\mathbb{N}}$, when F is not convex.

When (8.1) is solved using a deterministic scheme, the main advantage of KL condition lies in its ability to promote interesting asymptotic behavior when F is non-necessarily convex. In this context, KL has been used to prove convergence of proximal point algorithms [9, 10], of simple splitting algorithms such as the forward-backward algorithm and its variants [10, 64, 27, 103, 66, 198, 30, 31], as well as other algorithms based on the majorization-minimization principle [59, 56, 51]. A natural question is to investigate the transfer of the proof techniques from the deterministic setting, to the stochastic setting, for asymptotic analysis including a.s. convergence of stochastic processes. Such an extension is quite challenging, mainly due to the dynamics of the functions involved in KL conditions, that cannot be controlled in a stochastic environment. Although KL inequality has already been invoked for L^1 or L^2 convergences of some stochastic schemes [106], to the best of our knowledge, no generalized KL inequality has been developed yet in this context.

Hence, no asymptotic analysis including a.s. convergence of the stochastic processes has been properly formalized and completed. In particular, the authors of [152] mention that a stochastic formulation of the KL inequality would enable to study the accumulation points of the process, without however providing any theoretical results.

In this chapter, we investigate the conditions that a stochastic process $(\mathbf{x}_k)_{k\in\mathbb{N}}$ minimizing F must satisfy to ensure its a.s. convergence to a critical point, under the KL inequality. To this aim, we design a new KL condition for differentiable functions, that can be used in a stochastic setting. Using this new framework, we furthermore derive conditions to ensure the convergence of processes generated by a few generic stochastic schemes based on SGD and proximal iterations. In particular, we show that our conditions enable to ensure the convergence of some state-of-the-art SGD algorithms [212, 71, 162], stochastic proximal-gradient algorithms (also known as forward-backward) [71, 107, 151], and stochastic proximal algorithms [210].

The remainder of the chapter is organized as follows. Our general stochastic framework is described in section 8.2. In this section, we introduce general assumptions, and give preliminary results, including a first convergence result in a simplified setting. In section 8.3 we provide a deterministic extension of the uniformized KL property introduced in [9, Lemma 6]. We then use this extension to design

a KL condition that can be used in a stochastic setting. In section 8.4, using the stochastic KL inequality introduced in section 8.3, we show the a.s. convergence of a family of stochastic processes $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to a critical point. Finally, in section 8.5 and section 8.6, we apply the results of section 8.4 to exhibit conditions for convergence of SGD and stochastic proximal-based algorithms, respectively, in a non-convex context.

8.1.0.0.1 Notation Let $(\mathcal{H}, \langle\cdot|\cdot\rangle)$ be a finite dimensional Hilbert space. $\|\cdot\|$ denotes the canonical norm associated with $\mathcal{H}$. For any subset $E \subset \mathcal{H}$, the distance function to this set is denoted by $\mathrm{dist}(\cdot, E) = \inf_{\boldsymbol{x}\in E} \|\cdot - \boldsymbol{x}\|$. Bold letters as $\boldsymbol{x}$ are used for deterministic vectors, while straight bold letters as $\mathbf{x}$ are used for stochastic vectors. Similarly, x is used for denoting a deterministic scalar variable, while straight x denotes a stochastic scalar variable. Upper-case letters are used for functions. The variable $F(\mathbf{x})$ denotes the stochastic value of function F evaluated at the stochastic variable $\mathbf{x}$. We work on the probability space $(\Omega, \mathcal{F}, P)$. We remind that a condition holds *almost surely* (a.s.) if it holds on a probability-one event of $\mathcal{F}$. We denote by $\mathbb{E}[\cdot]$ the expectation operator, and $\mathbb{E}[\cdot|\mathcal{G}]$ the conditional expectation operator regarding a generic sub sigma-algebra $\mathcal{G} \subset \mathcal{F}$.

Let ∇F be the gradient of F. Then we denote by $\operatorname{zer}\nabla F$ the set of zeros of ∇F, i.e., the set of critical points of F. Finally, we introduce χ^∞, the set of accumulation points of sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$. So χ^∞ is a random variable from Ω to the set of subsets of $\mathcal{H}$.

8.2 General assumptions and preliminary results

In this section, we introduce the probabilistic setting and assumptions used in the remainder of this chapter.

8.2.1 Assumptions

The following generic assumptions will guide us throughout the remainder of this chapter.

Assumption 8.1. *F is coercive and continuously differentiable.*

Assumption 8.2. *Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a stochastic process.*

(i) *The sequence $(F(\mathbf{x}_k)_{k\in\mathbb{N}})_{k\in\mathbb{N}}$ converges a.s. to a random variable F_∞. In addition, F_∞ is finite a.s.*

(ii) *$(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$ almost surely converges to $\mathbf{0}_N$.*

These assumptions are very common, and are satisfied by a wide range of stochastic algorithms.

The first assumption is standard in the field of differentiable optimization and notably ensures that F admits at least one global minimizer [19]. Assumption 8.2 is usually satisfied by classical schemes. Indeed, convergence of $(F(\mathbf{x}_k)_{k\in\mathbb{N}})_{k\in\mathbb{N}}$ and $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$ are often obtained easily in both deterministic [178], and stochastic [202, 89] settings. Note that Assumption 8.2(ii) necessitates the almost sure convergence of the gradient to zero. This condition is verified in particular by stochastic schemes with a conditional strong growth condition, see for instance [212, 150].

8.2.2 Preliminary results

We will now give some preliminary results, with the aim to link the different sets related to convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ and $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$.

Proposition 8.1. *Under Assumption 8.1 and Assumption 8.2, we have*

(i) $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *is a.s. bounded.*

(ii) χ^∞ *is a.s. non empty and compact.*

(iii) $\chi^\infty \subset \operatorname{zer}\nabla F$ *a.s.*

(iv) $\left(\operatorname{dist}(\mathbf{x}_k, \chi^\infty)\right)_{k\in\mathbb{N}}$ *converges a.s. to* 0.

Proof. According to Assumption 8.2, there exists a set $\Lambda \subset \Omega$ of probability one where, for all $\omega \in \Lambda$,

$$\lim_{k\to+\infty} F(\mathbf{x}_k(\omega)) < +\infty, \tag{8.2a}$$

$$\lim_{k\to+\infty} \|\nabla F(\mathbf{x}_k(\omega))\| = 0. \tag{8.2b}$$

Inequality (8.2a) implies that $(F(\mathbf{x}_k(\omega)))_{k\in\mathbb{N}}$ is a bounded sequence. According to Assumption 8.1, F being coercive, $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ is also bounded. It follows that the set of cluster points $\chi^\infty(\omega)$ is non empty, bounded and closed in $\mathcal{H}$, hence compact.

Let $\omega \in \Lambda$. We will now show that $\chi^\infty(\omega) \subset \operatorname{zer}\nabla F$. Let $\boldsymbol{x}_\infty \in \chi^\infty(\omega)$. There exists a subsequence $\left(\mathbf{x}_{\psi(k)}(\omega)\right)_{k\in\mathbb{N}}$ converging to $\boldsymbol{x}_\infty$. Moreover, from (8.2b), $(\nabla F(\mathbf{x}_{\psi(k)}(\omega)))_{k\in\mathbb{N}}$ converges to $\mathbf{0}_N$. Since, according to Assumption 8.1, the gradient of F is continuous, we thus obtain $\nabla F(\boldsymbol{x}_\infty) = \mathbf{0}_N$, and hence $\boldsymbol{x}_\infty \in \operatorname{zer}\nabla F$.

We now show that (iv) holds. By contradiction, if $\left(\operatorname{dist}(\mathbf{x}_k(\omega), \chi^\infty(\omega)\right)_{k\in\mathbb{N}}$ does not converge to 0, then there exist $\varepsilon > 0$ and a subsequence $(\mathbf{x}_{\psi_1(k)}(\omega))_{k\in\mathbb{N}}$ such that

$$(\forall k \in \mathbb{N}) \quad \operatorname{dist}\left(\mathbf{x}_{\psi_1(k)}(\omega), \chi^\infty(\omega)\right) > \epsilon \tag{8.3}$$

Since $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ is bounded, $(\mathbf{x}_{\psi_1(k)}(\omega))_{k\in\mathbb{N}}$ is also bounded. So the set of cluster points of $(\mathbf{x}_{\psi_1(k)}(\omega))_{k\in\mathbb{N}}$ is non-empty and included in $\chi^\infty(\omega)$. Thus, there exists another subsequence $(\mathbf{x}_{(\psi_1\circ\psi_2)(k)}(\omega))_{k\in\mathbb{N}}$ and $\boldsymbol{x}'_\infty \in \chi^\infty(\omega)$ such that $\left\|\mathbf{x}_{(\psi_1\circ\psi_2)(k)}(\omega) - \boldsymbol{x}'_\infty\right\| \underset{k\to+\infty}{\longrightarrow} 0$.

Hence, $\operatorname{dist}\left(\mathbf{x}_{(\psi_1\circ\psi_2)(k)}(\omega), \chi^\infty(\omega)\right) \underset{k\to+\infty}{\longrightarrow} 0$, which contradicts (8.3) and thus concludes the proof. □

Proposition 8.2. *Under Assumption 8.1 and Assumption 8.2,* $\left\{\omega \in \Omega \;\middle|\; \mathrm{F}_\infty(\omega) \in F(\operatorname{zer}\nabla F)\right\}$ *is a probability-one set.*

Proof. According to Assumption 8.2 and Proposition 8.1, there exists a set $\Lambda' \subset \Omega$ of probability one where, for every $\omega \in \Lambda'$, $\lim_{k\to+\infty} F(\mathbf{x}_k(\omega)) = \mathrm{F}_\infty(\omega)$, $\chi^\infty(\omega) \neq \emptyset$, and $\chi^\infty(\omega) \subset \operatorname{zer}\nabla F$.

Then, for every $\omega \in \Lambda'$, there exist a vector $\boldsymbol{x}_\infty \in \chi^\infty(\omega)$ and a subsequence $\left(\mathbf{x}_{\psi(k)}(\omega)\right)_{k\in\mathbb{N}}$ such that $\mathbf{x}_{\psi(k)}(\omega) \underset{k\to+\infty}{\longrightarrow} \boldsymbol{x}_\infty$, and $F(\mathbf{x}_{\psi(k)}(\omega)) \underset{k\to+\infty}{\longrightarrow} \mathrm{F}_\infty(\omega)$. In addition, since F is continuous, we deduce that $F(\mathbf{x}_{\psi(k)}(\omega)) \underset{k\to+\infty}{\longrightarrow} F(\boldsymbol{x}_\infty)$.

Hence $\mathrm{F}_\infty(\omega) = F(\boldsymbol{x}_\infty) \in F(\chi^\infty)$. Using the fact that $\chi^\infty(\omega) \subset \operatorname{zer} \nabla F$, we then obtain $\mathrm{F}_\infty(\omega) \in F(\operatorname{zer} \nabla F)$, which concludes the proof. □

The next theorem is a first convergence result occurring in the particular case when $\operatorname{zer} \nabla F$ is at most countable, i.e., if it is either countable or finite.

Theorem 8.1. *Assume that Assumption 8.1 Assumption 8.2 are verified and that* $(\|\mathbf{x}_{k+1} - \mathbf{x}_k\|)_{k\in\mathbb{N}}$ *converges a.s. to* 0*. If* $\operatorname{zer} \nabla F$ *is at most countable, then sequence* $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *almost surely converges to a point belonging to this set.*

Proof. Since χ^∞ is compact a.s. and $(\|\mathbf{x}_{k+1} - \mathbf{x}_k\|)_{k\in\mathbb{N}}$ converges a.s. to 0, then, according to the Ostrowski's Lemma [182, 26.1], we can deduce that χ^∞ is connex a.s..

Moreover, χ^∞ is a.s. non empty, and at most countable as contained in $\operatorname{zer} \nabla F$. We thus deduce that χ^∞ is a.s. reduced to a singleton $\{\mathbf{x}^*\}$, where $\mathbf{x}^* \in \operatorname{zer} \nabla F$. Since $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is a.s. bounded, it converges a.s. to $\mathbf{x}^*$, hence concluding the proof.

□

8.3 KL theory as a baseline of improvement

When $\operatorname{zer} \nabla F$ is neither countable nor finite, there is a lack of information on the curvature of F to ensure the convergence of the general stochastic scheme to a critical point. The use of non-convex functions has encouraged the development of alternative theoretical tools, in particular in a deterministic context. One of the most famous is the class of KL inequalities [141, 25] enabling interesting gradient properties.

To this aim, we first need to introduce some notation.

For every $\zeta \in (0, +\infty]$, we denote by Φ_ζ the set of concave functions $\varphi\colon [0, \zeta) \mapsto \mathbb{R}_+$ such that $\varphi(0) = 0$, φ is continuous in 0, $\varphi \in C^1((0, \zeta))$, and, for every $s \in (0, \zeta)$, $\varphi'(s) > 0$.

The link between Φ_ζ (for $\zeta > 0$) and $\Phi_{+\infty}$ is described through the following proposition.

Proposition 8.3. *Let* $\zeta \in (0, +\infty)$*. Any function* φ *of* Φ_ζ *admits a bounded extension* $\widetilde{\varphi}$ *belonging to* $\Phi_{+\infty}$*.*

Proof. Let $\varphi \in \Phi_\zeta$. Since φ' is positive, it follows that φ is an increasing function. Then $l_1 = \lim_{s\to\zeta^-} \varphi(s)$ exists and lies in $[0, +\infty]$. Moreover the concavity and differentiability of φ with $\varphi(0) = 0$ ensure that, for every $s \in (0, \eta)$, $\varphi(s) \leq s\varphi'(0)$. Passing to the limit thus gives $l_1 \leq \eta\varphi'(0)$ and then $l_1 < +\infty$.

Moreover, since φ' is decreasing on $(0, \zeta)$ (due to the concavity of φ) and positive, we conclude that $l_2 = \lim_{s\to\zeta^-} \varphi'(s)$ exists and lies in $[0, +\infty)$.

Finally, we deduce that there exists a function $\widetilde{\varphi}\colon [0, +\infty) \to \mathbb{R}_+$ defined as (see figure 8.1)

$$(\forall s \in [0, +\infty)) \quad \widetilde{\varphi} = \begin{cases} \varphi(s) & \text{if } s \in [0, \zeta), \\ l_1 + l_2\zeta\left(1 - \frac{\zeta}{s}\right) & \text{otherwise.} \end{cases} \tag{8.4}$$

$\widetilde{\varphi}$ belongs to $\Phi_{+\infty}$ and is bounded. □

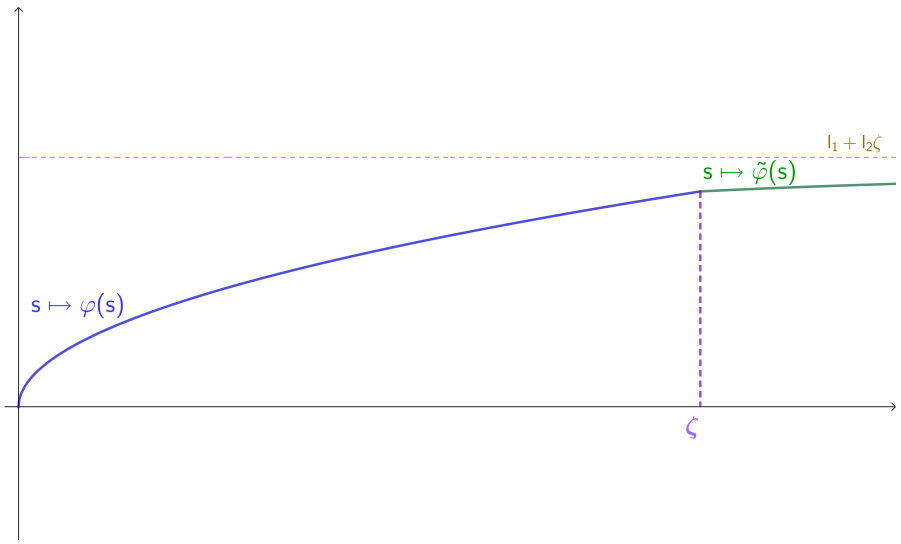


Figure 8.1: Graphical representation of the proof of Proposition 8.3, $\tilde{\varphi}$ is a C^1 extension of φ and especially admits $l_1 + \zeta l_2$ as a limit to $+\infty$.

8.3.1 Extension of the uniformized KL theorem

In this section we extend classical KL results for differentiable functions. To this aim, we first recall the generic definition of differentiable functions satisfying KL, as introduced in, e.g., [9, 27].

Definition 8.1. [KL inequality] *A differentiable function $f : \mathbb{R}^N \to \mathbb{R}$ satisfies the* Kurdyka-Lojasiewicz *property on $E \subset \mathcal{H}$, if for every $\widetilde{\boldsymbol{x}} \in E$, there exists a neighbourhood V of $\widetilde{\boldsymbol{x}}$, $\zeta > 0$ and $\varphi \in \Phi_\zeta$ such that $\|\nabla f(\boldsymbol{x})\|\varphi'(f(\boldsymbol{x}) - f(\widetilde{\boldsymbol{x}})) \geq 1$, for every $\boldsymbol{x} \in V$ satisfying $0 < f(\boldsymbol{x}) - f(\widetilde{\boldsymbol{x}}) < \zeta$.*

One major result following the KL inequality is the uniformized KL property introduced in [27, Lemma 6]. This extended KL condition has been used to prove convergence of deterministic algorithms in a non-convex setting, especially when considering block alternating approaches [27, 76].

Theorem 8.2. [Uniformized KL property] *Let C be a compact set in $\mathcal{H}$ and $f \colon \mathcal{H} \to \mathbb{R}$ be a differentiable function constant on C, satisfying the KL property on C. Then, there exist $(\varepsilon, \zeta) \in (0, +\infty)^2$ and $\varphi \in \Phi_\zeta$ such that*

$$(\forall \overline{\boldsymbol{x}} \in C)(\forall \boldsymbol{x} \in \mathcal{H}) \quad \|\nabla f(\boldsymbol{x})\|\, \varphi'(f(\boldsymbol{x}) - f(\overline{\boldsymbol{x}})) \geq 1, \tag{8.5}$$

when $\operatorname{dist}(\boldsymbol{x}, C) < \varepsilon$ *and* $0 < f(\boldsymbol{x}) - f(\overline{\boldsymbol{x}}) < \zeta$.

