UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

ACCELERATION OF OPTIMIZATION ALGORITHMS

TESIS PARA OPTAR AL GRADO DE
DOCTOR EN CIENCIAS DE LA INGENIERÍA,
MENCIÓN MODELACIÓN MATEMÁTICA
EN COTUTELA CON LA UNIVERSIDAD DE GRONINGEN

JUAN JOSÉ MAULÉN MUÑOZ

PROFESOR GUÍA 1:
JUAN PEYPOUQUET URBANEJA
PROFESOR GUÍA 2:
ARIS DANIILIDIS

MIEMBROS DE LA COMISIÓN:
FRANCISCO ARAGÓN-ARTACHO
CRISTÓBAL BERTOGLIO BELTRÁN
HECTOR RAMÍREZ CABRERA
MATHIAS STAUDIGL

Este trabajo ha sido parcialmente financiado por Beca Doctorado Nacional ANID 2021 21210993 y CMM ANID BASAL FB210005.

SANTIAGO DE CHILE
2023

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE
DOCTOR EN CIENCIAS DE LA INGENIERÍA,
MENCIÓN MODELACIÓN MATEMÁTICA
POR JUAN JOSÉ MAULÉN MUÑOZ
FECHA: 2023
PROF. GUÍAS: JUAN PEYPOUQUET, ARIS DANIILIDIS

ACELERACIÓN DE ALGORITMOS DE OPTIMIZACIÓN

El trabajo de tesis está enfocado en el estudio de la convergencia y desempeño de algoritmos de optimización combinando mecanismos de aceleración y estabilización. En particular, nos enfocamos en dos técnicas: un esquema de reinicio para una dinámica continua, y la inclusión de inercia en iteraciones de Krasnoselskii-Mann.

El capítulo 2 presenta un resultado sobre la convergencia de esquema de reinicio para función convexa a través de las soluciones de una ecuación diferencial con amortiguamiento definido por la curvatura. Esta dinámica en particular puede ser interpretada como la versión continua del método acelerado de Nesterov con un término que depende del Hessiano de una función convexa. La inclusión del término de segundo orden muestra, en la práctica, una reducción en las oscilaciones de los valores de la función, resultando en una convergencia más estable. Los resultados obtenidos en este capítulo pueden ser vistos como una extensión del esquema de reinicio propuesto por Su, Boyd y Candès en [83]. Se presentan también experimentos numéricos, mostrando el desempeño del esquema de reinicio, en su versión continua y sus consecuencias algorítmicas, y un teorema de existencia para las soluciones de la ecuación diferencial.

El capítulo 3 se enfoca en la inclusión de inercia en iteraciones del tipo Krasnoselskii-Mann (KM), gobernadas por una familia de operadores quasi-no expansivos sobre un espacio de Hilbert. Las iteraciones KM pueden ser vistas como una versión relajada de las iteraciones de Banach-Picard, y bajo hipótesis habituales, convergen a un punto fijo en común para la familia de operadores. Las iteraciones de punto fijo juegan un importante rol al definir algoritmos de optimización para una variada gama de problemas. En particular, se presentan resultados sobre la convergencia débil de las iteraciones hacia el punto fijo, junto con estimaciones para la tasa de convergencia no asintótica para los residuos. Se obtienen resultados sobre convergencia fuerte y lineal en el caso quasi-contractante, y se presentan simulaciones numéricas para versiones inerciales de un algoritmo primal dual y otro gobernado por tres operadores. En ambos casos, se observan mejoras en el desempeño con respecto a sus contrapartes no inerciales. Este capítulo también presenta resultados de una investigación en curso sobre el problema de estimación de desempeño (PEP) para iteraciones KM inerciales, lo cual nos lleva a conjeturar sobre las tasas de convergencia.

SUMMARY OF THE THESIS TO OBTAIN THE DEGREE OF
DOCTOR EN CIENCIAS DE LA INGENIERÍA,
MENCIÓN MODELACIÓN MATEMÁTICA
BY JUAN JOSÉ MAULÉN
DATE: 2023
ADVISORS: JUAN PEYPOUQUET, ARIS DANIILIDIS

ACCELERATION OF OPTIMIZATION ALGORITHMS

This thesis is focused on studying the convergence and performance of optimization algorithms combining acceleration and stabilization mechanisms. In particular, we focus on two techniques: a restart scheme for a continuous dynamics, and the inclusion of inertia on Krasnoselskii-Mann iterations.

Chapter 2 contains a result on the convergence of a restarting scheme for a convex function through the solution trajectories of a differential equation with curvature-defined damping. This particular dynamics can be interpreted as the continuous setting for Nesterov's accelerated method with a term that depends on the Hessian of the convex function. The inclusion of the Hessian term shows, in practice, a reduction in the oscillations of the values of the function, leading to a more stable convergence. The results displayed on this chapter can be interpreted as an extension of the restart scheme proposed by Su, Boyd and Candès in [83]. Numerical experiments are displayed, showing the performance of the restart routine both in the continuous setting and its algorithmic consequences, and an existence theorem for the solutions of the differential equation.