We propose an extension of Theorem 8.2 to apply KL on a finite union of compact sets, for functions that are piecewise constant on this union. This result will be instrumental to prove a.s. convergence of stochastic processes. To this aim, we introduce an additional notation. Let C be a non-empty subset of $\mathcal{H}$ and $f \colon \mathcal{H} \to \mathbb{R}$ be a function that is constant on C. Then we denote by f_C the value taken by f on C, i.e., for every $x \in C$, $f(x) = f_C$.

Theorem 8.3. [Extended uniformized KL property] *Let $C = \bigcup_{i=1}^{I} C_i$ be a union of $I \in \mathbb{N}_*$ non-empty disjoint and compact subsets $(C_i)_{1\leq i\leq I}$ of $\mathcal{H}$, and $f\colon \mathcal{H} \to \mathbb{R}$ be a differentiable function satisfying the KL property on C. We also suppose that f is constant on every C_i, for $i \in \{1,\dots,I\}$, with respective values $f_{C_1},\dots,f_{C_I}$. Then, there exist $(\varepsilon,\zeta) \in (0,+\infty)^2$ and $\varphi \in \Phi_\zeta$ such that*

$$(\forall i \in \{1,\dots,I\})(\forall \boldsymbol{x} \in \mathcal{H}) \quad \|\nabla f(\boldsymbol{x})\|\varphi'(f(\boldsymbol{x}) - f_{C_i}) \geq 1 \tag{8.6}$$

with $\operatorname{dist}(\boldsymbol{x}, C) < \varepsilon$ *and* $0 < f(\boldsymbol{x}) - f_{C_i} < \zeta$.

Proof. Without loss of generality we consider that $f_{C_1} \neq \cdots \neq f_{C_I}$.

We start by applying the uniform KL property to $C_1,\dots,C_I$. For every $i \in \{1,\dots,I\}$, there exist $(\varepsilon_i,\zeta_i) \in (0,+\infty)^2$ and $\varphi_i \in \Phi_{\zeta_i}$ such that for every $\boldsymbol{x} \in \mathcal{H}$ verifying $\operatorname{dist}(\boldsymbol{x}, C_i) < \varepsilon_i$ and $0 < f(\boldsymbol{x}) - f_{C_i} < \zeta_i$ we have

$$\|\nabla f(\boldsymbol{x})\|\varphi_i'(f(\boldsymbol{x}) - f_{C_i}) \geq 1. \tag{8.7}$$

Let $\widetilde{\zeta} = \upsilon \min_{j\neq i} |f_{C_i} - f_{C_j}|$ with $\upsilon \in]0,1/2[$. We have $\widetilde{\zeta} > 0$. In addition, using the continuity of f, for every $i \in \{1,\dots,I\}$, there exists $\widetilde{\varepsilon}_i \in]0,\varepsilon_i[$ such that, for every $\boldsymbol{x}$ satisfying $\operatorname{dist}(\boldsymbol{x}, C_i) < \widetilde{\varepsilon}_i$, we have

$$|f(\boldsymbol{x}) - f_{C_i}| < \widetilde{\zeta}. \tag{8.8}$$

Let $\delta \in]0,1[$ and $\varepsilon = \delta \min_{1\leq i\leq I} \widetilde{\varepsilon}_i$. Then

$$\{\boldsymbol{x} \in \mathcal{H} \mid \operatorname{dist}(\boldsymbol{x}, C) < \varepsilon\} \subset \bigcup_{i=1}^{I} \{\boldsymbol{x} \in \mathcal{H} \mid \operatorname{dist}(\boldsymbol{x}, C_i) < \widetilde{\varepsilon}_i\}. \tag{8.9}$$

We now show that (8.6) is satisfied for ε as defined above, $\zeta = \min(\zeta_1,\dots\zeta_I,\widetilde{\zeta})$ and $\varphi = \sum_{i=1}^{I} \varphi_i \in \Phi_\zeta$. Let $\boldsymbol{x} \in \mathcal{H}$ and $i \in \{1,\dots,I\}$ be such that $\operatorname{dist}(\boldsymbol{x}, C) < \varepsilon$ and $0 < f(\boldsymbol{x}) - f_{C_i} < \zeta$. For every $j \in \{1,\dots,I\} \setminus \{i\}$, since $\zeta \leq \widetilde{\zeta}$, using the definition of $\widetilde{\zeta}$, we have $0 < |f(\boldsymbol{x}) - f_{C_i}| \leq \upsilon |f_{C_i} - f_{C_j}|$. Since that $\upsilon \in]0,1/2[$, we obtain

$$\begin{aligned}
|f(\boldsymbol{x}) - f_{C_j}| &= |(f(\boldsymbol{x}) - f_{C_i}) + (f_{C_i} - f_{C_j})|, \\
&\geq \big|\, |f_{C_i} - f_{C_j}| - |f(\boldsymbol{x}) - f_{C_i}| \,\big| = |f_{C_i} - f_{C_j}| - |f(\boldsymbol{x}) - f_{C_i}|, \\
&\geq (1-\upsilon)|f_{C_i} - f_{C_j}| > \upsilon |f_{C_i} - f_{C_j}|, \\
&\geq \widetilde{\zeta}.
\end{aligned} \tag{8.10}$$

Then $\operatorname{dist}(\boldsymbol{x}, C_j) \geq \widetilde{\varepsilon}_j$ (otherwise this contradicts the continuity property of f in (8.8)). So, as $\operatorname{dist}(\boldsymbol{x}, C) < \varepsilon$, (8.9) necessarily implies that $\boldsymbol{x}$ belongs to the union of sets $\bigcup_{j=1}^{I} \{\boldsymbol{x} \in \mathcal{H} \mid \operatorname{dist}(\boldsymbol{x}, C_j) < \widetilde{\varepsilon}_j\}$. This leads to $\operatorname{dist}(\boldsymbol{x}, C_i) < \widetilde{\varepsilon}_i$ and we finally deduce that $\operatorname{dist}(\boldsymbol{x}, C_i) < \varepsilon_i$ (since $\widetilde{\varepsilon}_i \leq \varepsilon_i$). In addition, since $\zeta \leq \zeta_i$, we have $0 < f(\boldsymbol{x}) - f_{C_i} < \zeta_i$, and we can apply once more the uniform KL property at C_i (i.e., (8.7) is verified). Since $\varphi_1',\dots,\varphi_I'$ are all positives, we thus deduce that

$$\begin{aligned}
\|\nabla f(\boldsymbol{x})\|\varphi'(f(\boldsymbol{x}) - f_{C_i}) &= \|\nabla f(\boldsymbol{x})\| \sum_{j=1}^{I} \varphi_j'(f(\boldsymbol{x}) - f_{C_i}) \\
&\geq \|\nabla f(\boldsymbol{x})\|\varphi_i'(f(\boldsymbol{x}) - f_{C_i}) \geq 1.
\end{aligned}$$

This completes the proof. □

8.3.2 A stochastic KL property

Motivated by the results presented in the previous subsection, we make the following assumption on function F.

Assumption 8.3.

(i) *F satisfies the KL property on* $\operatorname{zer}\nabla F$*, the set of critical points of F.*

(ii) *There exist $I \in \mathbb{N}_*$ non-empty disjoint compact subsets $C_1, \ldots, C_I$ of $\mathcal{H}$, such that* $\operatorname{zer}\nabla F = \bigcup_{i=1}^{I} C_i$.

Note that assuming that F satisfies the KL property on $\operatorname{zer}\nabla F$ (i.e., Assumption 8.3(i)) is common in non-convex optimization. As emphasized, e.g., in [25], the KL inequality is satisfied for a wide class of functions, and in particular by real analytic, semi-algebraic[1] and log-exp functions. In addition, Assumption 8.3(ii) is a condition that will be used for the a.s. convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$, through the existence of a uniformized KL function with respect to all the trajectories of the process.

Let

$$\Pi = \liminf_{k\to+\infty} \{\omega \in \Omega \mid F(\mathbf{x}_k(\omega)) > \mathrm{F}_\infty(\omega)\}\,. \tag{8.11}$$

We then deduce the following properties.

Proposition 8.4. *Assume that Assumption 8.1, Assumption 8.2 and Assumption 8.3 hold. Then, over the set* Π*, there exist a bounded $\varphi \in \Phi_{+\infty}$ and an a.s. finite positive discrete random variable* K *such that, for every $k >$ K, $\|\nabla F(\mathbf{x}_k)\|\varphi'(F(\mathbf{x}_k) - \mathrm{F}_\infty) \geq 1$ a.s.*

Proof. Let us consider

$$\mathcal{C} = \bigcup_{\omega\in\Theta\cap\Pi} \chi^\infty(\omega) \tag{8.12}$$

where

$$\Theta = \Big\{F(\mathbf{x}_k) \underset{k\to+\infty}{\to} \mathrm{F}_\infty\Big\} \bigcap \Big\{\chi^\infty \subset \operatorname{zer}\nabla F\Big\} \bigcap \Big\{\operatorname{dist}(\mathbf{x}_k, \chi^\infty) \underset{k\to+\infty}{\to} 0\Big\} \bigcap \Big\{\mathrm{F}_\infty(\Omega) \subset F(\operatorname{zer}\nabla F)\Big\} \tag{8.13}$$

is a probability-one set, owing to Assumption 8.2, Proposition 8.1 and Proposition 8.2. The continuity of ∇F (Assumption 8.1) also ensures that $\operatorname{zer}\nabla F$ is closed and then $\overline{\mathcal{C}} \subset \operatorname{zer}\nabla F$, where $\overline{\mathcal{C}}$ denotes the closure of set $\mathcal{C}$. Considering $\mathbb{J} = \big\{i \in \{1, \ldots, I\} \mid \overline{\mathcal{C}} \cap C_i \neq \emptyset\big\}$, it follows that $\overline{\mathcal{C}} = \overline{\mathcal{C}} \cap \operatorname{zer}\nabla F = \bigcup_{i\in\mathbb{J}} \overline{\mathcal{C}} \cap C_i$,

where, for every $i \in \mathbb{J}$, the set $\overline{\mathcal{C}} \cap C_i$ is non-empty, bounded and closed. Moreover, for every $i \in \mathbb{J}$, we have $F(\overline{\mathcal{C}} \cap C_i) = \{F_{C_i}\}$.

According to Assumption 8.3, since F satisfies the KL property on $\operatorname{zer}\nabla F$, we can apply Theorem 8.3. As a consequence there exist $\varepsilon_{\mathcal{C}} > 0$, $\zeta_{\mathcal{C}} > 0$ and $\varphi_{\mathcal{C}} \in \Phi_{\zeta_{\mathcal{C}}}$, such that, for every $\boldsymbol{x} \in \mathcal{H}$ and $i \in \mathbb{J}$ satisfying $\operatorname{dist}(\boldsymbol{x}, \overline{\mathcal{C}}) < \varepsilon_{\mathcal{C}}$ and $0 < F(\boldsymbol{x}) - F_{C_i} < \zeta_{\mathcal{C}}$,

$$\|\nabla F(\boldsymbol{x})\|\ \varphi'_{\mathcal{C}}(F(\boldsymbol{x}) - F_{C_i}) \geq 1. \tag{8.14}$$

[1] A function is semi-algebraic if its graph is a finite union of sets defined by a finite number of polynomial inequalities.

According to Proposition 8.3, $\varphi_{\mathcal{C}}$ has a bounded extension $\widetilde{\varphi}_{\mathcal{C}}$ belonging to $\Phi_{+\infty}$. Then, $\widetilde{\varphi}_{\mathcal{C}}$ also satisfies, for any $\boldsymbol{x} \in \mathcal{H}$ and $i \in \mathbb{J}$ such that $\operatorname{dist}(\boldsymbol{x}, \overline{\mathcal{C}}) < \varepsilon_{\mathcal{C}}$ and $0 < F(\boldsymbol{x}) - F_{C_i} < \zeta_{\mathcal{C}}$,

$$\|\nabla F(\boldsymbol{x})\| \; \widetilde{\varphi}'_{\mathcal{C}}(F(\boldsymbol{x}) - F_{C_i}) \geq 1. \tag{8.15}$$

Considering this, we now define the positive discrete random variable

$$\mathrm{K} = \min\left\{ l \in \mathbb{N}_* \;\middle|\; (\forall p \geq l) \;\; \operatorname{dist}(\mathbf{x}_p, \overline{\mathcal{C}}) \leq \varepsilon_{\mathcal{C}} \text{ and } 0 < F(\mathbf{x}_p) - \mathrm{F}_\infty < \zeta_{\mathcal{C}} \right\}. \tag{8.16}$$

The latter is finite over $\Theta \cap \Pi$. In addition, according to Assumption 8.3, for every $\omega \in \Theta \cap \Pi$, there exists $i \in \mathbb{J}$ such that $\mathrm{F}_\infty(\omega) = F_{C_i}$, and (8.15) finally leads to

$$(\forall \omega \in \Theta \cap \Pi) \quad (\forall k > \mathrm{K}(w)) \quad \|\nabla F(\mathbf{x}_k(\omega))\| \; \widetilde{\varphi}'_{\mathcal{C}}\left(F(\mathbf{x}_k(\omega)) - \mathrm{F}_\infty(\omega)\right) \geq 1. \tag{8.17}$$

The fact that $\mathbb{P}(\Theta \cap \Pi) = P(\Pi)$ concludes the proof. □

In [152, Assumption 3.9, C.2], the authors have highlighted that the construction of a stochastic KL inequality was based on the set $\mathcal{C}$ as introduced in (8.12) without however going further in the reasoning. In addition, in [152, 95], the authors assumed the existence of a uniformized function with respect to all trajectories to enable their convergence results to hold.

8.4 An almost sure convergence result based on KL theory

In this section, we give conditions that a stochastic process must satisfy to ensure its convergence to a critical point of F. To this aim, we introduce additional notations that will be used in the remainder of this chapter. We associate the initial probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$. Moreover, for a given $k \in \mathbb{N}$ and subject to existence, we denote by $\mathbb{E}(.|\mathcal{F}_k)$ the conditional expectation operator associated with the sub sigma-algebra $\mathcal{F}_k$. We also denote by F^* the minimal value of F.

8.4.1 Main assumption and summability criterion

The following assumption is the backbone of our convergence theorem, given in subsection 8.4.2.

Assumption 8.4. *Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a stochastic process.*

(i) *$(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ and $\left(\|\nabla F(\mathbf{x}_k)\|^2\right)_{k\in\mathbb{N}}$ are two $\mathcal{F}_k$-measurable and integrable processes.*

(ii) *There exist $(u_k)_{k\in\mathbb{N}}, (v_k)_{k\in\mathbb{N}}, (r_k)_{k\in\mathbb{N}}, (s_k)_{k\in\mathbb{N}}, (t_k)_{k\in\mathbb{N}}$ five non-negative deterministic sequences and $(\mathrm{w}_k)_{k\in\mathbb{N}}$ a non-negative $\mathcal{F}_k$-measurable integrable process, such that*

$$\sum_{k=0}^{+\infty} u_k < +\infty, \;\; \inf_{k\in\mathbb{N}} v_k > 0, \;\; \sum_{k=0}^{+\infty} r_k < +\infty, \;\; \sum_{k=0}^{+\infty} t_k < +\infty, \;\; \sum_{k=0}^{+\infty} \mathrm{w}_k < \; +\infty \;\; a.s., \tag{8.18}$$

and such that, for every $k \in \mathbb{N}$, we have

$$\mathbb{E}[F(\mathbf{x}_{k+1}) - F^* | \mathcal{F}_k] \leq (1 + u_k)(F(\mathbf{x}_k) - F^*) - v_k \|\nabla F(\mathbf{x}_k)\|^2 + \mathrm{w}_k \;\; a.s., \tag{8.19}$$

and

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \mid \mathcal{F}_k] \leq r_k \sqrt{F(\mathbf{x}_k) - F^*} + s_k \|\nabla F(\mathbf{x}_k)\| + t_k \;\; a.s.. \tag{8.20}$$

Assumption 8.4(i) is a common assumption, satisfied in particular when the randomness appearing at iteration $k \in \mathbb{N}$ is independent from what happened during the previous iterations. In Assumption 8.4(ii), condition (8.19) ensures that $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ is a quasi-supermartingale [202]. Additionally, (8.20) ensures that the norm difference between consecutive iterates is upper bounded, relatively to the norm of the gradient of F. In practice, these two inequalities often imply an underlying assumption on the nature of the stochastic phenomenon at stake. Note that Assumption 8.4 is slightly stronger than Assumption 8.2, as emphasized by the next proposition.

Proposition 8.5. *If Assumption 8.1 and Assumption 8.4 are satisfied, then Assumption 8.2 holds.*

Proof. Condition (8.19) and the fact that F is bounded from below (since it is coercive) allows us to invoke the Robbins-Siegmund Lemma [202] which directly ensures that Assumption 8.2 is verified with $\sum_{k=0}^{+\infty} v_k \|\nabla F(\mathbf{x}_k)\|^2 < +\infty$ a.s..

The fact that $\inf_{k\in\mathbb{N}} v_k > 0$ then directly gives $\lim_{k\to+\infty} \|\nabla F(\mathbf{x}_k)\| = 0$ a.s.. □

Before going further, we need to introduce additional notation. Let $(p_k)_{k\in\mathbb{N}}$ be the positive sequence defined by

$$(\forall k \in \mathbb{N}) \quad p_k = \prod_{i=0}^{k} (1 + u_i), \tag{8.21}$$

where $(u_i)_{i\in\mathbb{N}}$ is defined in Assumption 8.4(ii). Let, for very $k \in \mathbb{N}_*$,

$$\begin{array}{rccl} L_k\colon & \mathcal{H} \times \mathbb{R}^k & \to & \mathbb{R} \\ & (\boldsymbol{x}, w_0, \dots, w_{k-1}) & \mapsto & (F(\boldsymbol{x}) - F^*) p_{k-1}^{-1} - \sum_{i=0}^{k-1} w_i p_i^{-1} \end{array} \tag{8.22}$$

Defining such a function enables to handle a supermartingale where the conditional decrease is only relative to the gradient, instead of the quasi-supermartingale appearing in condition (8.19).

In this context, for every $k \in \mathbb{N}_*$ and $(w_0, \dots, w_{k-1}) \in \mathbb{R}^k$, $L_k(\cdot, w_0, \dots, w_{k-1})$ plays the role of a Lyapunov function which can be use to get an equivalent formulation of (8.19).

Lemma 8.1. *Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a stochastic process, and let $(\mathrm{w}_k)_{k\in\mathbb{N}}$ be an integrable $\mathcal{F}_k$-measurable non-negative process satisfying* (8.18)*. Let also $(v_k)_{k\in\mathbb{N}}$ be a non-negative deterministic sequence satisfying* (8.18)*, and let $(p_k)_{k\in\mathbb{N}}$ be the sequence defined in* (8.21)*. Condition* (8.19) *is equivalent to*

$$\begin{aligned} (\forall k \in \mathbb{N}_*) \quad & \mathbb{E}[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \dots, \mathrm{w}_k) | \mathcal{F}_k] \\ & \leq L_k(\mathbf{x}_k, \mathrm{w}_0, \dots, \mathrm{w}_{k-1}) - \frac{v_k}{p_k} \|\nabla F(\mathbf{x}_k)\|^2 \ \text{a.s.}. \end{aligned} \tag{8.23}$$

Proof. For every $k \geq 1$, $\boldsymbol{x} \in \mathcal{H}$ and $(w_0, \dots, w_{k-1}) \in \mathbb{R}^k$, we have

$$F(\boldsymbol{x}) - F^* = p_{k-1}\Big(L_k(\boldsymbol{x}, w_0, \dots, w_{k-1}) + \sum_{i=0}^{k-1} w_i p_i^{-1}\Big).$$

Using this equality, for every $k \geq 1$, (8.19) is then equivalent to

$$p_k\Big(\mathbb{E}[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k] + \sum_{i=0}^{k} \mathrm{w}_i p_i^{-1}\Big) \\ \leq (1+u_k)p_{k-1}\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) + \sum_{i=0}^{k-1} \mathrm{w}_i p_i^{-1}\Big) - v_k\|\nabla F(\mathbf{x}_k)\|^2 + \mathrm{w}_k \text{ a.s.}.$$

Dividing this inequality by $p_k = p_{k-1}(1+u_k) > 0$, we obtain

$$\mathbb{E}[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k] \\ \leq L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \Big(\sum_{i=0}^{k} \mathrm{w}_i p_i^{-1} - \sum_{i=0}^{k-1} \mathrm{w}_i p_i^{-1}\Big) - \frac{v_k}{p_k}\|\nabla F(\mathbf{x}_k)\|^2 + \mathrm{w}_k p_k^{-1} \text{ a.s.} \quad (8.24)$$

The equivalence between (8.24) and (8.23) then follows from the fact that $\sum_{i=0}^{k} \mathrm{w}_i p_i^{-1} - \sum_{i=0}^{k-1} \mathrm{w}_i p_i^{-1} = \mathrm{w}_k p_k^{-1}$.

□

Remark 8.1. *Let $k \in \mathbb{N}_*$ and $(\mathrm{w}_i)_{0\leq i\leq k-1}$ be defined as in Assumption 8.4(ii). If the a.s. finite limit F_∞ of $(F(\mathbf{x}_l))_{l\in\mathbb{N}}$ exists, then the a.s. finite limit $\mathrm{L}_{k,\infty}$ of process $(L_k(\mathbf{x}_l, \mathrm{w}_1, \ldots \mathrm{w}_{k-1}))_{l\in\mathbb{N}}$ exists as well, and is given by*

$$\mathrm{L}_{k,\infty} = (\mathrm{F}_\infty - F^*)p_k^{-1} - \sum_{i=0}^{k-1} \mathrm{w}_i p_i^{-1}. \quad (8.25)$$

In the case when F_∞ exists, we define, for every $\gamma > 0$, the following set of events:

$$\Xi_\gamma = \Big\{\omega \in \Omega \;\Big|\; (\forall k \geq 1)\ \ \mathrm{F}_\infty(\omega) < F(\mathbf{x}_k(\omega)) \\ \text{and } \big|\mathrm{L}_{k,\infty} - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\big| \leq \gamma\frac{v_k}{p_k}\|\nabla F(\mathbf{x}_k(\omega))\|^2\Big\}. \quad (8.26)$$

In the remainder, to obtain summability and convergence results, we will assume that there exists $\gamma \in (0,1)$ such that $\mathbb{P}(\Xi_\gamma) = 1$. Intuitively, this assumption means that F_∞ must a.s. be a lower bound of process $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$.

The second inequality in (8.26) indicates that, at iteration $k \in \mathbb{N}$, the error on the identification of $\mathrm{L}_{k,\infty}$ (equivalently, of F_∞) is upper bounded by the squared norm of ∇F.

Proposition 8.6. *Suppose that Assumption 8.1, Assumption 8.3 and Assumption 8.4 hold. Moreover, assume that $\big(s_k p_k v_k^{-1}\big)_{k\in\mathbb{N}}$ is a non-increasing sequence and that there exists $\gamma \in (0,1)$ such that $\mathbb{P}(\Xi_\gamma) = 1$. Let $\varphi \in \Phi_{+\infty}$ be a bounded function, and*

$$\Gamma_\varphi := \sum_{k=1}^{+\infty} \frac{s_k p_k}{v_k}\Big(\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big) \\ - \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big)\Big). \quad (8.27)$$

Then $\Gamma_\varphi < +\infty$ a.s..

Proof. According to Proposition 8.5, Assumption 8.2 holds, so, for every $k \in \mathbb{N}_*$, $\mathrm{L}_{k,\infty}$, the a.s. limit of process $(L_k(\mathbf{x}_l, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}))_{l\in\mathbb{N}_*}$ exists and is given by (8.25).

According to Assumption 8.3 and Assumption 8.2, F_∞ belongs to a finite set. In addition, according to Assumption 8.4, $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ and $(\mathrm{w}_k)_{k\in\mathbb{N}}$ are integrable. Thus, for every $k \in \mathbb{N}_*$, $L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)$ and $\mathrm{L}_{k+1,\infty}$ are integrable, and hence $\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right]$ and $\mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]$ are well-defined.

On the one hand, there exists $\gamma \in (0, 1)$ such that $\mathbb{P}(\Xi_\gamma) = 1$. Using the definition of Ξ_γ, we therefore have $\mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k] - \mathrm{L}_{k,\infty} \geq -\gamma\frac{v_k}{p_k}\|\nabla F(\mathbf{x}_k)\|^2$ a.s..

Combining this inequality with (8.23), we obtain, for every $k \geq 1$,

$$L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty} - \Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big) \geq (1-\gamma)\frac{v_k}{p_k}\|\nabla F(\mathbf{x}_k)\|^2 \text{ a.s..} \quad (8.28)$$

Since $1 - \gamma > 0$, we deduce that

$$(\forall k \in \mathbb{N}) \quad L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty} \geq \mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k] \text{ a.s..} \quad (8.29)$$

On the other hand, following the definitions (8.22) and (8.25) with $\mathbb{P}(\Xi_\gamma) = 1$, for every $k \in \mathbb{N}_*$, the two random variables in (8.29), $(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty})$ and $(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k])$ are a.s. non-negative. Let $\varphi \in \Phi_{+\infty}$ be bounded. Then, the function φ can be applied to the a.s. positive random variables $(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty})$ and $(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k])$.

Furthermore, φ being an increasing function, (8.29) leads to

$$(\forall k \in \mathbb{N}) \quad \varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big) \geq \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big) \text{ a.s.,} \quad (8.30)$$

and consequently Γ_φ is a.s. well-defined, non-negative as an infinite sum of a.s. non-negative terms. We can subsequently take the expectation of these quantities, and use the monotone convergence theorem [145] to obtain

$$\mathbb{E}\left[\Gamma_\varphi\right] = \sum_{k=1}^{+\infty} \frac{s_k p_k}{v_k}\mathbb{E}\Big[\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big) - \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big)\Big]. \quad (8.31)$$

Since φ is bounded, we can use the linearity of the expectation, leading to

$$\mathbb{E}\left[\Gamma_\varphi\right] = \sum_{k=1}^{+\infty} \frac{s_k p_k}{v_k}\mathbb{E}\left[\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big)\right] - \frac{s_k p_k}{v_k}\mathbb{E}\left[\varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big)\right]. \quad (8.32)$$

Using the inverse conditional Jensen's inequality [145] applied to φ gives

$$\begin{aligned}
&\varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big), \\
&\quad = \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}|\mathcal{F}_k\right]\Big) \\
&\quad \geq \mathbb{E}\left[\varphi\Big(L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}\Big)|\mathcal{F}_k\right]. \qquad (8.33)
\end{aligned}$$

Taking the expectation $\mathbb{E}$ of these quantities leads to

$$\begin{aligned}
&\mathbb{E}\left[\varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big)\right], \\
&\quad \geq \mathbb{E}\left[\mathbb{E}\left[\varphi\Big(L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}\Big)|\mathcal{F}_k\right]\right] \\
&\quad = \mathbb{E}\left[\varphi\Big(L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}\Big)\right]. \qquad (8.34)
\end{aligned}$$

Combining (8.32) with (8.34), and since $\left(\frac{s_k p_k}{v_k}\right)_{k\in\mathbb{N}}$ is non-increasing, we obtain

$$\begin{aligned}
\mathbb{E}\left[\Gamma_\varphi\right] &\leq \sum_{k=1}^{+\infty} \frac{s_k p_k}{v_k} \mathbb{E}\left[\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big)\right] \\
&\quad - \frac{s_k p_k}{v_k} \mathbb{E}\left[\varphi\Big(L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}\Big)\right], \\
&\leq \sum_{k=1}^{+\infty} \frac{s_k p_k}{v_k} \mathbb{E}\left[\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big)\right] \\
&\quad - \frac{s_{k+1} p_{k+1}}{v_{k+1}} \mathbb{E}\left[\varphi\Big(L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}\Big)\right]. \qquad (8.35)
\end{aligned}$$

From (8.35), since φ is an increasing and bounded function on $\mathbb{R}_+$, we deduce that

$$0 \leq \mathbb{E}\left[\Gamma_\varphi\right] \leq \frac{s_1 p_1}{v_1} \mathbb{E}\left[\varphi\Big(L_1(\mathbf{x}_1, \mathrm{w}_0) - \mathrm{L}_{1,\infty}\Big)\right] < +\infty. \qquad (8.36)$$

Hence Γ_φ admits a finite expectation and consequently is a.s. finite. □

8.4.2 Main result

Theorem 8.4. *Under Assumption 8.1, Assumption 8.3 and Assumption 8.4, if* $\left(s_k p_k v_k^{-1}\right)_{k\in\mathbb{N}}$ *is a non-increasing sequence and if there exists* $\gamma \in (0,1)$ *such that* $\mathbb{P}(\Xi_\gamma) = 1$*, then*

(i) $\sum_{k=1}^{+\infty} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| < +\infty$ *a.s.,*

(ii) *the sequence* $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *converges a.s. to a critical point of* F*.*

Proof. According to Proposition 8.5, Assumption 8.2 holds. In addition, since $\Xi_\gamma \subset \Pi$, where Π is defined in (8.11), and $\mathbb{P}(\Xi_\gamma) = 1$, we have $\mathbb{P}(\Pi) = 1$. Thus Proposition 8.4 can be applied, and there exist $\varphi \in \Phi_{+\infty}$ and an a.s. finite non-negative discrete random variable K such that

$$(\forall k > \mathrm{K}) \quad \|\nabla F(\mathbf{x}_k)\| \varphi'(F(\mathbf{x}_k) - \mathrm{F}_\infty) \geq 1 \ \text{ a.s.}. \qquad (8.37)$$

According to the definitions of L_k and $\mathrm{L}_{k,\infty}$ in (8.22) and (8.25), respectively, we have

$$(\forall k \in \mathbb{N}_*) \quad F(\mathbf{x}_k) - \mathrm{F}_\infty = p_{k-1}\big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\big), \tag{8.38}$$

where $(\mathrm{w}_0, \ldots, \mathrm{w}_{k-1})$ are defined in Assumption 8.4(ii). Hence (8.37) rewrites

$$(\forall k > \mathrm{K}) \quad \|\nabla F(\mathbf{x}_k)\| \varphi'\Big(p_{k-1}\big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\big)\Big) \geq 1 \ \text{ a.s.}. \tag{8.39}$$

According to (8.21), $(p_k)_{k\in\mathbb{N}}$ is lower-bounded by 1. Then, since φ is concave, hence φ' is decreasing, and we have

$$(\forall k > \mathrm{K}) \quad \|\nabla F(\mathbf{x}_k)\| \varphi'\big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\big) \geq 1 \ \text{ a.s.}. \tag{8.40}$$

In addition, owing to the concavity of φ, for every $(a, b) \in \mathbb{R}^2$, we have $\varphi(a) - \varphi(b) \geq \varphi'(a)(a - b)$. So, for every $k > \mathrm{K}$, taking $a = L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}$ and $b = \mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]$ in (8.40), we obtain

$$\begin{aligned}
&\varphi\left(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\right) \\
&\quad - \varphi\left(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\right) \\
&\geq \varphi'(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}) \\
&\quad \times \Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty} \\
&\quad - \Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big)\Big) \ \text{a.s.}
\end{aligned} \tag{8.41}$$

Since Proposition 8.6, holds, we can use inequality (8.28). Hence, injecting successively (8.28) and (8.40) in (8.41) leads, for every $k > \mathrm{K}$, to

$$\begin{aligned}
&\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big) \\
&\quad - \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big), \\
&\geq (1-\gamma)\frac{v_k}{p_k}\|\nabla F(\mathbf{x}_k)\|^2 \varphi'\left(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\right), \\
&\geq (1-\gamma)\frac{v_k}{p_k}\|\nabla F(\mathbf{x}_k)\| \ \text{a.s.}.
\end{aligned} \tag{8.42}$$

Combining (8.20) with (8.42) almost surely gives, for every $k > \mathrm{K}$,

$$\begin{aligned}
\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\||\mathcal{F}_k] \leq (1-\gamma)^{-1} v_k^{-1} s_k p_k \Big(&\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big) \\
&- \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big)\Big) \\
&+ r_k\sqrt{F(\mathbf{x}_k) - F^*} + t_k.
\end{aligned} \tag{8.43}$$

Using Proposition 8.6, the summability of $(r_k)_{k\in\mathbb{N}}$ and $(t_k)_{k\in\mathbb{N}}$, and the fact that $\left(\sqrt{F(\mathbf{x}_k) - F^*}\right)_{k\in\mathbb{N}}$ a.s. converges to $\sqrt{\mathrm{F}_\infty - F^*}$, the right hand side term in (8.43) is a.s. summable. Then, since K is also a.s. finite, we can deduce that $\sum_{k=1}^{+\infty} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\||\mathcal{F}_k] < +\infty$ a.s.. Finally, since $(\|\mathbf{x}_{k+1} - \mathbf{x}_k\|)_{k\in\mathbb{N}}$ is a positive sequence, we can apply Levy's sharpening of Borel-Cantelli Lemma [163, Ch.1, Th.21], leading to $\sum_{k=1}^{+\infty} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| < +\infty$ a.s.. It then follows that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is a.s. a Cauchy sequence. Moreover, according to Proposition 8.1, $(\mathbf{x}_k)_{k\in\mathbb{N}}$ a.s. has a critical point of F as an accumulation point. Hence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ a.s. converges to this critical point. □

8.5 Stochastic gradient schemes

In the previous sections we have presented a general framework to show convergence of the iterates of stochastic schemes for solving (8.1) in a non-convex context, when F satisfies Assumption 8.1 and Assumption 8.3. In this section we give examples of stochastic gradient algorithms satisfying Assumption 8.4, hence with convergence guaranteed by Theorem 8.4.

A wide class of stochastic gradient schemes for solving (8.1) are of the form of

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{U}_k \mathbf{f}_k, \tag{8.44}$$

where, for every $k \in \mathbb{N}$, $\mathbf{f}_k \in \mathcal{H}$ usually denotes a stochastic approximation of the gradient of F at $\mathbf{x}_k$, $\alpha_k \in (0, +\infty)$ is the step-size (also called learning rate), and $\mathbf{U}_k \colon \mathcal{H} \to \mathcal{H}$ is a stochastic self-adjoint linear operator usually used as preconditioner (e.g., stochastic Newton or quasi-Newton schemes). Algorithms of the form of (8.44) include variants of the popular SGD [212, 206, 87, 137] as well as schemes incorporating second-order information like stochastic Hessian approximations [233, 162], or more generally preconditioned SGD algorithms [153].

8.5.1 Conditions on the approximated gradient

Let $(\mathcal{F}_k)_{k\in\mathbb{N}}$ be the canonical filtration, i.e., for every $k \in \mathbb{N}$ we have $\mathcal{F}_k = \sigma(\mathbf{x}_0, \dots \mathbf{x}_k)$.

For every $k \in \mathbb{N}$, $\mathbf{f}_k$ in (8.44) aims at approximating the true gradient $\nabla F(\mathbf{x}_k)$. In general, it is assumed to be integrable, and that there exist three non-negative deterministic sequences $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$, and $(c_k)_{k\in\mathbb{N}}$ such that

$$\mathbb{E}[\|\mathbf{f}_k\|^2 \mid \mathcal{F}_k] \leq a_k(F(\mathbf{x}_k) - F^*) + b_k\|\nabla F(\mathbf{x}_k)\|^2 + c_k \quad \text{a.s.}. \tag{8.45}$$

This condition is very generic, and is satisfied by multiple schemes from the literature. For instance, in [216, 133], condition (8.45) is satisfied when $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$, and $(c_k)_{k\in\mathbb{N}}$ are constant, equal to $a \geq 0$, $b \geq 0$ and $c \geq 0$, respectively.

When, in particular $a = 0$, and $c = 0$, condition (8.45) is equivalent to the conditional strong growth condition [212]:

$$(\forall k \in \mathbb{N}) \quad \mathbb{E}[\|\mathbf{f}_k\|^2 \mid \mathcal{F}_k] \leq b\|\nabla F(\mathbf{x}_k)\|^2 \quad \text{a.s.}. \tag{8.46}$$

A relaxed case, only assuming $a = 0$, has also been investigated in [20, 36]. Finally, condition (8.45) also extents those of [71], in the differentiable case, where, for every $k \in \mathbb{N}$, $a_k = 0$, but without assuming that $(b_k)_{k\in\mathbb{N}}$ and $(c_k)_{k\in\mathbb{N}}$ are constant.

8.5.2 Conditions on the step-size

In (8.44), $(\alpha_k)_{k\in\mathbb{N}}$ is a positive sequence playing the role of step-sizes. When they are chosen to be constant, it has been shown in [212] that Assumption 8.4 is verified if the unknown gradients $(\mathbf{f}_k)_{k\in\mathbb{N}}$ satisfy the strong growth condition (8.46). If this condition is not satisfied, in particular for non-constant step-sizes, but a more general condition holds (e.g., (8.45)), $(\alpha_k)_{k\in\mathbb{N}}$ must be decreasing to ensure the robustness of the perturbation on the induced variance [36].