Chapter 3 focuses on the inclusion of inertia on Krasnoselskii-Mann iterations, governed by a family of quasi-nonexpansive operators defined over a Hilbert space. Krasnoselskii-Mann iterations can be seen as a relaxed setting for Banach-Picard iterations, and under standard hypotheses, they converge towards a common fixed point of the family of operators. Fixed point iterations play a central role on defining optimization algorithms for a wide range of problems. In particular, we provide results on the weak convergence of KM iterations towards a common fixed point of the famility of operators, along with estimates for the non-asymptotic rate at which the residuals vanish. Strong and linear convergence are obtained in the quasi-contractive setting, and numerical illustrations are displayed for an inertial primal-dual method and an inertial three-operator splitting algorithm, whose performance is superior to that of their non-inertial counterparts. Also this chapter presents some results of an ongoing research about the performance estimation problem (PEP) for inertial KM iterations, leading us to conjecture about rates of convergence.

A mi familia

Agradecimientos

Primero que todo, agradezco a mis dos supervisores que permitieron que este trabajo se llevara a cabo. Gracias a Juan Peypouquet por guiarme, por confiar en mí, por sus consejos y por constantemente sacar lo mejor de mí. Gracias a Aris Daniilidis por supervisar mi trabajo y por sus comentarios para que lograra su mejor forma posible. Gracias a Luis Briceño y Fernando Roldán por su gentileza en ayudarnos en las simulaciones, resolviendo dudas y compartiendo códigos. Gracias a Ignacio Fierro, con quien trabajamos codo a codo todo un verano procesando imágenes. Gracias a Juan Pablo Contreras, con quien trabajamos en la última parte del proyecto. Gracias a Sebastián Zamorano, mi primer guía y agente fundamental en mi decisión de querer dedicarme a la matemática. Finalmente, gracias a Francisco Aragón-Artacho, Mathias Stuidigl, Cristóbal Bertoglio y Héctor Ramírez por concederme el honor de revisar mi trabajo.

A riesgo de sonar repetitivo, en el plano personal quiero agradecer nuevamente a Juan Peypouquet. Gracias por darme la oportunidad de viajar a Groningen y vivir una de las mejores experiencias de mi vida. Gracias por las cenas, los vinos y tu invaluable conocimiento sobre funciones convexas y teoría de operadores. Gracias por abrirme las puertas de tu casa y recibirme como a uno más de tu familia. Extiendo los agradecimientos a Alexis, Emi y Ñarki por aceptarme dentro del clan durante ese año.

Gracias a toda la gente en la Universidad de Chile que me acogió desde el primer día, compañeros de Doctorado y funcionarios que hicieron que mi estadía en Beauchef fuera una muy linda experiencia. Gracias compañeros por todos los momentos y cafés compartidos: Jessica, Jorge, Daniel, Yamit, Hugo, JP, Víctor, Melanie, Seba, Emilio. Especiales gracias a Natacha, Silvia y Karen por toda su ayuda. Extiendo mi gratitud a todos los profesores con quien tuve el privilegio de asistir a clases. Ha sido un honor estudiar en este departamento.

Very special thanks to all the people I shared with in Groningen, in what was one of the most special years of my life. I have a special affection for the city and every person I met there. To my paranymphs and office mates Edo and Maaike, thanks for your friendship and support on the defense. Special thanks to Niki, Rodrigo, Abby, Cornée, Rennik, Vale and to everyone who has been a part of my experience in this beautiful city. I want to express my appreciation to all the colleagues and professors on Bernoulliborg, specially the people from SCO group, who contributed to making me feel warmly welcomed during my stay.

Gracias a mi familia, a mi padre Juan y mi madre Carmen por apoyarme en todo momento, en todas mis decisiones y sobre todo, por toda la ayuda que me brindaron el último año. Gracias a mis hermanos, Rodrigo, Carolina y Claudio por siempre estar para mí. Especialmente agradezco a Rodrigo, quien durante los últimos años fue la mayor inspiración para pararse cada vez y hacer frente a cualquier obstáculo que se interponga. Gracias por mostrarme que por difícil que se vea el panorama, se va de frente y se da la pelea. Gracias a mis sobrinos, primos y familia en general por su infinito e incondicional cariño. En particular, gracias a mi primo Renato Peña por ayudarme en el diseño de la portada.

Finalmente, pero no menos importante, dedicar unas palabras al más grande tesoro que tengo el lujo de poseer: mis amistades. No los puedo nombrar a todos, pero Andy, Cristopher, Luchin y Jack se merecen la mayor de las gratitudes. Hicieron que la distancia desde los pastos de la Usach hasta Groningen pareciera ínfima. No creo que haya gente capaz de hacer una locura similar. Yo no, desde luego. Gracias a todos los que me bancaron la irrisoria decisión de renunciar al banco y entrar a un Doctorado. Gracias a Mirko y a Dani por estar ahí siempre, incluso en los momentos más difíciles. Para no dejar a nadie fuera, extiendo los agradecimientos a todos los que me han acompañado durante estos años, el camino fue largo y no olvido a todos quienes han estado ahí.

Gracias a ANID, al Centro de Modelamiento Matemático (CMM) y la Universidad de Groningen por financiar este trabajo.

Contents

Chapter 1

Introduction

Mathematical Optimization, the science behind finding the best solution in a range of possibilities, is the core of endless applications and one of the main supports of modern mathematics. Convex analysis and operator theory provide mathematical properties and background to tackle optimization problems. These two fields of study allow to build efficient and effective algorithms that converge to the optimal solution of the problem, leading researchers to study conditions and features that guarantee convergence properties.