8.5.3 Conditions on the preconditioning operator

In (8.44), $(\mathbf{U}_k)_{k\in\mathbb{N}}$ is a sequence of stochastic preconditioning self-adjoint linear operators, with an a.s. uniformly bounded spectrum, i.e., there exist two positive sequences $(\mu_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ such that

$$(\forall k \in \mathbb{N})(\forall \boldsymbol{x} \in \mathcal{H}) \quad \mu_k \|\boldsymbol{x}\|^2 \leq \langle \boldsymbol{x} \mid \mathbf{U}_k \boldsymbol{x} \rangle \leq \nu_k \|\boldsymbol{x}\|^2 \quad \text{a.s..} \tag{8.47}$$

The most famous class of preconditioning matrices for gradient descent schemes is the Quasi-Newton one. In this context, for every $k \in \mathbb{N}$, $\mathbf{U}_k$ is chosen to be an approximation of the Hessian of F evaluated at the current iterate $\mathbf{x}_k$ (assuming that F is twice-differentiable). The celebrated quasi-Newton BFGS method is one of them. The convergence behavior of $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ has been studied in [233] in a non-convex and stochastic setting. However, to the best of our knowledge, the a.s. convergence of the sequence of iterates $(\mathbf{x}_k)_{k\in\mathbb{N}}$ has not been studied yet in the literature.

8.5.4 A general framework for convergence of SGD algorithms in a non-convex context

In this section, we link the noisy gradient condition (8.45) to the descent conditions of Assumption 8.4. In particular, we show that under mild conditions on the parameters involved in (8.44), (8.45) and (8.47), Theorem 8.4 holds.

Proposition 8.7. *Let F be a β-Lipschitz differentiable function, for some $\beta > 0$, and let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a process verifying* (8.44), (8.45), *and* (8.47) *with respect to $(\mathcal{F}_k)_{k\in\mathbb{N}}$. Assume that*

$$\inf_{k\in\mathbb{N}} \mu_k > 0, \quad \sum_{k=0}^{+\infty} \alpha_k \nu_k a_k^{1/2} < +\infty, \quad \sum_{k=0}^{+\infty} \alpha_k \nu_k c_k^{1/2} < +\infty, \tag{8.48}$$

and that one of the two following statement holds:

(i) *The sequence $(\mathbf{f}_k)_{k\in\mathbb{N}}$ is conditionally unbiased, i.e., for every $k \in \mathbb{N}$, $\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] = \nabla F(\mathbf{x}_k)$ a.s.; and*

$$\inf_{k\in\mathbb{N}} \alpha_k \left(\mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2} \right) > 0. \tag{8.49}$$

(ii) *The sequence $(\mathbf{f}_k)_{k\in\mathbb{N}}$ verifies*

$$\sum_{k=0}^{+\infty} \frac{\alpha_k \nu_k^2}{\mu_k} \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|^2 < +\infty \quad \text{a.s.} \tag{8.50}$$

and there exists $\varrho \in (0,1)$ such that

$$\inf_{k\in\mathbb{N}} \alpha_k \left(\varrho\mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2} \right) > 0. \tag{8.51}$$

If the sequence $(\mathbf{U}_k)_{k\in\mathbb{N}}$ is chosen to be adapted to $(\mathcal{F}_k)_{k\in\mathbb{N}}$ (i.e., for every $k \in \mathbb{N}$, $\mathbf{U}_k$ is $\mathcal{F}_k$-measurable), then $(\mathbf{x}_k)_{k\in\mathbb{N}}$ satisfies Assumption 8.4.

Proof. Following a similar proof as for [57, Lemma 2], it can be shown that the conditional expectations of all manipulated random variable are well-defined and satisfy Assumption 8.4(i). To show that Assumption 8.4(ii) is satisfied, we need to show that both conditions (8.19) and (8.20) hold. To this aim, we will first derive generic inequalities. F being β-Lipschitz differentiable, we can apply the descent lemma [19], and we obtain

$$(\forall k \in \mathbb{N}) \quad F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \alpha_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbf{f}_k \rangle + \frac{\alpha_k^2 \beta}{2} \|\mathbf{U}_k \mathbf{f}_k\|^2. \tag{8.52}$$

Subtracting F^* in (8.52), and passing to the conditional expectation lead to

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad & \mathbb{E}[F(\mathbf{x}_{k+1}) - F^* \mid \mathcal{F}_k], \\ & \leq \mathbb{E}\Big[F(\mathbf{x}_k) - F^* - \alpha_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbf{f}_k \rangle + \frac{\alpha_k^2 \beta}{2} \|\mathbf{U}_k \mathbf{f}_k\|^2 \;\Big|\; \mathcal{F}_k\Big] \quad \text{a.s.}, \\ & = F(\mathbf{x}_k) - F^* - \alpha_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbb{E}[\mathbf{f}_k | \mathcal{F}_k] \rangle + \frac{\alpha_k^2 \beta}{2} \mathbb{E}\big[\|\mathbf{U}_k \mathbf{f}_k\|^2 | \mathcal{F}_k\big] \quad \text{a.s.}, \end{aligned} \tag{8.53}$$

where the equality is obtained using the linearity of the conditional expectation, and the $\mathcal{F}_k$-measurability of $F(\mathbf{x}_k)$, $\nabla F(\mathbf{x}_k)$ and $\mathbf{U}_k$, for all $k \in \mathbb{N}$.

Moreover, from (8.45) and (8.47), it follows that, for every $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{U}_k \mathbf{f}_k\|^2 | \mathcal{F}_k\big] & \leq \nu_k^2 \mathbb{E}\big[\|\mathbf{f}_k\|^2 | \mathcal{F}_k\big], \\ & \leq \nu_k^2 a_k \big(F(\mathbf{x}_k) - F^*\big) + \nu_k^2 b_k \|\nabla F(\mathbf{x}_k)\|^2 + \nu_k^2 c_k \quad \text{a.s.}. \end{aligned} \tag{8.54}$$

Injecting (8.54) in (8.53) gives, for every $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}_{k+1}) - F^* | \mathcal{F}_k] \leq & \left(1 + \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}\right) \big(F(\mathbf{x}_k) - F^*\big) \\ & - \alpha_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbb{E}[\mathbf{f}_k | \mathcal{F}_k] \rangle \frac{\alpha_k^2 \beta \nu_k^2 b_k}{2} \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2}. \end{aligned} \tag{8.55}$$

We will now first show that condition (8.20) is verified in both cases (i) and (ii). We will then show that (8.19) is also satisfied, using separately case (i) or case (ii).

According to the scheme in definition (8.44), we have

$$(\forall k \in \mathbb{N}) \quad \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{F}_k] = \alpha_k^2 \mathbb{E}[\|\mathbf{U}_k \mathbf{f}_k\|^2 | \mathcal{F}_k]. \tag{8.56}$$

Hence, injecting (8.56) in (8.54), we obtain for all $k \in \mathbb{N}$

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{F}_k] \leq \alpha_k^2 \nu_k^2 a_k (F(\mathbf{x}_k) - F^*) + \alpha_k^2 \nu_k^2 b_k \|\nabla F(\mathbf{x}_k)\|^2 + \alpha_k^2 \nu_k^2 c_k \quad \text{a.s.}. \tag{8.57}$$

Using the conditional version of Jensen's inequality [145] leads to

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad & \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\| | \mathcal{F}_k], \\ & \leq \sqrt{\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{F}_k]}, \\ & \leq \sqrt{\alpha_k^2 \nu_k^2 a_k (F(\mathbf{x}_k) - F^*) + \alpha_k^2 \nu_k^2 b_k \|\nabla F(\mathbf{x}_k)\|^2 + \alpha_k^2 \nu_k^2 c_k} \\ & \leq \alpha_k \nu_k a_k^{1/2} \sqrt{F(\mathbf{x}_k) - F^*} + \alpha_k \nu_k b_k^{1/2} \|\nabla F(\mathbf{x}_k)\| + \alpha_k \nu_k c_k^{1/2} \quad \text{a.s.}. \end{aligned} \tag{8.58}$$

Let, for every $k \in \mathbb{N}$, $r_k := \alpha_k \nu_k a_k^{1/2}$, $s_k = \alpha_k \nu_k b_k^{1/2}$, and $t_k = \alpha_k \nu_k c_k^{1/2}$. Sequences $(r_k)_{k\in\mathbb{N}}$, $(s_k)_{k\in\mathbb{N}}$, $(t_k)_{k\in\mathbb{N}}$ are non-negative, and the summabilities of $(r_k)_{k\in\mathbb{N}}$ and $(t_k)_{k\in\mathbb{N}}$ are due to (8.48). Hence (8.20) is satisfied.

We now distinguish case (i) from case (ii) to show that condition (8.20) is also satisfied.
Case (i): Since $(\mathbf{f}_k)_{k\in\mathbb{N}}$ is conditionally unbiased in (8.55), we obtain, for every $k \in \mathbb{N}$,

$$
\begin{aligned}
&\mathbb{E}[F(\mathbf{x}_{k+1}) - F^* | \mathcal{F}_k], \\
&\leq \left(1 + \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}\right) \big(F(\mathbf{x}_k) - F^*\big) \\
&\quad - \alpha_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \nabla F(\mathbf{x}_k) \rangle + \frac{\alpha_k^2 \beta \nu_k^2 b_k}{2} \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2}, \\
&\leq \left(1 + \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}\right) \big(F(\mathbf{x}_k) - F^*\big) \\
&\quad - \alpha_k \mu_k \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 b_k}{2} \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2}, \\
&= \left(1 + \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}\right) \big(F(\mathbf{x}_k) - F^*\big) \\
&\quad - \alpha_k \left(\mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2}\right) \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2} \quad \text{a.s.},
\end{aligned}
\tag{8.59}
$$

where majoration of $\langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \nabla F(\mathbf{x}_k) \rangle$ directly comes from (8.47). Majoration in (8.59) has the same structure as in (8.19), where, for every $k \in \mathbb{N}$,

$$
u_k := \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}, \quad v_k := \alpha_k \left(\mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2}\right), \quad \mathrm{w}_k := \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2}. \tag{8.60}
$$

Moreover, thanks to (8.48) and (8.49), we deduce that $(u_k)_{k\in\mathbb{N}}, (v_k)_{k\in\mathbb{N}}, (\mathrm{w}_k)_{k\in\mathbb{N}}$ are non-negative sequences, and that $\sum_k u_k < +\infty$, $\inf_k v_k > 0$, and $\sum_k \mathrm{w}_k < +\infty$. Hence condition (8.19) is satisfied under condition (i).
Case (ii): Using (8.47) and the Cauchy-Schwarz inequality, we have, for every $k \in \mathbb{N}$,

$$
\begin{aligned}
&\langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbb{E}[\mathbf{f}_k | \mathcal{F}_k] \rangle, \\
&= \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \nabla F(\mathbf{x}_k) \rangle + \Big\langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \Big(\mathbb{E}[\mathbf{f}_k | \mathcal{F}_k] - \nabla F(\mathbf{x}_k) \Big) \Big\rangle, \\
&\geq \mu_k \|\nabla F(\mathbf{x}_k)\|^2 - \big\|\nabla F(\mathbf{x}_k)\big\| \, \Big\| \mathbf{U}_k \Big(\mathbb{E}[\mathbf{f}_k | \mathcal{F}_k] - \nabla F(\mathbf{x}_k) \Big) \Big\|, \\
&\geq \mu_k \|\nabla F(\mathbf{x}_k)\|^2 - \nu_k \|\nabla F(\mathbf{x}_k)\| \, \|\mathbb{E}[\mathbf{f}_k | \mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|.
\end{aligned}
\tag{8.61}
$$

We consider inequality $ab \leq a^2 c + b^2/(4c)$, that holds for any $a \geq 0$, $b \geq 0$ and $c > 0$. Let $k \in \mathbb{N}$, and take $a = \|\nabla F(\mathbf{x}_k)\|$, $b = \|\mathbb{E}[\mathbf{f}_k | \mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|$, $c = (1 - \varrho)\mu_k \nu_k^{-1}$. Then, for every $k \in \mathbb{N}$, we have

$$
\begin{aligned}
&\|\nabla F(\mathbf{x}_k)\| \, \|\mathbb{E}[\mathbf{f}_k | \mathcal{F}_k] - \nabla F(\mathbf{x}_k)\| \\
&\qquad \leq \frac{(1 - \varrho)\mu_k}{\nu_k} \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\nu_k}{4(1 - \varrho)\mu_k} \|\mathbb{E}[\mathbf{f}_k | \mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|^2.
\end{aligned}
\tag{8.62}
$$

Combining this inequality with (8.61) leads to

$$(\forall k \in \mathbb{N}) \quad \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbb{E}[\mathbf{f}_k | \mathcal{F}_k] \rangle \\ \geq \varrho \mu_k \|\nabla F(\mathbf{x}_k)\|^2 - \frac{\nu_k^2}{4(1-\varrho)\mu_k} \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|^2. \quad (8.63)$$

Injecting (8.63) in (8.55), we obtain

$$\mathbb{E}[F(\mathbf{x}_{k+1}) - F^* | \mathcal{F}_k] \\ \leq \Big(1 + \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}\Big)\big(F(\mathbf{x}_k) - F^*\big) - \alpha_k \Big(\varrho \mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2}\Big) \|\nabla F(\mathbf{x}_k)\|^2 \\ + \frac{\alpha_k \nu_k^2}{4(1-\varrho)\mu_k} \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2} \quad \text{a.s..} \quad (8.64)$$

Let, for every $k \in \mathbb{N}$, $u_k = \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}$, $v_k = \alpha_k \Big(\varrho \mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2}\Big)$, and $\mathrm{w}_k = \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2} + \frac{\alpha_k \nu_k^2}{4\mu_k(1-\varrho)} \|\mathbb{E}(\mathbf{f}_k|\mathcal{F}_k) - \nabla F(\mathbf{x}_k)\|^2$.

According to (8.48) (8.50) and (8.51), $(u_k)_{k\in\mathbb{N}}$, $(v_k)_{k\in\mathbb{N}}$, $(\mathrm{w}_k)_{k\in\mathbb{N}}$ are non-negative sequences, such that $\sum_k u_k < +\infty$, $\inf_k v_k > 0$, and $\sum_k \mathrm{w}_k < +\infty$ a.s. Hence (8.19) is satisfied under condition (ii). □

8.5.5 *Application to some state-of-the-art algorithms*

In this section we review a few state-of-the-art SGD algorithms of the form of (8.44), those convergence in a non-convex setting is ensured by Proposition (8.7).

8.5.5.1 *SGD with constant stepsize from [212]*

The algorithm proposed in [212] consists of an SGD scheme of the form of (8.44), with constant stepsize. Precisely, for every $k \in \mathbb{N}$, $\alpha_k = \alpha$, with $\alpha > 0$, and without preconditioning matrix, i.e.,

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\, \mathbf{f}_k. \quad (8.65)$$

The a.s. convergence of the iterates (8.65) can be shown using Proposition 8.7. First, the authors in [212] show that scheme (8.65) verifies a descent condition in a non-necessary convex setting assuming that (8.45) holds with, for every $k \in \mathbb{N}$, $a_k = c_k = 0$ and $b_k = b$, for some $b > 0$. This ensures the convergence of $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$. Second, condition (8.47) is directly satisfied with, for every $k \in \mathbb{N}$, $\mu_k = \nu_k = 1$. Finally, to the extent that the gradient approximations $(\mathbf{f}_k)_{k\in\mathbb{N}}$ are unbiased, condition (i) of Proposition 8.7 holds for $\alpha \in \big(0, 2/(Lb)\big)$. Hence, according to Proposition 8.7, Assumption 8.4 is satisfied and the a.s. convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to a critical point is ensured by Theorem 8.4.

8.5.5.2 *SGD from [71]*

In [71], the authors provide a stochastic version of the forward-backward algorithm for minimizing the sum of a Lipschitz-differentiable function, with constant $\beta > 0$, and a non-necessarily smooth function,

both being convex. Assuming that the non-smooth function is zero, then this algorithm boils down to an SGD algorithm of the form of (8.44), without preconditioning matrices, i.e.,

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \lambda_k \gamma_k \mathbf{f}_k, \tag{8.66}$$

where, for every $k \in \mathbb{N}$, $\lambda_k \in]0, 1]$ and $\gamma_k \in (0, 2/\beta)$. By contrast with the usual SGD algorithms, in [71] the sequence of approximated gradients $(\mathbf{f}_k)_{k\in\mathbb{N}}$ is not supposed to be unbiased, but to verify the following relaxed condition :

$$\sum_{k=0}^{+\infty} \sqrt{\lambda_k} \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\| < +\infty \quad \text{a.s..} \tag{8.67}$$

The a.s. convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to a minimizer of F is ensured by [71, Thm. 2.1], under few technical assumptions (including F convex).

Under the same assumptions, using Proposition 8.7 and Theorem 8.4, we can show that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ a.s. converges to a critical point of F, without needing F to be convex. Indeed, first, [71, Thm. 2.1 - cond. (c)] ensures that (8.45) is satisfied, where $a_k \equiv 0$, $(b_k)_{k\in\mathbb{N}}$ is a bounded sequence, and $(c_k)_{k\in\mathbb{N}}$ satisfies $\sum_{k=0}^{+\infty} \lambda_k c_k < +\infty$. Second, condition (8.47) is satisfied with, for every $k \in \mathbb{N}$, $\mu_k = \nu_k = 1$. Finally, if $\inf_{k\in\mathbb{N}} \lambda_k > 0$, as in [71, Thm. 2.1 - cond. (f)], then (8.67) implies condition (8.50), hence condition (ii) in Proposition 8.7 holds. Consequently, according to Proposition 8.7, Assumption 8.4 is satisfied and the a.s. convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to a critical point is ensured by Theorem 8.4. This result is novel, up to our knowledge. Actually, almost-sure convergence based on sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ for stochastic gradient schemes remains scarcely studied in the non-convex case. Recent results in [105, 57] proposed an alternative proof technique without KL inequality by constraining the topology of the stationary points of F.