Nowadays, the increasing computer power and new coding techniques allow to provide a wide range of applications for convex optimization and algorithms in engineering and industry problems. Some applications include classification and regression models for Machine Learning, signal processing, risk modeling and portfolio management in finance, and control theory related applications such as transport, biological systems and aerospace engineering.

Both in convex optimization and in operator theory, convergence results can be established regarding the speed of convergence for algorithms, commonly known as *rates of convergence*, which are upper bounds for some metric that measures proximity to an optimality condition, for example the difference between the function values and the optimal value, the distance between iterations to the theoretical solution, or the distance between two consecutive iterations. These kind of results have a great significance, because they ensure beforehand the speed at which the algorithm will converge, allowing to establish stopping criteria for a given tolerance. The rates of convergence, in practice do not provide information about the execution time of the algorithm, but they provide valuable information about the precision reached in a fixed amount of iterations. For complex problems, dealing with larger volumes of data, an improvement on the execution time, memory consumption or precision defines a crucial and valuable objective.

This thesis presents results regarding the study of acceleration, stabilization and performance of optimization algorithms, in the context of convex analysis and fixed-point iterations. The main work is divided into two mayor sections, both of them motivated by an important minimization algorithm to be presented next.

Nesterov's accelerated gradient method

The fundamental problem in convex optimization is to find the minimum of a convex function. For that purpose, first-order methods are among the most popular tools, being the gradient method [27] one of the essentials, mostly because it is easy to implement and is computationally efficient. This method can be interpreted as a finite-difference discretization of the differential equation

$$\dot{x}(t) + \nabla\phi\big(x(t)\big) = 0, \tag{1.1}$$

describing the steepest descent dynamics. The gradient method is applicable to smooth functions, but there are more contemporary variations that can deal with nonsmooth ones, and even exploit the functions' structure to enhance the algorithm's per iteration complexity, or overall performance. A key example is the *proximal-gradient* (or *forward-backward*) method [57, 73], (see also [61, 79]), which is in close relationship with a nonsmooth version of (1.1). See [19] for a more detailed reading about first-order methods.

In 1983, Yurii Nesterov proposed an accelerated version of the gradient method [67], which can be formulated as follows. Let $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a differentiable convex function that attains its minimum at x^*. Given two starting points x_0, $x_1 \in \mathbb{R}^n$, define the sequences

$$\left\{ \begin{array}{rl} y_k & = x_k + \alpha_k(x_k - x_{k-1}), \\ x_{k+1} & = y_k - s\nabla\phi(y_k), \end{array} \right. \tag{1.2}$$

where $s > 0$ is the step-size and α_k acts as an extrapolation sequence given by

$$\alpha_k = \frac{k-1}{k+\alpha-1} \approx 1 - \frac{\alpha}{k}.$$

If $\nabla\phi$ is L-Lipschitz, $s \leq 1/L$ and $\alpha \geq 3$, this method exhibits the rate of convergence

$$\phi(x_k) - \phi^* \leq \mathcal{O}\left(\frac{1}{k^2}\right),$$

where $\phi^* = \phi(x^*)$. This scheme exhibits a faster worst-case convergence rate than the regular gradient method (which corresponds to consider $\alpha_k \equiv 0$ on (1.2)), which exhibits a rate of convergence of $\mathcal{O}(1/k)$. Then, the acceleration of the convergence is given by the inclusion of the sequence α_k, which is usually referred as *momentum* or *inertia*. The motivation for considering this inertial term can be tracked to different algorithms and physical motivations (see [76, 4]). Notice that the accelerated version (1.2) has a similar computational complexity than the usual gradient method, because it only adds an extrapolation of two points.

Nesterov's scheme, and its convergence properties motivate the work presented on this thesis. The work is divided into two mayor parts, each of which can be seen as a branch of extensions of the acceleration ideas provided by Nesterov.

1.1 Part I: Restart of a dynamics with Hessian damping

In [83], Su, Boyd and Candès studied the following differential equation:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla\phi(x(t)) = 0, \tag{AVD}$$

with $\alpha > 0$ and $t > 0$. This equation can be interpreted as a continuous setting for Nesterov's method, given that the method can be derived by discretize (AVD). Equation (AVD) is commonly known as AVD (Asymptotic Vanishing Damping), and [83] shows that the function values along the trajectories converge towards the minimum at a rate of $\mathcal{O}(1/t^2)$ for $\alpha \geq 3$. Moreover in the case $\alpha > 3$, [12] shows that $x(t)$ converges weakly towards x^* and [63] that the rate of convergence is $o(1/t^2)$.

On [15], Attouch, Peypouquet and Redont proposed a Dynamic Inertial Newton system with Asymptotically Vanishing Damping, given by

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla\phi(x(t)) + \beta\nabla^2\phi(x(t))\dot{x}(t) = 0, \tag{DIN-AVD}$$

where $\alpha, \beta > 0$. Similar to the case (AVD), exhibits a rate of convergence of $\mathcal{O}(1/t^2)$ for the function values along the solutions, when $\alpha \geq 3$. Although this system asks to ϕ being twice differentiable, and it seems more difficult to deal with second-order derivatives, the authors show that it can be transformed into an equivalent first-order equation system in time and space, which can be extended to a differential inclusion that is well posed whenever ϕ is closed and convex.

One of the most interesting facts of the solutions of (DIN-AVD) is that the inclusion of the Hessian term can be interpreted as a damping agent for the oscillations in the convergence. As a very simple example, consider the convex function $\phi : \mathbb{R}^2 \to \mathbb{R}$ given by $\phi(x_1, x_2) = \frac{1}{2}(x_1^2 + 100x_2^2)$, which has a global minimum at $(0, 0)$ and optimal value 0. Figure 1.1 shows the convergence both for the trajectories and function values obtained by solving (DIN-AVD) and (AVD) (which corresponds to (DIN-AVD) with $\beta = 0$). In both cases, the convergence towards the minimizer and the optimal value can be observed, with a remarkable neutralizer effect of the Hessian term.

The erratic behavior displayed by (AVD), and mostly by its discrete counterpart, Nesterov's method, has motivated a strategy which aims to avoid it: restart schemes. This heuristic involves running the algorithm until a stopping criteria is achieved, with the most common choice being whenever $\phi(x_{k+1}) > \phi(x_k)$, that is, the point when the function starts to increase instead of decrease (see Figure 1.1 bottom left, the points where the function "bounces"). Then, restart k to 0 and perform the algorithm again using as initial point the last point of the previous cycle. O'Donaghue and Candès [72], proposed this strategy which shows a monotonic and faster convergence in practice, but there are no rates of convergence for this method that can explain the acceleration obtained. In order to address that problem, there are several works that provide rates of convergence using different criteria or approaches. One of those is the result given by Su, Boyd and Candès [83], where they propose a restart scheme for the continuous setting of Nesterov's method, that is, equation (AVD), using as restart criteria the first moment where the speed of the solution trajectories

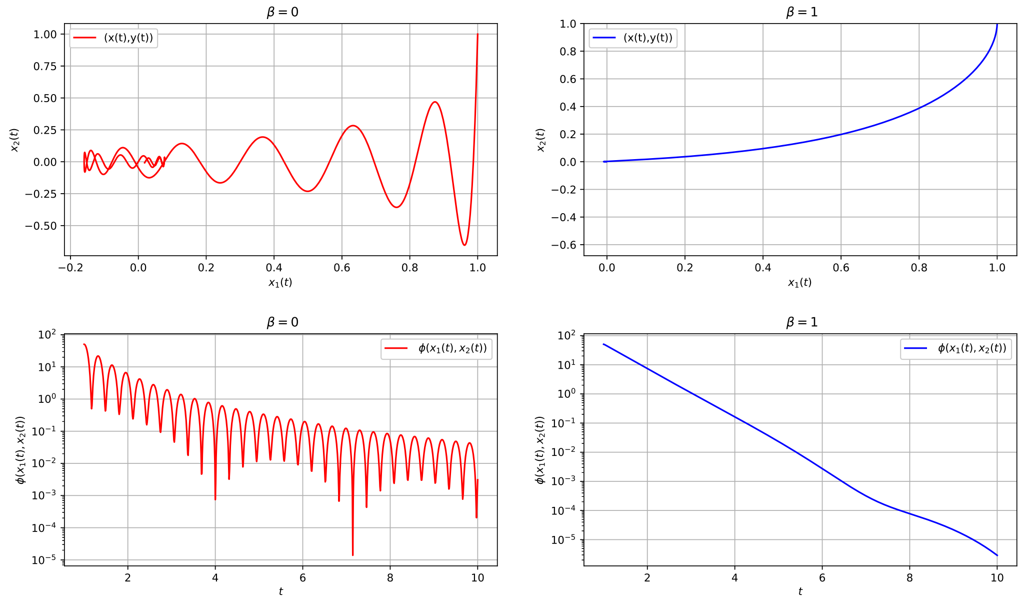


Figure 1.1: Convergence of solutions of (DIN-AVD) in the cases $\beta = 0$ and $\beta = 1$, using $\alpha = 3.1$ on the interval $[1, 10]$. Trajectories (top) and function values (bottom).

decreases. The motivation behind this criteria is to keep a high velocity along the trajectory. Although a rate of convergence for the algorithm is not provided, the result in the continuous case illustrates the behavior of a restart routine in an algorithmic context.

The main objective of this first part is to provide a speed restart scheme for the dynamics given by (DIN-AVD), that is, to extend the results obtained in [83] to the case involving the Hessian of the function. Our result states that the trajectories generated by this method converge weakly as $t \to \infty$ and the values converge to the optimum linearly. Equation (DIN-AVD) leads to first-order optimization algorithms which involve a gradient-correction term, being studied recently in [10, 11, 1], for example. The continuous speed restart result obtained will hopefully give an insight into the use of a restart routine on this kind of algorithms.

This first part is displayed on Chapter 2 as follows: Section 2.2 presents the main result as a theorem for the convergence and the rate for the restart scheme for equation (DIN-AVD). Sections 2.3 and 2.4 present technical results that are used to prove the main result, the first one provides a bound for the restarting time, and the second one proves the linear convergence between two restarts. Using the linear convergence obtained, along with an estimation of how many times the trajectory has been restarted, the proof of the main theorem is straightforward. A few numerical examples are displayed on Section 2.5 on the continuous case, and Section 2.6 on the discrete case. In both cases, the convergence of the function values is stabilized by the restart routine. By the end of the chapter, the proof of the existence theorem and a useful bound are provided in Appendix 2.7 and Appendix 2.8, respectively.

The results presented in this first part led to the writing of the article [62], published in *Journal of Optimization: Theory and applications*.