8.5.5.3 *Stochastic quasi-Newton with constant stepsize from [162]*

In the case where F is twice differentiable, the author of [162] proposed an alternative version of the usual stochastic Quasi-Newton algorithm [233] of the form

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \eta \mathbf{U}_k \mathbf{f}_k, \tag{8.68}$$

where $\eta > 0$ is a constant stepsize. In [162], the authors show that any process generated by scheme (8.68) satisfies Assumption 8.1. Under the same assumptions, we can show that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ a.s. converges to a critical point of F, using Proposition 8.7 and Theorem 8.4. First, the authors of [233] assume that the gradients $(\mathbf{f}_k)_{k\in\mathbb{N}}$ are unbiased and satisfy (8.45), with, for every $k \in \mathbb{N}$, $a_k = c_k = 0$ and $b_k = b$, for some $b > 0$. Second, condition (8.47) is satisfied with, for every $k \in \mathbb{N}$, $\mu_k = \mu > 0$ and $\nu_k = \nu > 0$, since the preconditioning matrix $\mathbf{U}_k$ is built to approximate the Hessian of F at $\mathbf{x}_k$ and to have a uniformly bounded spectrum. Finally, (i) in Proposition 8.7 is verified when $\alpha \in \big(0, 2\mu/(\beta b \nu^2)\big)$.

8.6 Stochastic proximal-gradient schemes

In this section, we will give examples of stochastic proximal algorithms from the literature satisfying Assumption 8.4, hence benefiting from the convergence guarantees given in Theorem 8.4. Similarly to section 8.5, we will present a general framework for investigating convergence of stochastic proximal schemes in a non-convex setting.

We consider the case where problem (8.1) can be written as

$$\underset{\boldsymbol{x}\in\mathcal{H}}{\text{minimize}}\, F(\boldsymbol{x}) = G(\boldsymbol{x}) + H(\boldsymbol{x}), \tag{8.69}$$

where $F\colon \mathcal{H} \to \mathbb{R}$ verifies Assumption 8.1 and Assumption 8.3, and is expressed as the sum of two functions $G\colon \mathcal{H} \to \mathbb{R}$ and $H\colon \mathcal{H} \to]-\infty, +\infty]$, where G is assumed to be differentiable, and H is assumed to be convex, proper and lower semi-continuous. In this context, a suitable scheme to solve (8.69) is to consider a proximal-gradient algorithm (also known as forward-backward, or ISTA). It alternates, at each iteration, between a gradient-descent step on the differentiable function and a proximal step on the non-smooth function. The proximity operator of a proper, lower semi-continuous, convex function H at a point $\boldsymbol{x} \in \mathcal{H}$ is denoted by $\operatorname{prox}_H(\boldsymbol{x})$, and is defined as the unique minimizer of $H + \frac{1}{2\gamma_k}\|\boldsymbol{x} - \cdot\|^2$ on $\mathcal{H}$ [167, 14].

A stochastic proximal-gradient algorithm can be expressed as

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \left(\mathbf{P}_k(\mathbf{x}_k - \gamma_k \mathbf{g}_k) - \mathbf{x}_k\right), \tag{8.70}$$

where for every $k \in \mathbb{N}$, $\mathbf{g}_k \in \mathcal{H}$ is a stochastic approximation of the gradient of G at $\mathbf{x}_k$, γ_k and λ_k are two positive stepsizes, and $\mathbf{P}_k\colon \mathcal{H} \to \mathcal{H}$ is a stochastic approximation of the proximity operator of H computed at the output of the gradient descent step $(\mathbf{x}_k - \gamma_k \mathbf{g}_k)$.

The stochastic approximations of the proximity operators can be useful in the context when the computation of the proximity operator itself is too demanding, e.g., due to the structure of H (for instance where H is a sum of a very large number of components). A typical example of operators $(\mathbf{P}_k)_{k\in\mathbb{N}}$ is the one encountered in federated algorithms [210]. In this context, H reads as a convex combination $H = \sum_{i=1}^{I} \omega_i H_i$, where $I > 1$, $(\omega_i)_{1\leq i\leq I} \in (0,1)^I$ with $\sum_{i=1}^{I} \omega_i = 1$, and, for every $i \in \{1, \ldots, I\}$, $H_i\colon \mathcal{H} \to]-\infty, +\infty]$. Then, the most common choice consists in adopting, at every iteration $k \in \mathbb{N}$, $\mathbf{P}_k = \sum_{i\in\mathbb{I}_k} \omega_i \operatorname{prox}_{\gamma_k H_i}$, where $\mathbb{I}_k$ is a random subset of $\{1, \ldots, I\}$. Another popular choice of approximation relies on the (deterministic) notion of ϵ-subdifferentiability [211].

Scheme (8.70) can be seen as a stochastic gradient scheme by rewriting it as

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k \lambda_k \mathbf{f}_k, \tag{8.71}$$

where, at each iteration $k \in \mathbb{N}$, $\mathbf{f}_k = \gamma_k^{-1}\left(\mathbf{x}_k - \mathbf{P}_k(\mathbf{x}_k - \gamma_k \mathbf{g}_k)\right)$. Thus, $\mathbf{f}_k$ can be interpreted as an estimation of the gradient of the whole function $\nabla F(\mathbf{x}_k)$. In the literature, this viewpoint is adopted to prove the convergence of some versions of the standard stochastic proximal algorithm (i.e., when, for every $k \in \mathbb{N}$, $\lambda_k = 1$) [166].

8.6.1 Conditions on the approximations

Let $(\mathcal{F}_k)_{k\in\mathbb{N}}$ be the canonical filtration as defined in subsection 8.5.1. Let $\mathbf{y}$ be an integrable random variable from Ω to $\mathcal{H}$. We assume that approximation sequences $(\mathbf{g}_k)_{k\in\mathbb{N}}$ and $\big(\mathbf{P}_k(\mathbf{y})\big)_{k\in\mathbb{N}}$ are conditionally unbiased, i.e.,

$$(\forall k \in \mathbb{N}) \quad \mathbb{E}[\mathbf{g}_k|\mathcal{F}_k] = \nabla G(\mathbf{x}_k) \quad \text{and} \quad \mathbb{E}[\mathbf{P}_k(\mathbf{y}) - \operatorname{prox}_{\gamma_k H}(\mathbf{y})|\mathcal{F}_k] = 0, \tag{8.72}$$

and that there exist two deterministic sequences $(d_k)_{k\in\mathbb{N}}$ and $(e_k)_{k\in\mathbb{N}}$ such that

$$(\forall k \in \mathbb{N}) \quad \mathbb{E}\big[\|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|^2|\mathcal{F}_k\big] \leq d_k\|\nabla F(\mathbf{x}_k)\|^2 + e_k, \tag{8.73}$$

$$\mathbb{E}\big[\|\mathbf{P}_k(\mathbf{y}) - \operatorname{prox}_{\gamma_k H}(\mathbf{y})\|^2|\mathcal{F}_k\big] \leq d_k\|\nabla F(\mathbf{x}_k)\|^2 + e_k. \tag{8.74}$$

8.6.2 A general framework for convergence of differentiable proximal-gradient algorithms in a non-convex context

In this section we introduce a result to enable the convergence study of proximal-gradient based methods, for non-convex minimization problems. In order to use the results developed in section 8.3 and section 8.4, we will consider differentiable functions G and H.

Proposition 8.8. *Let F be of the form of* (8.69)*, where G and H are Lipschitz-differentiable functions, with respective constants $\beta_G > 0$ and $\beta_H > 0$, and H is convex. Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a process generated by* (8.70)*, satisfying conditions* (8.72)*,* (8.73)*, and* (8.74)*, with respect to $(\mathcal{F}_k)_{k\in\mathbb{N}}$. Assume that*

$$\inf_{k\in\mathbb{N}} \lambda_k\gamma_k > 0, \quad \sup_{k\in\mathbb{N}} \lambda_k \leq 1, \quad \sup_{k\in\mathbb{N}} \gamma_k < 1/\beta_H, \quad \sum_{k=0}^{+\infty} e_k^{1/2} < +\infty. \tag{8.75}$$

Let $\beta = \beta_G + \beta_H$, and $(\rho_k)_{k\in\mathbb{N}}$, $(\sigma_k)_{k\in\mathbb{N}}$ be two sequences defined as

$$(\forall k \in \mathbb{N}) \quad \rho_k = (1 - \beta_H\gamma_k)^{-1} \quad \textit{and} \quad \sigma_k = \frac{\gamma_k\rho_k}{2}\Big((\sqrt{2}+1)\beta_H + 4\beta\lambda_k\rho_k\Big), \tag{8.76}$$

with

$$\sup_{k\in\mathbb{N}} \quad \sigma_k + d_k\Big(\sigma_k + \lambda_k\gamma_k^{-1}\beta\Big) < 1. \tag{8.77}$$

Then $(\mathbf{x}_k)_{k\in\mathbb{N}}$ verifies Assumption 8.4.

Proof. To prove that Assumption 8.4 holds, we will use the equivalent form (8.71) of the proximal-gradient scheme (8.70). Hence the proof will be very similar to the one of Proposition 8.7. Precisely, we consider scheme (8.44), where, for every $k \in \mathbb{N}$, $\alpha_k = \gamma_k\lambda_k$ and $\mathbf{U}_k = \boldsymbol{I}$, and the sequence $(\mathbf{f}_k)_{k\in\mathbb{N}}$ is an approximation of $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$ given by

$$(\forall k \in \mathbb{N}) \quad \mathbf{f}_k = \mathbf{d}_k + \Delta_k, \tag{8.78}$$

$$\mathbf{d}_k = \gamma_k^{-1}\left(\mathbf{x}_k - \operatorname{prox}_{\gamma_k H}(\mathbf{x}_k - \gamma_k\mathbf{g}_k)\right), \tag{8.79}$$

$$\Delta_k = \gamma_k^{-1}\left(\operatorname{prox}_{\gamma_k H}(\mathbf{x}_k - \gamma_k\mathbf{g}_k) - \mathbf{P}_k(\mathbf{x}_k - \gamma_k\mathbf{g}_k)\right). \tag{8.80}$$

The first step of the proof consists in giving a majoration of the approximated gradient sequence $(\mathbf{f}_k)_{k\in\mathbb{N}}$ as a function of the true gradient and the error terms involved in the assumptions.

Since H is differentiable, according to the Fermat's rule [14] we have

$$(\mathbf{x}_k - \gamma_k \mathbf{g}_k) - \operatorname{prox}_{\gamma_k H}(\mathbf{x}_k - \gamma_k \mathbf{g}_k) = \gamma_k \nabla H\left(\operatorname{prox}_{\gamma_k H}(\mathbf{x}_k - \gamma_k \mathbf{g}_k)\right). \tag{8.81}$$

Combining (8.79) and (8.81), it follows that

$$\mathbf{d}_k = \mathbf{g}_k + \nabla H\left(\operatorname{prox}_{\gamma_k H}(\mathbf{x}_k - \gamma_k \mathbf{g}_k)\right) = \mathbf{g}_k + \nabla H(\mathbf{x}_k - \gamma_k \mathbf{d}_k). \tag{8.82}$$

Since $\nabla F = \nabla G + \nabla H$, using the triangular inequality, and the fact that H has a β_H-Lipschitzian gradient

$$\begin{aligned} \|\mathbf{d}_k\| &= \|\nabla F(\mathbf{x}_k) + (\mathbf{g}_k - \nabla G(\mathbf{x}_k)) + \nabla H(\mathbf{x}_k - \gamma_k \mathbf{d}_k) - \nabla H(\mathbf{x}_k)\|, \\ &\leq \|\nabla F(\mathbf{x}_k)\| + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\| + \beta_H \gamma_k \|\mathbf{d}_k\|. \end{aligned} \tag{8.83}$$

According to the definition of ρ_k in (8.76), this leads to

$$\|\mathbf{d}_k\| \leq \rho_k \left(\|\nabla F(\mathbf{x}_k)\| + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|\right). \tag{8.84}$$

Combining (8.78) and (8.84) gives

$$\|\mathbf{f}_k\| \leq \rho_k \left(\|\nabla F(\mathbf{x}_k)\| + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|\right) + \|\Delta_k\|. \tag{8.85}$$

Moreover, passing to the square, we also deduce that

$$\|\mathbf{f}_k\|^2 \leq 4\rho_k^2 \left(\|\nabla F(\mathbf{x}_k)\|^2 + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|^2\right) + 2\|\Delta_k\|^2. \tag{8.86}$$

The second step of the proof aims at establishing stochastic descent properties for the scheme at stake, to show that conditions (8.19) and (8.20) hold.

First, we apply the descent Lemma [20] to G, using the fact that $\mathbf{x}_{k+1} - \mathbf{x}_k = -\gamma_k \mathbf{f}_k$,

$$G(\mathbf{x}_{k+1}) \leq G(\mathbf{x}_k) - \gamma_k \lambda_k \langle \nabla G(\mathbf{x}_k) \mid \mathbf{f}_k \rangle + \frac{\beta_G \gamma_k^2 \lambda_k^2}{2} \|\mathbf{f}_k\|^2, \tag{8.87}$$

Then, adding $H(\mathbf{x}_{k+1})$ on both sides of (8.87) gives

$$F(\mathbf{x}_{k+1}) \leq G(\mathbf{x}_k) - \gamma_k \lambda_k \langle \nabla G(\mathbf{x}_k) \mid \mathbf{f}_k \rangle + H(\mathbf{x}_{k+1}) + \frac{\beta_G \gamma_k^2 \lambda_k^2}{2} \|\mathbf{f}_k\|^2. \tag{8.88}$$

Since H is convex and differentiable, using again $\mathbf{x}_{k+1} - \mathbf{x}_k = -\gamma_k \mathbf{f}_k$, we also have

$$H(\mathbf{x}_{k+1}) \leq H(\mathbf{x}_k) - \gamma_k \lambda_k \langle \nabla H(\mathbf{x}_{k+1}) \mid \mathbf{f}_k \rangle. \tag{8.89}$$

Combining (8.88) and (8.89) gives a descent inequality, given by, for every $k \in \mathbb{N}$,

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \left(\gamma_k \lambda_k \langle \nabla G(\mathbf{x}_k) + \nabla H(\mathbf{x}_{k+1}) \mid \mathbf{f}_k \rangle - \frac{\beta_G \gamma_k^2 \lambda_k^2}{2} \|\mathbf{f}_k\|^2\right). \tag{8.90}$$

Note that, for every $k \in \mathbb{N}$, we have

$$\begin{aligned}
&\gamma_k\lambda_k\langle \nabla G(\mathbf{x}_k) + \nabla H(\mathbf{x}_{k+1}) \mid \mathbf{f}_k\rangle,\\
&\quad = \gamma_k\lambda_k\langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k\rangle - \gamma_k\lambda_k\langle \nabla H(\mathbf{x}_k) - \nabla H(\mathbf{x}_{k+1}) \mid \mathbf{f}_k\rangle,\\
&\quad \geq \gamma_k\lambda_k\langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k\rangle - \gamma_k\lambda_k\big\|\nabla H(\mathbf{x}_k) - \nabla H(\mathbf{x}_{k+1})\big\|\,\|\mathbf{f}_k\|, \qquad (8.91)\\
&\quad \geq \gamma_k\lambda_k\langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k\rangle - \frac{\beta_H\gamma_k^2\lambda_k^2}{2}\|\mathbf{f}_k\|^2, \qquad (8.92)
\end{aligned}$$

where (8.91) is obtained using the Cauchy-Schwarz inequality, and (8.92) using the Lipschitz continuity of ∇H and (8.71). Hence, since $\beta = \beta_G + \beta_H$, we have

$$\gamma_k\lambda_k\langle \nabla G(\mathbf{x}_k) + \nabla H(\mathbf{x}_{k+1}) \mid \mathbf{f}_k\rangle - \frac{\beta_G\gamma_k^2\lambda_k^2}{2}\,\|\mathbf{f}_k\|^2 \geq \gamma_k\lambda_k\langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k\rangle - \frac{\gamma_k^2\lambda_k^2\beta}{2}\|\mathbf{f}_k\|^2. \qquad (8.93)$$

On the one hand, using (8.86), we obtain

$$-\ \frac{\gamma_k^2\lambda_k^2\beta}{2}\|\mathbf{f}_k\|^2 \quad \geq \quad -\gamma_k^2\lambda_k^2\beta\Big(\|\Delta_k\|^2 \ +\ 2\rho_k^2\Big(\|\nabla F(\mathbf{x}_k)\|^2 \ +\ \|\mathbf{g}_k \ -\ \nabla G(\mathbf{x}_k)\|^2\Big)\Big). \qquad (8.94)$$

On the other hand, using the definition of $\mathbf{f}_k$ in (8.78), equation (8.82), and the fact that $\nabla F = \nabla G + \nabla H$, we have

$$\begin{aligned}
\langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k\rangle &= \langle \nabla F(\mathbf{x}_k) \mid \mathbf{d}_k + \Delta_k\rangle,\\
&= \langle \nabla F(\mathbf{x}_k) \mid \nabla F(\mathbf{x}_k) - \nabla G(\mathbf{x}_k) - \nabla H(\mathbf{x}_k) + \mathbf{g}_k + \nabla H\big(\mathbf{x}_k - \gamma_k\mathbf{d}_k\big) + \Delta_k\rangle,\\
&= \|\nabla F(\mathbf{x}_k)\|^2 - \big\langle \nabla F(\mathbf{x}_k) \mid \nabla H(\mathbf{x}_k) - \nabla H\big(\mathbf{x}_k - \gamma_k\mathbf{d}_k\big)\big\rangle\\
&\qquad + \langle \nabla F(\mathbf{x}_k) \mid \mathbf{g}_k - \nabla G(\mathbf{x}_k) + \Delta_k\rangle. \qquad (8.95)
\end{aligned}$$