1.2 Part II: Inertial Krasnoselskii-Mann iterations

Let $\mathcal{H}$ be a real Hilbert space and consider the classical gradient method, that is: from a starting point $x_0 \in \mathcal{H}$ and $s > 0$, recursively define $x_{k+1} = x_k - s\nabla\phi(x_k)$. Notice, that the algorithm can be defined in an operator form, as $x_{k+1} = Tx_k$, with $T = I - s\nabla\phi$. Then, finding a minimum of the function is the same as finding a fixed point of the operator.

Given an operator T and $x_0 \in \mathcal{H}$, iterations of the form

$$x_{k+1} = Tx_k, \tag{1.3}$$

introduced in [75], are commonly known as Banach-Picard iterations and they converge to the unique fixed point of the operator, when T is a strict contraction (L-Lipschitz with $L < 1$) thanks to Banach's fixed point theorem [17]. The convergence of iterations (1.3) towards a fixed point of the operator can fail when this is not a strict contraction. A simple example is to consider the operator $Tx = -x$, which has 0 as the unique fixed point. In that case, iterations (1.3) with $x_0 \neq 0$ do not converge.

In order to relax the strict contraction hypothesis, Krasnoselskii-Mann (KM) iterations [52, 60] were introduced. Given a sequence $\lambda_k \in (0, 1)$, generate recursively

$$x_{k+1} = (1 - \lambda_k)x_k + \lambda_k Tx_k, \tag{1.4}$$

with $x_0 \in \mathcal{H}$. In [78] it is proved that if T is a nonexpansive operator (L-Lipschitz with $L \leq 1$) with $\text{Fix}(T) \neq \emptyset$, and $\sum \lambda_k(1 - \lambda_k) = \infty$, then iteration (1.4) converges weakly towards a fixed point of T. Then, λ_k acts as a sequence of relaxation or averaging parameters that can improve the convergence with respect to Banach-Picard iterations. As a very simple illustration, consider again $Tx = -x$ and $\lambda_k \equiv \lambda$. Then, iteration (1.4) converges linearly to 0 for every value of $\lambda \in (0, 1)$ and moreover, in the case $\lambda = 1/2$ converges in one step.

As shown before, the gradient method, which can be seen as a particular instance of fixed-point iterations, improves its convergence by the inclusion of an inertial term. Then, the purpose of this part is to develop further insight into the convergence properties of inertial Krasnoselskii-Mann iterations in their general form

$$\left\{ \begin{array}{rcl} y_k & = & x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} & = & (1 - \lambda_k)y_k + \lambda_k T_k y_k, \end{array} \right. \tag{1.5}$$

where (T_k) is a family of operators defined on a real Hilbert space $\mathcal{H}$, and the positive sequences (α_k) and (λ_k) are the *inertial* and *relaxation* (or *averaging*) parameters, respectively.

The general aim of this part is to provide conditions on the parameter sequences and the family of operators to ensure that the sequences generated by (1.5) converges (weakly or strongly) to a common fixed point of the T_k's, provided there are any. More specifically, our contribution consists on a weak convergence result simplifying the existing hypotheses on the literature, a strong convergence result with a rate of convergence, and numerical simulations where two inertial schemes for existing fixed point algorithms are provided. We shall also see that adding the inertial term does not always make algorithms faster (this is reflected in the worst-case convergence rates), but may boost their convergence in some relevant instances.

The results presented in this second part corresponds to a joint work with Ignacio Fierro[1], and led to the article [45], submitted to *Set-Valued and Variational Analysis*.

[1] PhD, BIOCORE team, Centre INRIA de l'Université de la Côte d'Azur, France.

In addition to the theoretical and numerical convergence results leading to the article mentioned, we will also include results from an ongoing research project. This project is motivated for an open question that has emerged in our study of KM iterations. The rate of convergence found for the strong convergence of the iterations, implies that the optimal performance is achieved when inertia is not employed. One challenging matter about rates is to evaluate their *tightness*, that is, how close they approximate the actual convergence. Since the rate is an upper bound, in practice it can converge to zero significantly slower than the algorithm and will be still an upper bound. Then, a motivating open question in this matter is to obtain a rate of convergence that truly explains the inclusion of inertia on KM iterations.

Following ideas of Drori and Teboulle on [43], the problem of estimating the worst-case speed of convergence for some iterations can be stated itself as an optimization problem, called Performance Estimation Problem (PEP). Using this approach, the problem of estimating rates of convergence for inertial KM iterations can be stated as a PEP. An overview of the PEP modeling is provided along with preliminary results from our ongoing numerical experiments.

This second part is displayed on Chapter 3 as follows: in Section 3.2 we establish the weak convergence of KM iterations towards a common fixed point of the family of operators in the quasi-nonexpansive case, along with a non-asymptotic rate at which the residuals vanish. Section 3.3 is devoted to the strong and linear convergence in the quasi-contractive setting. In both cases, we highlight the relationship with the non-inertial case, and show that passing from one regime to the other is a continuous process in terms of parameter hypotheses and convergence rates. In Section 3.4, we discuss several instances of KM iterations, which are relevant to the numerical illustrations provided in Section 3.5, concerning an inertial primal-dual method and an inertial three-operator splitting algorithm. The results of the PEP analysis for KM iterations are displayed at Section 3.6.

Chapter 2

Restart of a Hessian dynamics

2.1 Introduction

Let $\phi : \mathbb{R}^n \to \mathbb{R}$ a twice continuously differentiable convex function, which attains its minimum value ϕ^*, and whose gradient $\nabla\phi$ is Lipschitz-continuous with constant $L > 0$. In [83], Su, Boyd and Candès studied the following differential equation:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla\phi(x(t)) = 0, \tag{AVD}$$

with $\alpha > 0$ and $t > 0$. Despite its rate of convergence guarantees a faster decay than the steepest descent dynamics (1.1), trajectories satisfying (AVD) exhibit a somewhat chaotic behavior, especially if the objective function is ill-conditioned. In particular, the function values tend not to decrease monotonically, but to present an oscillatory behavior, instead.

Example 2.1.1. *We consider the quadratic function $\phi : \mathbb{R}^3 \to \mathbb{R}$, defined by*

$$\phi(x_1, x_2, x_3) = \frac{1}{2}(x_1^2 + \rho x_2^2 + \rho^2 x_3^2), \tag{2.1}$$

Figure 2.1 shows the behavior of the solution to (AVD)*, with $x(1) = (1, 1, 1)$ and $\dot{x}(1) = -\nabla\phi\big(x(1)\big)$ (the direction of maximum descent).*

The oscillatory nature exhibited by the trajectories arises a problem when an approximation of the solution is needed. In that case, the precision of the approximation will vary depending on where the time interval concludes. In order to avoid this undesirable behavior, and partly inspired by a continuous version of Newton's method [6], Attouch, Peypouquet and Redont [15] proposed a Dynamic Inertial Newton system with Asymptotically Vanishing Damping, given by

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla\phi(x(t)) + \beta\nabla^2\phi(x(t))\dot{x}(t) = 0, \tag{DIN-AVD}$$

where $\alpha, \beta > 0$. The authors presented (DIN-AVD) as a continuous-time model for the design of new algorithms, a line of research already outlined in [15], and continued in [10]. Although (DIN-AVD) may initially appear more intricate to deal with, due to the second order information required on ϕ, a remarkable feature is that it can be rewritten as a first

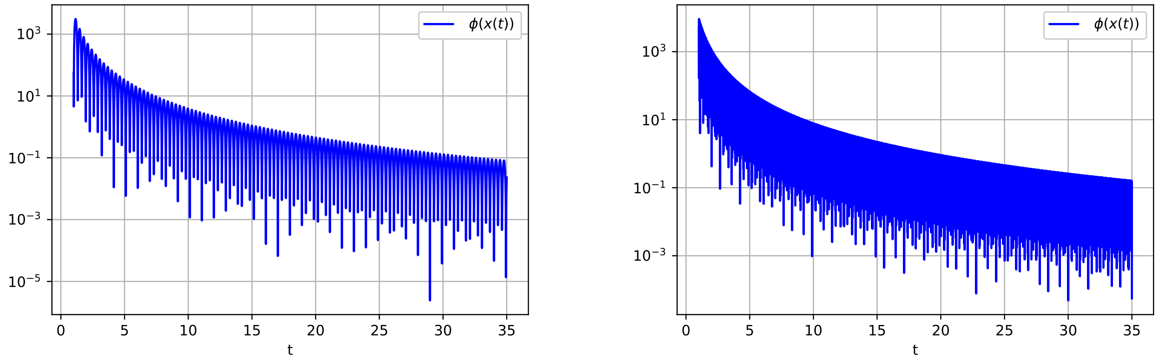


Figure 2.1: Depiction of the function values according to Example 2.1.1, on the interval $[1, 35]$, for $\alpha = 3.1$, and $\rho = 10$ (left) and $\rho = 100$ (right).

order system depending only on the gradient of the function. In terms of the convergence, the function values vanish along the solutions, with the same rates as for (AVD). Nevertheless, in contrast with the solutions of (AVD), the oscillations are tame.

Example 2.1.2. *In the context of Example 2.1.1, Figure 2.2 shows the behavior of the solution to* (AVD) *in comparison with that of* (DIN-AVD), *both with* $x(1) = (1, 1, 1)$ *and* $\dot{x}(1) = -\nabla\phi\big(x(1)\big)$.

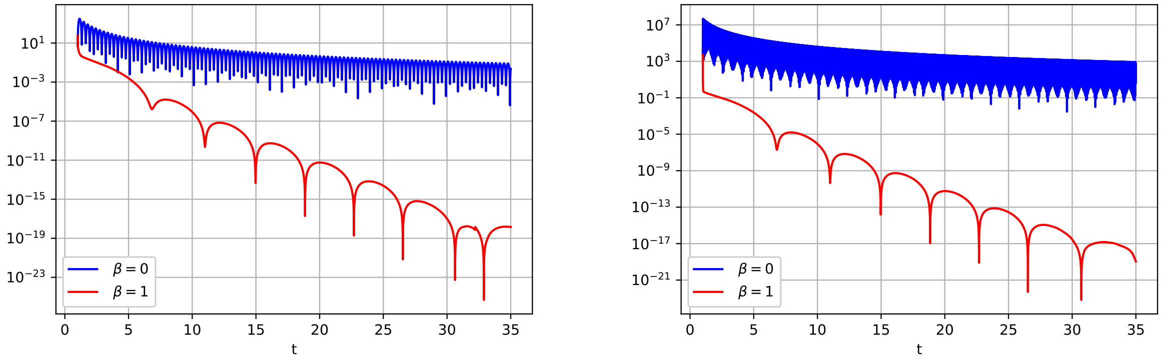


Figure 2.2: Depiction of the function values according to Example 2.1.2, on the interval $[1, 35]$, for $\alpha = 3.1$, $\beta = 1$, and $\rho = 10$ (left) and $\rho = 100$ (right).