Using the Cauchy-Schwarz inequality and the Lipschitz continuity of ∇H, we obtain

$$\begin{aligned}
\langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k\rangle &\geq \|\nabla F(\mathbf{x}_k)\|^2 - \beta_H\gamma_k\|\nabla F(\mathbf{x}_k)\|\|\mathbf{d}_k\|,\\
&\qquad + \langle \nabla F(\mathbf{x}_k) \mid \mathbf{g}_k - \nabla G(\mathbf{x}_k) + \Delta_k\rangle,\\
&\geq \|\nabla F(\mathbf{x}_k)\|^2 - \beta_H\gamma_k\rho_k\|\nabla F(\mathbf{x}_k)\|\Big(\|\nabla F(\mathbf{x}_k)\| + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|\Big)\\
&\qquad + \langle \nabla F(\mathbf{x}_k) \mid \mathbf{g}_k - \nabla G(\mathbf{x}_k) + \Delta_k\rangle, \qquad (8.96)\\
&= \big(1 - \beta_H\gamma_k\rho_k\big)\|\nabla F(\mathbf{x}_k)\|^2 - \beta_H\gamma_k\rho_k\|\nabla F(\mathbf{x}_k)\|\,\|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|\\
&\qquad + \langle \nabla F(\mathbf{x}_k) \mid \mathbf{g}_k - \nabla G(\mathbf{x}_k) + \Delta_k\rangle. \qquad (8.97)
\end{aligned}$$

where (8.96) is obtained using (8.84). For any $(a, b, c) \in \mathbb{R}^3_+$, we have $ab \leq ca^2 + b^2/(4c)$. Taking $a = \|\nabla F(\mathbf{x}_k)\|, b = \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|, c = (\sqrt{2} - 1)/2$, we obtain

$$\begin{aligned}
\langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k\rangle \geq \Big(1 - \beta_H\gamma_k\rho_k\frac{\sqrt{2}+1}{2}\Big)\|\nabla F(\mathbf{x}_k)\|^2&\\
- \beta_H\gamma_k\rho_k\frac{\sqrt{2}+1}{2}\|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|^2 + \langle \nabla F(\mathbf{x}_k) \mid \mathbf{g}_k - \nabla G(\mathbf{x}_k) + \Delta_k\rangle.& \qquad (8.98)
\end{aligned}$$

Combining (8.93), (8.94), and (8.98), we obtain

$$\begin{aligned}
&\gamma_k\lambda_k\big\langle\nabla G(\mathbf{x}_k)+\nabla H\big(\mathbf{x}_{k+1}\big)\mid\mathbf{f}_k\big\rangle-\frac{\beta_G\gamma_k^2\lambda_k^2}{2}\,\|\mathbf{f}_k\|^2,\\
&\qquad\geq\gamma_k\lambda_k\Big(1-\beta_H\gamma_k\rho_k\frac{\sqrt{2}+1}{2}\Big)\|\nabla F(\mathbf{x}_k)\|^2-\beta_H\lambda_k\gamma_k^2\rho_k\frac{\sqrt{2}+1}{2}\|\mathbf{g}_k-\nabla G(\mathbf{x}_k)\|^2\\
&\qquad+\gamma_k\lambda_k\langle\nabla F(\mathbf{x}_k)\mid\mathbf{g}_k-\nabla G(\mathbf{x}_k)+\Delta_k\rangle\\
&\qquad-\gamma_k^2\lambda_k^2\beta\bigg(\|\Delta_k\|^2+2\rho_k^2\Big(\|\nabla F(\mathbf{x}_k)\|^2+\|\mathbf{g}_k-\nabla G(\mathbf{x}_k)\|^2\Big)\bigg).
\end{aligned}\tag{8.99}$$

Using the definition of $(\sigma_k)_{k\in\mathbb{N}}$ given in (8.76) we thus obtain

$$\begin{aligned}
\gamma_k\lambda_k\langle\nabla G(\mathbf{x}_k)+\nabla H(\mathbf{x}_{k+1})\mid\mathbf{f}_k\rangle-\frac{\beta_G\gamma_k^2\lambda_k^2}{2}\|\mathbf{f}_k\|^2,&\\
\geq\gamma_k\lambda_k\left(1-\sigma_k\right)\|\nabla F(\mathbf{x}_k)\|^2-\gamma_k\lambda_k\sigma_k\|\mathbf{g}_k-\nabla G(\mathbf{x}_k)\|^2&\\
-\gamma_k^2\lambda_k^2\beta\|\Delta_k\|^2+\gamma_k\lambda_k\langle\nabla F(\mathbf{x}_k)\mid\mathbf{g}_k-\nabla G(\mathbf{x}_k)+\Delta_k\rangle.&
\end{aligned}\tag{8.100}$$

Since, for every $k\in\mathbb{N}$, $\|\nabla F(\mathbf{x}_k)\|^2$ is $\mathcal{F}_k$-measurable, passing to conditional expectation in (8.100) and using successively (8.72), (8.73) and (8.74) gives

$$\begin{aligned}
&\mathbb{E}\left[\gamma_k\lambda_k\langle\nabla G(\mathbf{x}_k)+\nabla H(\mathbf{x}_{k+1})\mid\mathbf{f}_k\rangle-\frac{\beta_G\gamma_k^2\lambda_k^2}{2}\|\mathbf{f}_k\|^2\;\big|\mathcal{F}_k\right],\\
&\quad\geq\gamma_k\lambda_k\left(1-\sigma_k\right)\|\nabla F(\mathbf{x}_k)\|^2-\gamma_k\lambda_k\sigma_k\;\mathbb{E}\big[\|\mathbf{g}_k-\nabla G(\mathbf{x}_k)\|^2\;\big|\mathcal{F}_k\big]-\gamma_k^2\lambda_k^2\beta\;\mathbb{E}\big[\|\Delta_k\|^2\;\big|\mathcal{F}_k\big],\\
&\quad\geq\gamma_k\lambda_k\left(1-\sigma_k\right)\|\nabla F(\mathbf{x}_k)\|^2-\gamma_k\lambda_k\sigma_k\Big(d_k\|\nabla F(\mathbf{x}_k)\|^2+e_k\Big)-\lambda_k^2\beta\Big(d_k\|\nabla F(\mathbf{x}_k)\|^2+e_k\Big),\\
&\quad=\gamma_k\lambda_k\Big(1-\sigma_k\big(1+d_k\big)-\lambda_k\gamma_k^{-1}\beta d_k\Big)\|\nabla F(\mathbf{x}_k)\|^2-\gamma_k\lambda_k e_k\Big(\sigma_k+\lambda_k\gamma_k^{-1}\beta\Big)\;\text{ a.s..}
\end{aligned}\tag{8.101}$$

Taking the conditional expectation in (8.90), combining it with (8.101), we obtain

$$\begin{aligned}
\mathbb{E}\big[F(\mathbf{x}_{k+1})\;\big|\mathcal{F}_k\big]\leq F(\mathbf{x}_k)-\gamma_k\lambda_k\Big(1-\sigma_k\big(1+d_k\big)-\lambda_k\gamma_k^{-1}\beta d_k\Big)\|\nabla F(\mathbf{x}_k)\|^2&\\
+\gamma_k\lambda_k e_k\Big(\sigma_k+\lambda_k\gamma_k^{-1}\beta\Big)\;\text{ a.s..}&
\end{aligned}\tag{8.102}$$

Let, for every $k\in\mathbb{N}$, $u_k=0$, $v_k=\gamma_k\lambda_k\Big(1-\sigma_k\big(1+d_k\big)-\lambda_k\gamma_k^{-1}\beta d_k\Big)$, and $\mathrm{w}_k=\gamma_k\lambda_k e_k\Big(\sigma_k+\lambda_k\gamma_k^{-1}\beta\Big)$. We have $\sum_k u_k<+\infty$, according to (8.75) and (8.77) we have $\inf_k v_k>0$, and $\sum_k \mathrm{w}_k<+\infty$. Then, inequality (8.102) is of the form of (8.19).

It remains to show that (8.20) holds. According to (8.71) and (8.85), and using Jensen's inequality, for every $k\in\mathbb{N}$, we have almost-surely

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{x}_{k+1}-\mathbf{x}_k\|\;|\mathcal{F}_k],\\
&\quad\leq\gamma_k\lambda_k\rho_k\Big(\|\nabla F(\mathbf{x}_k)\|+\mathbb{E}\big[\|\mathbf{g}_k-\nabla G(\mathbf{x}_k)\|\;\big|\mathcal{F}_k\big]+\gamma_k\lambda_k\mathbb{E}\big[\|\Delta_k\|\;\big|\mathcal{F}_k\big],\\
&\quad\leq\gamma_k\lambda_k\rho_k\Big(\|\nabla F(\mathbf{x}_k)\|+\sqrt{\mathbb{E}\big[\|\mathbf{g}_k-\nabla G(\mathbf{x}_k)\|^2\;\big|\mathcal{F}_k\big]}\Big)+\gamma_k\lambda_k\sqrt{\mathbb{E}\big[\|\Delta_k\|^2\;\big|\mathcal{F}_k\big]}.
\end{aligned}\tag{8.103}$$

According to conditions (8.73) and (8.74), we then obtain

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{x}_{k+1}-\mathbf{x}_k\| \ |\mathcal{F}_k],\\
&\leq \gamma_k\lambda_k\rho_k\left(\|\nabla F(\mathbf{x}_k)\|+\sqrt{d_k\|\nabla F(\mathbf{x}_k)\|^2+e_k}\right)+\gamma_k\lambda_k\sqrt{d_k\|\nabla F(\mathbf{x}_k)\|^2+e_k},\\
&\leq \gamma_k\lambda_k\left(\rho_k+\rho_k d_k^{1/2}+d_k^{1/2}\right)\|\nabla F(\mathbf{x}_k)\|+\gamma_k\lambda_k(\rho_k+1)e_k^{1/2} \quad \text{a.s.},
\end{aligned} \tag{8.104}$$

where the last majoration is obtained using identity $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$, that holds for any $a,b\geq 0$. Let, for every $k\in\mathbb{N}$, $r_k=0$, $s_k=\gamma_k\lambda_k\left(\rho_k+\rho_k d_k^{1/2}+d_k^{1/2}\right)$, and $t_k=\gamma_k\lambda_k(\rho_k+1)e_k^{1/2}$. We have $\sum_k r_k<+\infty$ and, according to (8.75) and (8.76), $\sum_k t_k<+\infty$. Hence, (8.104) is of the form of condition (8.20).

Hence we conclude that Assumption 8.4 holds.

□

8.6.3 Application to some state-of-the-art algorithms

The stochastic aspect of process (8.70) enables two types of uncertainties. First, it takes into account uncertainty arising from stochastic approximations of the gradient of G. Second, it can handle uncertainty resulting from the use of an approximation of the proximity operator of H. The combination of these two types of uncertainties in a single scheme has been taken into account in some works in the literature.

8.6.3.1 *Stochastic FB scheme*

The authors of [70, 71] show the a.s. convergence of the iterates of a stochastic FB scheme of the form of (8.70), assuming that G is convex, and considering the following error on true proximity operator, i.e.,

$$(\forall k\in\mathbb{N})(\forall \boldsymbol{u}\in\mathcal{H})\quad \mathbf{P}_k(\boldsymbol{u})=\operatorname{prox}_{\gamma_k H_k}(\boldsymbol{u})+\mathbf{p}_k, \tag{8.105}$$

where $(H_k)_{k\in\mathbb{N}}$ are successive convex approximations of H and $(\mathbf{p}_k)_{k\in\mathbb{N}}$ is a conditionally summable sequence. The latter study generalises those of [205] for which no error term on the proximal operator was considered.

It is worth noticing that, up to our knowledge, the a.s. convergence of the iterates when G is non-convex has only been studied when $\lambda_k=1$ and without any error on the proximal mapping [107, 151], i.e.,

$$(\forall k\in\mathbb{N})(\forall \boldsymbol{u}\in\mathcal{H})\quad \mathbf{P}_k(\boldsymbol{u})=\operatorname{prox}_{\gamma_k H}(\boldsymbol{u}). \tag{8.106}$$

Hence, the general results we presented in subsection 8.6.2 ensuring the a.s. convergence of the iterates when G is non-convex, and allowing stochastic approximations of the proximity operator of H, appear to be new.

8.6.3.2 *Stochastic proximal scheme*

Let, for every $k\in\mathbb{N}$, $\lambda_k=1$, $\gamma_k=1$, and let $G=0$, and $H=\sum_{i=1}^I\omega_i H_i$, where $(\omega_i)_{1\leq i\leq I}\in]0,1[^I$ with $\sum_{i=1}^I\omega_i=1$ and, for every $i\in\{1,\dots,I\}$, H_i Lipschitz-differentiable. Then scheme (8.70) boils

down to a stochastic proximal algorithm of the form of

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{P}_k(\mathbf{x}_k). \tag{8.107}$$

Such algorithm is very similar to the federated-prox algorithm introduced in [69, 210]. In particular, in [69, 210], the authors propose to choose

$$(\forall k \in \mathbb{N})(\forall \boldsymbol{u} \in \mathcal{H}) \quad \mathbf{P}_k(\boldsymbol{u}) = \mathrm{prox}_{\sum_{i \in \mathbb{S}_k} \omega_i H_i}(\boldsymbol{u}), \tag{8.108}$$

where, for every $k \in \mathbb{N}$, $\mathbb{S}_k$ is a subset of $\{1, \dots, I\}$. The authors show that, assuming the functions $(H_i)_{1 \leq i \leq I}$ satisfy some bounded dissimilarity condition, the expectation of the global objective function conditionally to the subsets of indices, decreases at each iteration.

The results presented in subsection 8.6.2 provide stronger convergence guaranties than in [210], at the price of slightly stronger assumptions on the choice of $(\mathbf{P}_k)_{k \in \mathbb{N}}$.

Chapter 9

Conclusion

9.1 Summary of contributions

This thesis has been devoted to extent the knowledge of quadratic MM algorithms, both theoretically and numerically, particularly when facing large-scale problems. Our developments in the stochastic framework have also led us to move beyond this initial topic and to address more general schemes for non-convex stochastic optimization.

Chapters 2 and 6, relatively similar in their structure, were intended to offer the reader a pedagogical approach to differentiable optimization in the deterministic and the then stochastic settings. Our objectives were to describe and explain various tools that we have used over this PhD to carry out our various theoretical investigations. Our contributions in these two chapters are therefore didactic in nature. Although the scientific content presented was already available in the literature, we have endeavoured to redefine certain implicitly known notions, in particular that of descent (deterministic or stochastic) conditions, in order to provide a more versatile strategy for studying optimization algorithms asymptotically.

The basis for our work on quadratic MM (QMM) algorithms was introduced in Chapter 3 through a general overview on the existing methods adopting such a principle. Two points in particular were highlighted. The first one concerns the flexibility of these algorithms from a structural point of view as they can be redesigned to incorporate acceleration strategies, remaining essential to deal with very high dimensional problems. The second one emphasises their main advantage, namely their natural stability independently of any convexity assumption, and which subsequently motivated a great part of this PhD work. Finally, it should also be noted that some results, although sometimes already used in the past, are explained and demonstrated for the first time in this chapter. In particular for the first time, we exhibited a sufficient and necessary condition for the existence of quadratic tangent majorization approximations, and a justification enabling the use of subspace quadratic MM (SQMM) scheme.

Chapter 4 and 5 naturally follow Chapter 3. They constitute our main contributions on MM algorithms to handle large-scale optimization problems. Chapter 4 aimed to formalize the convergence analysis of a block version of the SQMM algorithm, named B2MS. In particular, we use the strategy of proof introduced in Chapter 2 based on the Kurdyka-Łojasiewciz theory to overcome the non-convexity

of the cost function. In Chapter 5, we proposed a distributed version of the B2MS scheme under the acronym BD3MG. This enabled asynchronous data processing by dividing tasks between several machine cores without wasting any information. On the one hand, we were able to combine theoretical results from Chapter 2 and 3 with those of [76] to prove the global convergence of the latter algorithm and to exhibit some convergence rates always in a non-convex framework. On the other hand, several numerical experiments in a large-scale context, were conducted to show the interest of our method for image reconstruction, both on academic data and on real samples, coming from two-photons excitation microscopy acquisitions.

In Chapter 7, we introduced the scheme SABRINA as a new stochastic extension of the SQMM scheme. The core of our theoretical analysis presented was the establishment of a quasi-supermartingale simple descent condition verified by any Lipschitz continuous gradient but non-necessary convex cost function. We also provided more precise results under additional assumptions and especially, a convergence rate in the strongly-convex case. In order to illustrate the numerical interest of our work, we also presented two sets of numerical experiments. The first one, in line with statistical learning, presents an academical supervised binary classification problem while, the second one, about blur identification, lies within the field of inverse problems for image processing.

Our last contribution, presented in Chapter 8, stepped beyond the MM framework by proposing a new methodology for analysing the convergence of stochastic algorithms for non-convex problems. More specifically, by means of a new probabilistic version of the uniform Kurdyka-Łojasiewicz inequality, our proof strategy can be seen as a generalization of the one presented in Chapter 2 section 2.5.2 for the deterministic setting. Chapter 8 gathered our most recent work, and we hope it will opens several ways for improvement. Especially, its last two sections constituted a preliminary benchmark, identifying several algorithms in the literature for which our proof methodology could be applied.

9.2 Perspectives

In view of the works presented throughout this manuscript, a number of promising avenues of exploration are conceivable and discussed hereafter.

Relaxing the stepsize in SABRINA algorithm

Because of its structure, the SABRINA scheme can be considered as a stochastic gradient extension which would have been conditioned by a majorization matrix sequence $(\boldsymbol{A}_k)_{k\in\mathbb{N}}$ combined with a stepsize incorporation. The convergence of the latter to zero acts as an additional condition and is imposed in order to control the noise variance and thus to ensure a certain robustness of the resulting algorithm. In our case, however, this constraint tends to dominate the MM aspect of our scheme to the point of almost completely masking it during the asymptotic analysis phase. The convergence of the stepsize to zero makes it easier to obtain classical convergence results via the descent condition, but without really using the majorizing character of $(\boldsymbol{A}_k)_{k\in\mathbb{N}}$ (actually, we only used the fact that it has a uniform bounded spectrum). One possibility would thus be to relax the stepsize constraint by typically assuming that it is bounded. Obtaining behaviours such as convergence of the gradient to zero in the sense of (at least) a subsequence would be trickier to obtain, but would undoubtedly make full use of the majorization property. This would be an advantage over the usual stochastic gradient methods,

which generally do not easily allow to work with such a flexible stepsize, and it may be possible, in the longer term, to obtain convergence rates even in a non-convex setting. The difficulty that would be added to this relaxation would then reside in the nature of the noise likely to affect the gradient. This leads to the new problematic: will the only majorization property be enough to ensure similar or even better convergence guarantees without restricting the nature of the noise ?

Combining BD3MG and SABRINA to create a new distributed stochastic algorithm.