An alternative way to avoid (or at least moderate) the oscillations exemplified in Figure 2.1 for the solutions of (AVD) is to stop the evolution and restart it with zero initial velocity, from time to time. The simplest option is to do so periodically, at fixed intervals. This idea is used in [69] for the accelerated gradient method, where the number of iterations between restarts that depends on the parameter of strong convexity of the function. See also [66, 3, 16], where the problem of estimating the appropriate restart times is addressed. An adaptive policy for the restarting of Nesterov's Method was proposed by O'Donoghue and Candès in [71], where the algorithm is restarted at the first iteration k such that $\phi(x_{k+1}) > \phi(x_k)$, which prevents the function values to increase locally. This kind of restarting criteria shows a remarkable performance, although convergence rate guarantees have not been established, some partial steps in this direction have been made in [47, 56]. Moreover, the authors of [71] observe that this heuristic displays an erratic behavior when the difference $\phi(x_k) - \phi(x_{k+1})$ is small, due to the prevalence of cancellation errors. Therefore, this method must be handled with care if high accuracy is desired. A different restarting scheme, based on the speed of the trajectories, is proposed for (AVD) in [83], where rates of convergence are established. Also, in [83], the authors also perform numerical tests using Nesterov's inertial gradient method, with this restarting scheme as a heuristic, and observe a faster convergence to the optimal value.

The aim of this first part is to analyze the impact that the speed restarting scheme has on the solutions of (DIN-AVD), in order to set the theoretical foundations to further accelerate Hessian driven inertial algorithms (like the ones in [10]) by means of a restarting policy. This approach combines two oscillation mitigation principles that result in a monotonic and fast convergence of the function values. Linear convergence rates for functions with quadratic growth are provided, and it is observed a noticeable improvement in the behavior of the trajectories in terms of stability and convergence speed, both in comparison with the non-restarted trajectories, and with the restarted solutions of (AVD). As a byproduct, we generalize and improve some of the results in [83]. It is worth noticing that the convergence rate result holds for all values of $\alpha > 0$ and $\beta \geq 0$, in contrast with those in [12, 15, 10].

Consider the ordinary differential equation (DIN-AVD) with initial conditions $x(0) = x_0$, $\dot{x}(0) = 0$, and parameters $\alpha > 0$ and $\beta \geq 0$. A *solution* is a function in $\mathcal{C}^2((0, +\infty); \mathbb{R}^n) \cap \mathcal{C}^1([0, +\infty); \mathbb{R}^n)$, such that $x(0) = x_0$, $\dot{x}(0) = 0$ and (DIN-AVD) holds for every $t > 0$. Existence and uniqueness of such a solution is not straightforward due to the singularity at $t = 0$, but can be established by a limiting procedure. As shown in Appendix 2.7, we have the following:

Theorem 2.1.1. *For every $x_0 \in \mathbb{R}^n$, the differential equation* (DIN-AVD)*, with initial conditions $x(0) = x_0$ and $\dot{x}(0) = 0$, has a unique solution.*

We are concerned with the design and analysis of a restart scheme to accelerate the convergence of the solutions of (DIN-AVD) to minimizers of ϕ, based on the method proposed in [83].

2.2 A speed restarting scheme and the main theoretical result

Since the damping coefficient α/t goes to 0 as $t \to \infty$, large values of t result in a smaller stabilization of the trajectory. The idea is thus to restart the dynamics at the point where the speed ceases to increase.

Given $z \in \mathbb{R}^n$, let y_z be the solution of (DIN-AVD), with initial conditions $y_z(0) = z$ and $\dot{y}_z(0) = 0$. Set

$$T(z) = \inf \left\{ t > 0 \ : \ \frac{d}{dt} \|\dot{y}_z(t)\|^2 \leq 0 \right\}. \tag{2.2}$$

Remark 2.2.1. *Take $z \notin \operatorname{argmin}(\phi)$, and define y_z as above. For $t \in \big(0, T(z)\big)$, we have*

$$\begin{aligned}\frac{d}{dt}\phi(y_z(t)) &= \langle \nabla\phi(y_z(t)), \dot{y}_z(t)\rangle \\ &= -\langle \ddot{y}_z(t), \dot{y}_z(t)\rangle - \frac{\alpha}{t}\left\|\dot{y}_z(t)\right\|^2 - \beta\langle \nabla^2\phi(y_z(t))\dot{y}_z(t), \dot{y}_z(t)\rangle.\end{aligned}$$

But $\langle \nabla^2\phi(y_z(t))\dot{y}_z(t), \dot{y}_z(t)\rangle \geq 0$ by convexity, and $\langle \ddot{y}_z(t), \dot{y}_z(t)\rangle \geq 0$ by the definition of $T(z)$. Therefore,

$$\frac{d}{dt}\phi(y_z(t)) \leq -\frac{\alpha}{t}\left\|\dot{y}_z(t)\right\|^2. \tag{2.3}$$

In particular, $t \mapsto \phi\big(y_z(t)\big)$ decreases on $[0, T(z)]$.