The algorithms we have presented and analyzed in this manuscript can be divided according to whether or not they are stochastic or asynchronous. SABRINA algorithm presented in Chapter 7 bridges a first gap between the usual deterministic method and stochastic approximation algorithms. However, its structure does not easily allow for a distributed implementation. Let us stress out that our developments on SABRINA method were done in the beginning of this thesis, and no asynchronous extension was envisioned at that stage. In view of our recent progresses around BD3MG, we feel that it would be feasible to propose an asynchronous version of SABRINA, in the spirit of the stochastic PALM [76]. On a theoretical point of view, the proof for stochastic descent condition may be similar of the one used in [76] and the majorization property might serve to build a relaxed assumption either on the cost function or on the statistics of the noise. As a second step and for deeper investigation, one might consider the proof methodology from Chapter 8, to obtain almost-sure convergence to a stationary point. Of course, this future work would greatly depend on the feasible progresses around the novel stochastic KL theory.

Investigating alternative assumptions for convergence of stochastic algorithms in a non-convex setting.

In a very general way, the cornerstone of Chapter 8 is the construction of a new method to improve convergence guarantees of stochastic process $(\mathbf{x}_k)_{k\in\mathbb{N}}$, supposed well-built enough to verify elementary asymptotical properties, as the convergence of its gradient to zero or those of $F(\mathbf{x}_k)_{k\in\mathbb{N}}$ to a finite limit. In this respect, we felt it was particularly appropriate to work on a new version of the KL property, as it is currently one of the most highly developed tools in the deterministic literature for overcoming non-convexity obstacles. The stochastic extension we are proposing allows to obtain a convergence theorem for iterates, albeit under (almost-sure) monotonicity assumptions which may appear relatively restrictive, particularly in the stochastic framework. Such limitations are from technical order and are directly related to the inner structure of KL inequality. More precisely the non-derivability of the φ function at zero reveals to be a very challenging obstacle. It typically prevents from building a concave extension of φ over all $\mathbb{R}$, that would make it possible to remove the monotonicity conditions we were forced to impose. As things stand, it is difficult to say whether this non-derivability constraint can be lifted, as this would require further developments of the KL theory itself beyond an optimization framework. If results were to be obtained in this direction, the monotonicity assumptions could be attenuated or even eliminated, and a convergence theorem will be obtained that is even more generic, and ultimately applicable to a larger number of stochastic algorithms especially SABRINA or even one of its eventual asynchronous extension.

Investigating the non-differentiable case.

It should also be noted that, in this thesis, all our results have been established in a differentiable optimization framework. However, it turns out that the original KL theory, in this sense, is more flexible because it allows us to work not only with gradients but more generally with subdifferential [27] sets. This aspect could not be addressed in depth during this work but, at first sight, it appears to be relatively accessible and does not seem to raise major technical obstacles. The resulting convergence theorems could thus be applied to algorithms of the proximal stochastic type or even more generally employing monotone operators [70].

List of Figures

List of Tables

Bibliography

[1] F. Abboud, M. Stamm, E. Chouzenoux, J.-C. Pesquet, and H. Talbot. Distributed algorithms for proximity operator computation with applications to video processing. *Digital Signal Processing*, 128:103610, Aug. 2022.

[2] P.-A. Absil and K. A. Gallivan. Accelerated line-search and trust-region methods. *SIAM Journal on Numerical Analysis*, 47(2):997–1018, 2009.

[3] P.-A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

[4] Ö. D. Akyildiz, E. Chouzenoux, V. Elvira, and J. Míguez. A probabilistic incremental proximal gradient method. *IEEE Signal Processing Letters*, 26(8):1257–1261, 2019.

[5] M. Allain, J. Idier, and Y. Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Transactions on Image Processing*, 15(5):1130–1142, 2006.

[6] L. Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

[7] Y. F. Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *Journal on Machine Learning Research*, 18:1–33, 2017.

[8] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116:5–16, 2009.

[9] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[10] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

[11] A. Auslender. Noncoercive optimization problems. *Mathematics of Operations Research*, 21(4):769–782, 1996.

[12] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

[13] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS 2011)*, pages x–x+8, Granada, Spain, Dec. 12 - 17 2011.

[14] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011.

[15] S. Becker and Y. Le Cun. Improving the convergence of back-propagation learning with second-order methods. In *Proceedings of the 1988 Connectionist Models Summer School*, 1988.

[16] R. Bellman and R. Kalaba. A mathematical theory of adaptive control processes. *Proceedings of the National Academy of Sciences*, 45(8):1288–1290, 1959.

[17] M. Bertero. *Introduction to Inverse Problems in Imaging*. CRC Press, 2017.

[18] D. P. Bertsekas. Gradient convergence in gradient methods. Technical report, 1997. https://core.ac.uk/download/pdf/4381121.pdf.

[19] D. P. Bertsekas. *Nonlinear Programming*. Taylor & Francis, 1997.

[20] D. P. Bertsekas and J. N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[21] E. Bierstone and P. D. Milman. Semianalytic and subanalytic sets. *Publications Mathématiques de l'IHÉS*, 67:5–42, 1988.

[22] P. Billingsley. *Probability and Measure*. John Wiley & Sons, 2008.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Real Algebraic Geometry*, volume 36. Springer Science & Business Media, 2013.

[24] D. Böhning and B. G. Lindsay. Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663, 1988.

[25] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of lojasiewicz inequalities and applications. *arXiv preprint arXiv:0802.0826*, 2008.

[26] J. Bolte and E. Pauwels. Majorization-minimization procedures and convergence of sqp methods for semi-algebraic and tame programs. *Mathematics of Operations Research*, 41(2):442–465, 2016.

[27] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[28] S. Bonettini, F. Porta, M. Prato, S. Rebegoldi, V. Ruggiero, and L. Zanni. *Recent Advances in Variable Metric First-Order Methods*, pages 1–31. Springer International Publishing, Cham, 2019.

[29] S. Bonettini, M. Prato, and S. Rebegoldi. A block coordinate variable metric linesearch based proximal gradient method. *Computational Optimization and Applications*, 71(1):5–52, 2018.

[30] S. Bonettini, M. Prato, and S. Rebegoldi. Convergence of inexact forward–backward algorithms using the forward–backward envelope. *SIAM Journal on Optimization*, 30(4):3069–3097, 2020.

[31] S. Bonettini, M. Prato, and S. Rebegoldi. New convergence results for the inexact variable metric forward–backward method. *Applied Mathematics and Computation*, 392:125719, 2021.

[32] J.-F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical Optimization: Theoretical and Practical Aspects*. Springer Science & Business Media, 2006.

[33] A. Bordes, L. Bottou, and P. Gallinari. Sgd-qn: Careful quasi-Newton stochastic gradient descent. *Journal of Machine Learning Research*, 10:1737–1754, 2009.

[34] R. I. Boţ and E. R. Csetnek. An inertial tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems. *Journal of Optimization Theory and Applications*, 171(2):600–616, 2016.

[35] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of International Conference on Computational Statistics (COMPSTAT 2010)*, pages 177–186. Springer, 2010.

[36] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[37] G. Bouchard. Efficient bounds for the softmax function and applications to approximate inference in hybrid models. In *Proceedings of the Neural Information Processing Systems (NIPS 2008)*, volume 31, 2008.

[38] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[39] L. M. Briceño-Arias, G. Chierchia, E. Chouzenoux, and J.-C. Pesquet. A random block-coordinate douglas–rachford splitting method with low computational complexity for binary logistic regression. *Computational Optimization and Applications*, 72(3):707–726, 2019.

[40] C. G. Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[41] R. H. Byrd, G. M. Chin, W. Neveitt, and J. Nocedal. On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, 21(3):977–995, 2011.

[42] R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

[43] S. Cadoni, E. Chouzenoux, J.-C. Pesquet, and C. Chaux. A block parallel majorize-minimize memory gradient algorithm. In *23rd IEEE Int. Conf. Image Process. (ICIP 2016)*, pages 3194–3198, Phoenix, AZ, 25-28 Sep. 2016.

[44] E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[45] L. Cannelli, F. Facchinei, G. Scutari, and V. Kungurtsev. Asynchronous optimization over graphs: Linear convergence under error bound conditions. *IEEE Transactions on Automatic Control*, 66(10):4604–4619, 2020.

[46] J. W. Cantrell. Relation between the memory gradient method and the Fletcher-Reeves method. *Journal of Optimization Theory and Applications*, 4(1):67–71, 1969.

[47] C. Castera, J. Bolte, C. Fevotte, and E. Pauwels. An inertial Newton algorithm for deep learning. Technical report, 2019. https://arxiv.org/abs/1905.12278.

[48] A. Chakrabarti, T. Zickler, and W. T. Freeman. Analyzing spatially-varying blur. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, pages 2512–2519, San Francisco, CA, USA, 13-18 June 2010.

[49] M. Chalvidal and E. Chouzenoux. Block distributed 3MG algorithm and its application to 3D image restoration. In *Proceedings of the IEEE International Conference on Image Processing (ICIP 2020)*, pages 938–942, Abu Dhabi, United Arab Emirates (virtual), 25-28 Oct. 2020.

[50] M. Chalvidal and E. Chouzenoux. Python toolbox for block distributed majorize-minimize memory gradient algorithm, 2022. https://github.com/mathieuchal/BD3MG.

[51] M. Chalvidal, E. Chouzenoux, J.-B. Fest, and C. Lefort. Block delayed majorize-minimize subspace algorithm for large scale image restoration. *Inverse Problems*, 39(4):044002, 2023.

[52] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011. Software available at *http://www.csie.ntu.edu.tw/ cjlin/libsvm.*

[53] A. Cherni, E. Chouzenoux, L. Duval, and J.-C. Pesquet. SPOQ lp-over-lq regularization for sparse signal recovery applied to mass spectrometry. *IEEE Transactions on Signal Processing*, 68:6070–6084, 2020.

[54] F. Chorobura and I. Necoara. Random coordinate descent methods for nonseparable composite optimization. Technical report, 2022. https://arxiv.org/abs/2203.14368.

[55] E. Chouzenoux. *Recherche de pas par Majoration-Minoration. Application à la résolution de problèmes inverses.* PhD thesis, Ecole Centrale de Nantes (ECN), 2010.

[56] E. Chouzenoux and J.-B. Fest. Convergence analysis of block majorize-minimize subspace approach. Technical report, 2022. https://hal.science/hal-03920026.

[57] E. Chouzenoux and J.-B. Fest. SABRINA: a stochastic subspace majorization-minimization algorithm. *Journal of Optimization Theory and Applications*, 195:919–952, 2022.

[58] E. Chouzenoux, J. Idier, and S. Moussaoui. A majorize–minimize strategy for subspace optimization applied to image restoration. *IEEE Transactions on Image Processing*, 20(6):1517–1528, 2010.

[59] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot. A majorize-minimize subspace approach for $\ell_2 - \ell_0$ image regularization. *SIAM Journal on Imaging Sciences*, 6(1):563–591, 2013.

[60] E. Chouzenoux, S. Martin, and J. Pesquet. A local MM subspace method for solving constrained variational problems in image recovery. Technical report, 2021. http://www.optimization-online.org/DB_HTML/2021/09/8595.html.

[61] E. Chouzenoux and J.-C. Pesquet. Convergence rate analysis of the majorize-minimize subspace algorithm. *IEEE Signal Processing Letters*, 23(9):1284–1288, 2016.

[62] E. Chouzenoux and J.-C. Pesquet. Convergence rate analysis of the majorize-minimize subspace algorithm. *IEEE Signal Processing Letters*, 23(9):1284–1288, Sep. 2016.

[63] E. Chouzenoux and J.-C. Pesquet. A stochastic majorize-minimize subspace algorithm for online penalized least squares estimation. *IEEE Transactions on Signal Processing*, 65(18):4770–4783, 2017.

[64] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. *Journal of Optimization Theory and Applications*, 162(1):107–132, 2014.

[65] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. A block coordinate variable metric forward-backward algorithm. *Journal of Global Optimization*, pages 1–29, Feb. 2016.

[66] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. A block coordinate variable metric forward–backward algorithm. *Journal of Global Optimization*, 66(3):457–485, 2016.

[67] E. Chouzenoux, T. Tsz-Kit Lau, C. Lefort, and J.-C. Pesquet. Optimal multivariate Gaussian fitting with applications to PSF modeling in two-photon microscopy imaging. *Journal of Mathematical Imaging and Vision*, 61(7):1037–1050, Sept. 2019.

[68] J. Chung, S. Knepper, and J. G. Nagy. *Large-Scale Inverse Problems in Imaging*, pages 47–90. Springer New York, New York, NY, 2015.

[69] P. L. Combettes and J.-C. Pesquet. A proximal decomposition method for solving convex variational inverse problems. *Inverse problems*, 24(6):065014, 2008.

[70] P. L. Combettes and J.-C. Pesquet. Stochastic approximations and perturbations in forward-backward splitting for monotone operators. *Pure and Applied Functional Analysis*, 1(1):13–37, 2016.

[71] P. L. Combettes and J.-C. Pesquet. Stochastic forward-backward and primal-dual approximation algorithms with application to online image restoration. In *Proceedings of the 24th European Signal Processing Conference (EUSIPCO 2016)*, pages 1813–1817, 2016.

[72] M. Coste. *An Introduction to Semialgebraic Geometry.* Istituti Editoriali e Poligrafici Internazionali, 2000.

[73] M. Coste. Ensembles semi-algébriques. In *Géométrie Algébrique Réelle et Formes Quadratiques: Journées SMF, Université de Rennes 1, Mai 1981*, pages 109–138. Springer, 2006.

[74] E. Cragg and A. Levy. Study on a supermemory gradient method for the minimization of functions. *Journal of Optimization Theory and Applications*, 4(3):191–205, 1969.

[75] W. C. Davidon. Variable metric method for minimization. *SIAM Journal on Optimization*, 1(1):1–17, 1959.

[76] D. Davis. The asynchronous palm algorithm for nonsmooth nonconvex problems. *arXiv preprint arXiv:1604.00526*, 2016.

[77] D. Davis, M. Udell, and B. Edmunds. The sound of APALM clapping: Faster nonsmooth nonconvex optimization with stochastic asynchronous palm. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, 5-10 Dec. 2016.

[78] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27, 2014.

[79] A. Defazio, J. Domke, et al. Finito: A faster, permutable incremental gradient method for big data problems. In *Proceedings of the International Conference on Machine Learning (ICML 2014)*, pages 1125–1133. PMLR, 2014.

[80] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

[81] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[82] L. Denis, E. Thiébaut, and F. Soulez. Fast model of space-variant blurring and its application to deconvolution in astronomy. In *2011 18th IEEE International Conference on Image Processing*, pages 2817–2820. IEEE, 2011.

[83] L. Denis, E. Thiébaut, F. Soulez, J. M. Becker, and R. Mourya. Fast approximations of shift-variant blur. *International Journal of Computer Vision*, 115:253–278, 2015.

[84] J. E. Dennis, Jr and J. J. Moré. Quasi-Newton methods, motivation and theory. *SIAM Review*, 19(1):46–89, 1977.

[85] A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. Technical report, 2020. https://arxiv.org/abs/1707.06386.

[86] J. L. Doob. *Stochastic Processes*. John Wiley & Sons, 1953.

[87] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

[88] V. Dudar, G. Chierchia, E. Chouzenoux, J.-C. Pesquet, and V. Semenov. A two-stage subspace trust region approach for deep neural network training. In *Proceedings of the 25th European Signal Processing Conference (EUSIPCO 2017)*, Kos Island, Greece, 28 Aug.-2 Sep. 2017.

[89] M. Duflo. *Random Iterative Models*, volume 34. Springer Science & Business Media, 2013.

[90] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems*. SIAM, 1999.

[91] V. Elvira and E. Chouzenoux. Optimized population Monte Carlo. Technical report, 2021. https://hal.archives-ouvertes.fr/hal-03136318.

[92] J. M. Ermoliev and Z. V. Nekrylova. The method of stochastic gradients and its application. In *Seminar: Theory of Optimal Solutions. No. 1 (Russian)*, pages 24–47. Akad. Nauk Ukrain. SSR, Kiev, 1967.

[93] P. Escande and P. Weiss. Sparse wavelet representations of spatially varying blurring operators. *SIAM Journal on Imaging Sciences*, 8:2976–3014, 2015.

[94] T. Fan and T. Murphey. Majorization minimization methods for distributed pose graph optimization with convergence guarantees. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2020)*, pages 5058–5065, Las Vegas, USA, 2020.

[95] I. Fatkhullin, J. Etesami, N. He, and N. Kiyavash. Sharp analysis of stochastic optimization under global Kurdyka-Lojasiewicz inequality. Technical report, 2022. https://arxiv.org/abs/2210.01748.

[96] B. Fehrman, B. Gess, and A. Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21, 2020.

[97] J. Fernandez-Bes, V. Elvira, and S. Van Vaerenbergh. A probabilistic least-mean-squares filter. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2199–2203. IEEE, 2015.

[98] J. A. Fessler. Grouped coordinate descent algorithms for robust edge-preserving image restoration. In T. J. Schulz, editor, *Image Reconstruction and Restoration II, International Society for Optics and Photonics*, volume 3170, pages 184–194, 1997.

[99] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, 1970.

[100] R. Fletcher and M. J. Powell. A rapidly convergent descent method for minimization. *The Computer Journal*, 6(2):163–168, 1963.

[101] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, 7(2):149–154, 1964.

[102] A. Florescu, E. Chouzenoux, J.-C. Pesquet, P. Ciuciu, and S. Ciochina. A majorize-minimize memory gradient method for complex-valued inverse problems. *Signal Processing*, 103:285–295, 2014.

[103] P. Frankel, G. Garrigos, and J. Peypouquet. Splitting methods with variable metric for kurdyka–łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165:874–900, 2015.

[104] S. Gadat. Stochastic optimization algorithms, non asymptotic and asymptotic behaviour. Technical report, 2017. https://perso.math.univ-toulouse.fr/m2r/files/2016/02/B5-2016-2017-new.pdf.

[105] S. Gadat and I. Gavra. Asymptotic study of stochastic adaptive algorithm in non-convex landscape. *Journal of Machine Learning Research*, 23:1–54, 2022.

[106] S. Gadat and F. Panloup. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. Technical report, 2017. https://arxiv.org/abs/1709.03342.

[107] C. Geiersbach and T. Scarinci. Stochastic proximal gradient methods for nonconvex problems in Hilbert spaces. *Computational Optimization and Applications*, 78(3):705–740, 2021.

[108] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(3):367–383, 1992.

[109] D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing*, 4(7):932–946, 1995.

[110] M. Gharbi, E. Chouzenoux, J.-C. Pesquet, and L. Duval. GPU-based implementations of MM algorithms. Application to spectroscopy signal restoration. In *Proceedings of the 29th European Signal Processing Conference (EUSIPCO 2021)*, 23-27 Aug. 2021.

[111] I. Gitman, H. Lang, P. Zhang, and L. Xiao. Understanding the role of momentum in stochastic gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[112] E. Gladyshev. On stochastic approximation. *Theory of Probability & Its Applications*, 10(2):275–278, 1965.

[113] W. Göbel, B. M. Kampa, and F. Helmchen. Imaging cellular network dynamics in three dimensions using fast 3d laser scanning. *Nature Methods*, 4(1):73–79, 2007.

[114] H. J. Greenberg and W. P. Pierskalla. A review of quasi-convex functions. *Operations Research*, 19(7):1553–1570, 1971.

[115] D. Grishchenko, F. Iutzeler, J. Malick, and M.-R. Amini. Asynchronous distributed learning with sparse communications and identification. Technical report, 2018. https://arxiv.org/abs/1812.03871.

[116] S. B. Hadj and L. Blanc-Féraud. Modeling and removing depth variant blur in 3d fluorescence microscopy. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 689–692. IEEE, 2012.

[117] R. Hannah and W. Yin. On unbounded delays in asynchronous parallel fixed-point algorithms. *Journal of Scientific Computing*, 76(1):299–326, Dec 2017.

[118] S. Haykin. *Blind Deconvolution*. 1994.

[119] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving. *Journal of Research of the National Bureau of Standards*, 49(6):409, 1952.

[120] G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Technical report, 2012. https://www.cs.toronto.edu.

[121] M. Hong, M. Razaviyayn, Z. Luo, and J. Pang. A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing. *IEEE Signal Processing Magazine*, 33(1):57–77, Jan 2016.

[122] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang. A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing. *IEEE Signal Processing Magazine*, 33(1):57–77, 2015.

[123] S. Horváth, M. Sanjabi, L. Xiao, P. Richtárik, and M. Rabbat. Fedshuffle: Recipes for better use of local work in federated learning. Technical report, 2022. https://arxiv.org/abs/2204.13169.

[124] P. J. Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[125] D. R. Hunter and K. Lange. [optimization transfer using surrogate objective functions]: Rejoinder. *Journal of Computational and Graphical Statistics*, 9(1):52–59, 2000.

[126] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[127] M. W. Jacobson and J. A. Fessler. Properties of MM algorithms on convex feasible sets: extended version. Technical report, 2004. https://www.eecs.umich.edu/techreports/systems/cspl/cspl-353.pdf.

[128] M. W. Jacobson and J. A. Fessler. An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms. *IEEE Transactions on Image Processing*, 16(10):2411–2422, Oct. 2007.

[129] M. A. Jendoubi. A simple unified approach to some convergence theorems of L. Simon. *journal of Functional Analysis*, 153(1):187–202, 1998.

[130] A. Jezierska, H. Talbot, and J.-C. Pesquet. Spatially variant psf modeling in confocal macroscopy. In *Proceedings of the 15th IEEE International Symposium on Biomedical Imaging (ISBI 2018)*, pages 489–492. Washington, DC, USA, 4-7 April 2018.

[131] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.

[132] J. P. Junior and W. De Melo. *Geometric theory of dynamical systems: an introduction*. Springer-Verlag, 1982.

[133] A. Khaled and P. Richtárik. Better theory for SGD in the nonconvex world. Technical report, 2020. https://arxiv.org/abs/2002.03329.

[134] A. Khintchine. Korrelationstheorie der stationären stochastischen prozesse. *Mathematische Annalen*, 109(1):604–615, 1934.

[135] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

[136] B. Kim and T. Naemura. Blind depth-variant deconvolution of 3D data in wide-field fluorescence microscopy. *Scientific Reports*, 5(1):1–9, 2015.

[137] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. Technical report, 2014. https://arxiv.org/abs/1412.6980.

[138] J. Konečnỳ, J. Liu, P. Richtárik, and M. Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255, 2015.

[139] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. Technical report, 2016. https://arxiv.org/abs/1610.02527.

[140] D.-J. Kroon. Showvol isosurface render, 2023. Matlab Central File Exchange.

[141] K. Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–783, 1998.

[142] K. Kurdyka, T. Mostowski, and A. Parusinski. Proof of the gradient conjecture of r. thom. *Annals of Mathematics*, pages 763–792, 2000.

[143] T. L. Lai. Stochastic approximation. *The Annals of Statistics*, 31(2):391–406, 2003.

[144] K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000.

[145] J.-F. Le Gall. Intégration, probabilités et processus aléatoires. *Ecole Normale Supérieure de Paris*, 2006.

[146] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems (NIPS 2000)*, page 535541, Denver, Colorado, 2000.

[147] C. Lefort, M. Chalvidal, A. Parente, V. Blanquet, H. Massias, L. Magnol, and E. Chouzenoux. FAMOUS: a fast instrumental and computational pipeline for multiphoton microscopy applied to 3d imaging of muscle ultrastructure. *Journal of Physics D: Applied Physics*, 54(27):274005, 2021.

[148] C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. Technical report, 2015. https://arxiv.org/abs/1512.07666.

[149] G. Li and T. K. Pong. Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, 18(5):1199–1232, 2018.

[150] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

[151] X. Li and A. Milzarek. A unified convergence theorem for stochastic optimization methods. In *Proceedings of Thirty-Sixth Conference on Neural Information Processing Systems (NEURIPS 2022)*, 2022.

[152] X. Li, A. Milzarek, and J. Qiu. Convergence of random reshuffling under the Kurdyka-Lojasiewicz inequality. Technical report, 2021. https://arxiv.org/abs/2110.04926.

[153] X.-L. Li. Preconditioned stochastic gradient descent. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1454–1466, 2017.

[154] Z. Li, Z. Dong, Z. Liang, and Z. Ding. Surrogate-based distributed optimisation for expensive black-box functions. *Automatica*, 125:109407, 2021.

[155] X. Lian, Y. Huang, Y. Li, and J. Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. Technical report, 2015. https://arxiv.org/abs/1506.08272.

[156] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

[157] N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.

[158] N. Loizou and P. Richtárik. Revisiting randomized gossip algorithms: General framework, convergence rates and novel block and accelerated protocols. *IEEE Transactions on Information Theory*, 67(12):8300–8324, 2021.

[159] S. Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[160] A. M. Lyapunov. The general problem of the stability of motion. *International Journal of Control*, 55(3):531–534, 1992.

[161] J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. Technical report, 2013. https://arxiv.org/abs/1306.4650.

[162] S. Y. Meng. *Stochastic second-order optimization for over-parameterized machine learning models*. PhD thesis, University of British Columbia, 2020.

[163] P.-A. Meyer. *Martingales and stochastic integrals I*, volume 284. Springer, 2006.

[164] A. Miele and J. Cantrell. Study on a memory gradient method for the minimization of functions. *Journal of Optimization Theory and Applications*, 3(6):459–470, 1969.

[165] K. Mishchenko, F. Iutzeler, and J. Malick. A distributed flexible delay-tolerant proximal gradient algorithm. *SIAM Journal on Optimization*, 30(1):933–959, 2020.

[166] K. Mishchenko, A. Khaled, and P. Richtárik. Proximal and federated random reshuffling. In *Proceedings of the International Conference on Machine Learning (ICML 2022)*, pages 15718–15749, 2022.

[167] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[168] E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24, 2011.

[169] K. P. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT press, 2012.

[170] J. Nagy and D. O'Leary. Restoring images degraded by spatially variant blur. *SIAM Journal on Scientific Computing*, 19(4):1063–1082, 1998.

[171] A. S. Nemirovskij and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

[172] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003.

[173] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[174] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), CORE Discussion Papers*, 22, Jan. 2010.

[175] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

[176] M. Nikolova and M.-K. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal of Scientific Computing*, 27:937–966, 2005.

[177] F. Niu, B. Recht, C. Re, and S. Wright. Hogwild: A lockfree approach to parallelizing stochastic gradient descent. In *Proceedings of the 25th Conference on Advances in Neural Information Processing Systems (NIPS 2011)*, pages 693–701, Granada, Spain, 12-17 Dec. 2011.

[178] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 1999.

[179] J. Nutini, M. Schmidt, I. H. Laradji, M. Friedlander, and H. Koepke. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML 2015)*, volume 37, page 1632–1641, 2015.

[180] A. Onose, R. E. Carrillo, A. Repetti, J. D. McEwen, J.-T. Thiran, J.-C. Pesquet, and Y. Wiaux. Scalable splitting algorithms for big-data interferometric imaging in the SKA era. *Monthly Notices of the Royal Astronomical Society*, 462(4):4314–4335, 2016.

[181] J. Ortega and W. Rheinboldt. Iterative solution of nonlinear equations in banach spaces, 1970.

[182] A. M. Ostrowski. *Solution of Equations in Euclidean and Banach Spaces*, volume 9. Academic press, 1973.

[183] D. A. Patterson, J. L. Hennessy, and D. Goldberg. *Computer Architecture: a Quantitative Approach*, volume 2. Morgan Kaufmann San Mateo, CA, 1990.

[184] M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. O. Hero, and S. McLaughlin. A survey of stochastic simulation and optimization methods in signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):224–241, 2015.

[185] J.-C. Pesquet and A. Repetti. A class of randomized primal-dual algorithms for distributed optimization. *Journal of Nonlinear and Convex Analysis*, 16(12):2353–2490, 2014.

[186] G. C. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

[187] E. Polak and G. Ribiere. Note sur la convergence de méthodes de directions conjuguées. *Revue française d'informatique et de recherche opérationnelle. Série rouge*, 3(16):35–43, 1969.

[188] B. T. Polyak. *Introduction to Optimization*. Translations Series in Mathematics and Engineering. New York: Optimization Software Inc. Publications Division, 1987.

[189] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[190] J. Ponstein. Seven kinds of convexity. *Siam Review*, 9(1):115–119, 1967.

[191] M. J. Powell. Rank one methods for unconstrained optimization. Technical report, Atomic Energy Research Establishment, Harwell (England), 1969.

[192] M. Prato, A. La Camera, S. Bonettini, S. Rebegoldi, M. Bertero, and P. Boccacci. A blind deconvolution method for ground based telescopes and Fizeau interferometers. *New Astronomy*, 40:1–13, 2015.

[193] C. Preza and J.-A. Conchello. Depth-variant maximum-likelihood restoration for three-dimensional fluorescence microscopy. *Journal of the Optical Society of America A*, 21(9):1593–1601, 2004.

[194] W. Rasband. ImageJ, 2018. https://imagej.nih.gov/ij/.

[195] M. Razaviyayn, M. Hong, and Z.-Q. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

[196] A. Repetti, M. Pham, L. Duval, E. Chouzenoux, and J.-C. Pesquet. Euclid in a Taxicab: Sparse blind deconvolution with smoothed l1/l2 regularization. *IEEE Signal Processing Letters*, 22(5):pages 539–543, May 2015.

[197] A. Repetti and Y. Wiaux. A forward-backward algorithm for reweighted procedures: Application to radio-astronomical imaging. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, pages 1434–1438, 4-8 May 2020.

[198] A. Repetti and Y. Wiaux. Variable metric forward-backward algorithm for composite minimization problems. *SIAM Journal on Optimization*, 31(2):1215–1241, May 2021.

[199] E. Richardson, R. Herskovitz, B. Ginsburg, and M. Zibulevsky. Seboost-boosting stochastic learning using subspace optimization techniques. Technical report, 2016. https://arxiv.org/abs/1609.00629.

[200] P. Richtárik and M. Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, Apr 2015.

[201] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[202] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier, 1971.

[203] M. C. Robini and Y. Zhu. Generic half-quadratic optimization for image reconstruction. *SIAM Journal on Imaging Sciences*, 8(3):1752–1797, 2015.

[204] R. T. Rockafellar. *Convex Analysis*, volume 18. Princeton university press, 1970.

[205] L. Rosasco, S. Villa, and B. C. Vũ. Convergence of stochastic proximal gradient algorithm. *Applied Mathematics and Optimization*, 82:891–917, 2014.

[206] N. Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. *Advances in Neural Information Processing Systems*, 25, 2012.

[207] S. Ruder. An overview of gradient descent optimization algorithms. Technical report, 2016. https://arxiv.org/abs/1609.04747.

[208] W. Rudin et al. *Principles of Mathematical Analysis*, volume 3. McGraw-Hill New York, 1976.

[209] D. Saad. Online algorithms and stochastic approximations. *Online Learning*, 5(3):6, 1998.

[210] A. K. Sahu, T. Li, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith. On the convergence of federated optimization in heterogeneous networks. Technical report, 2018. https://arxiv.org/abs/1812.06127.

[211] S. Salzo and S. Villa. Inexact and accelerated proximal point algorithms. *Journal of Convex Analysis*, 19(4):1167–1192, 2012.

[212] M. Schmidt and N. L. Roux. Fast convergence of stochastic gradient descent under a strong growth condition. Technical report, 2013. https://inria.hal.science/hal-00855113.

[213] N. N. Schraudolph, J. Yu, and S. Günter. A stochastic quasi-Newton method for online convex optimization. In *Artificial Intelligence and Statistics*, pages 436–443. PMLR, 2007.

[214] L. Schwartz. *Analyse IV: Applications à la Théorie de la Mesure*. Editions Hermann, 1993.

[215] G. Scutari and Y. Sun. *Parallel and Distributed Successive Convex Approximation Methods for Big-Data Optimization*. Springer Verlag Series. C.I.M.E Lecture Notes in Mathematics, 2018.

[216] O. Sebbouh, R. M. Gower, and A. Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Proceedings of the Conference on Learning Theory (COLT 2021)*, pages 3935–3971, 2021.

[217] M. Sghaier, E. Chouzenoux, J.-C. Pesquet, and S. Muller. A novel task-based reconstruction approach for digital breast tomosynthesis. 2020.

[218] S. Shalev-Shwartz. SDCA without duality, regularization, and individual convexity. In *Proceedings of the International Conference on Machine Learning (ICML 2016)*, pages 747–754. PMLR, 2016.

[219] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2021.

[220] Z.-J. Shi and J. Shen. Convergence of supermemory gradient method. *Journal of Applied Mathematics and Computing*, 24(1):367–376, 2007.

[221] P. Sonneveld. CGS: A fast Lanczos-type solver for nonsymmetric linear systems. *SIAM Journal on Scientific Statistical Computing*, 10(1):36–52, Jan. 1989.

[222] S. Sotthivirat and J. A. Fessler. Image recovery using partitioned-separable paraboloidal surrogate coordinate ascent algorithms. *IEEE Transactions on Signal Processing*, 11(3):306–317, 2002.

[223] Y. Sun, P. Babu, and D. P. Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2016.

[224] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the International conference on machine learning (ICML 2013)*, pages 1139–1147, 2013.

[225] P. Thouvenin, N. Dobigeon, and J.-Y. Tourneret. Partially asynchronous distributed unmixing of hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4), Apr. 2019.

[226] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization, Theory and Applications*, 109(3):475–494, 2001.

[227] J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986.

[228] J. Tuck, D. Hallac, and S. Boyd. Distributed majorization-minimization for Laplacian regularized problems. *IEEE/CAA Journal of Automatica Sinica*, 6(1):45–52, 2019.

[229] L. Van den Dries and C. Miller. Geometric categories and o-minimal structures. Technical report, 1996. https://people.math.osu.edu/miller.1987/newg.pdf.

[230] R. G. Vuchkov, C. G. Petra, and N. Petra. On the derivation of quasi-newton formulas for optimization in function spaces. *Numerical Functional Analysis and Optimization*, 41(13):1564–1587, 2020.

[231] A. Wald and T. Schuster. Sequential subspace optimization for nonlinear inverse problems. *Journal of Inverse and Ill-posed Problems*, 25(1):99–117, 2017.

[232] J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, et al. A field guide to federated optimization. Technical report, 2021. https://arxiv.org/abs/2107.06917.

[233] X. Wang, S. Ma, D. Goldfarb, and W. Liu. Stochastic quasi-Newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956, 2017.

[234] D. Williams. *Probability with Martingales*. Cambridge university press, 1991.

[235] A. Wilson. *Lyapunov Arguments in Optimization*. University of California, Berkeley, 2018.

[236] O. Wirjadi and T. Breuel. Approximate separable 3D anisotropic Gauss filter. In *Proceedings of the 12nd IEEE International Conference on Image Processing (ICIP 2005)*, volume 2, pages 149–52, Genoa, Italy, 11-14 Sept 2005.

[237] P. Wolfe. Convergence conditions for ascent methods. *SIAM Review*, 11(2):226–235, 1969.

[238] L. Xiao. Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

[239] A. B.-B. Y. Marnissi, E. Chouzenoux and J.-C. Pesquet. Majorize-minimize adapted Metropolis-Hastings algorithm. *IEEE Transactions on Signal Processing*, (68):2356–2369, 2020.

[240] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

[241] P. Yu, G. Li, and T. K. Pong. Kurdyka–łojasiewicz exponent via inf-projection. *Foundations of Computational Mathematics*, 22(4):1171–1217, 2022.

[242] Y. Yuan. Subspace techniques for nonlinear optimization. In *Some Topics in Industrial and Applied Mathematics*, pages 206–218. World Scientific, 2007.

[243] Y.-X. Yuan. Subspace methods for large scale nonlinear equations and nonlinear least squares. *Optimization and Engineering*, 10(2):207–218, nov 2008.

[244] S. Zavriev and F. Kostyuk. Heavy-ball method in nonconvex optimization problems. *Computational Mathematics and Modeling*, 4(4):336–341, 1993.

[245] R. Zhang and J. T. Kwok. Asynchronous distributed ADMM for consensus optimization. In *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML 2014)*, volume 32, 2014.

[246] Z. Zhang, J. T. Kwok, and D.-Y. Yeung. Surrogate maximization/minimization algorithms and extensions. *Machine Learning*, 69:1–33, 2007.

[247] P. Zheng, E. Chouzenoux, and L. Duval. PENDANTSS:penalized norm-ratios disentangling additive noise, trend and sparse spikes. Technical report, 2023. https://arxiv.org/abs/2301.01514.

[248] Y. Zheng and Q. Liu. A review of distributed optimization: Problems, models and algorithms. *Neurocomputing*, 483:446–459, 2021.

[249] M. Zibulevsky. SESOP-TN: Combining sequential subspace optimization with truncated Newton method. Technical report, 2008. http://www.optimization-online.org/DB_FILE/2008/09/2098.pdf.

[250] G. Zoutendijk. Nonlinear programming, computational methods. *Integer and Nonlinear Programming*, pages 37–86, 1970.