If $z \notin \operatorname{argmin}(\phi)$, then $T(z)$ cannot be 0. In fact, we shall prove (see Corollaries 2.3.2 and 2.3.3) that

$$0 < \inf\big\{T(z) : z \notin \operatorname{argmin}(\phi)\big\} \leq \sup\big\{T(z) : z \notin \operatorname{argmin}(\phi)\big\} < \infty. \tag{2.4}$$

Definition 2.2.1. *Given $x_0 \in \mathbb{R}^n$, the restarted trajectory $\chi_{x_0} : [0, \infty) \to \mathbb{R}^n$ is defined inductively:*

1. *First, compute y_{x_0}, $T_1 = T(x_0)$ and $S_1 = T_1$, and define $\chi_{x_0}(t) = y_{x_0}(t)$ for $t \in [0, S_1]$.*
2. *For $i \geq 1$, having defined $\chi_{x_0}(t)$ for $t \in [0, S_i]$, set $x_i = \chi_{x_0}(S_i)$, and compute y_{x_i}. Then, set $T_{i+1} = T(x_i)$ and $S_{i+1} = S_i + T_{i+1}$, and define $\chi_{x_0}(t) = y_{x_i}(t - S_i)$ for $t \in (S_i, S_{i+1}]$.*

In view of (2.4), S_i is defined for all $i \geq 1$, $\inf_{i\geq 1}(S_{i+1} - S_i) > 0$ and $\lim_{i\to\infty} S_i = \infty$. Moreover, in view of Remark 2.2.1, we have

Proposition 2.2.1. *The function $t \mapsto \phi\big(\chi_{x_0}(t)\big)$ is nonincreasing on $[0, \infty)$.*

Our main theoretical result establishes that $\phi\big(\chi_{x_0}(t)\big)$ converges linearly to ϕ^*, provided there exists $\mu > 0$ such that

$$\mu(\phi(z) - \phi^*) \leq \frac{1}{2}\left\|\nabla\phi(z)\right\|^2 \tag{2.5}$$

for all $z \in \mathbb{R}^n$. The Łojasiewicz inequality (2.5) is equivalent to quadratic growth and is implied by strong convexity (see [21]). More precisely, we have the following:

Theorem 2.2.1. *Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be convex and twice continuously differentiable. Assume $\nabla\phi$ is Lipschitz-continuous with constant $L > 0$, that there exists $\mu > 0$ such that (2.5) holds, and that the minimum value ϕ^* of ϕ is attained. Given $\alpha > 0$ and $\beta \geq 0$, let be the restarted trajectory defined by (DIN-AVD) from an initial point $x_0 \in \mathbb{R}^n$. Then, there exist constants $C, K > 0$ such that*

$$\phi\big(\chi_{x_0}(t)\big) - \phi^* \leq Ce^{-Kt}\big(\phi(x_0) - \phi^*\big) \leq \frac{CL}{2}e^{-Kt}\operatorname{dist}\left(x_0, \operatorname{argmin}(\phi)\right)^2$$

for all $t > 0$.

The rather technical proof is split into several parts and presented in the next subsections.

2.3 Technicalities

Throughout this section, we fix $z \notin \operatorname{argmin}(\phi)$ and, in order to simplify the notation, we denote by x (instead of y_z) the solution of (DIN-AVD) with initial condition $x(0) = z$ and $\dot{x}(0) = 0$.

2.3.1 A few useful bounds

We begin by defining some useful auxiliary functions and point out the main relationships between them.

To this end, we first rewrite equation (DIN-AVD) as

$$\frac{d}{dt}(t^\alpha \dot{x}(t)) = -t^\alpha \nabla\phi(x(t)) - \beta t^\alpha \nabla^2\phi(x(t))\dot{x}(t). \tag{2.6}$$

Integrating (2.6) over $[0, t]$, we get

$$\begin{aligned} t^\alpha \dot{x}(t) &= -\int_0^t u^\alpha \nabla\phi(x(u))\,du - \beta\int_0^t u^\alpha \nabla^2\phi(x(u))\dot{x}(u)\,du \\ &= -\left[\int_0^t u^\alpha(\nabla\phi(x(u)) - \nabla\phi(z))\,du\right] - \left[\beta\int_0^t u^\alpha \nabla^2\phi(x(u))\dot{x}(u)\,du\right] \\ &\quad - \frac{t^{\alpha+1}}{\alpha+1}\nabla\phi(z). \end{aligned} \tag{2.7}$$

In order to obtain an upper bound for the speed $\dot{x}$, the integrals

$$I_z(t) = \int_0^t u^\alpha(\nabla\phi(x(u)) - \nabla\phi(z))\,du, \quad J_z(t) = \beta\int_0^t u^\alpha \nabla^2\phi(x(u))\dot{x}(u)\,du \tag{2.8}$$

will be majorized using the function

$$M_z(t) = \sup_{u\in(0,t]}\left[\frac{\|\dot{x}(u)\|}{u}\right], \tag{2.9}$$

which is positive, nondecreasing and continuous.

Lemma 2.3.1. *For every $t > 0$, we have*

$$\|I_z(t)\| \leq \frac{LM_z(t)t^{\alpha+3}}{2(\alpha+3)} \qquad \text{and} \qquad \|J_z(t)\| \leq \frac{\beta LM_z(t)t^{\alpha+2}}{\alpha+2}.$$

Proof. For the first estimation, we use the Lipschitz-continuity of $\nabla\phi$ and the fact that M in nondecreasing, to obtain

$$\begin{aligned} \|\nabla\phi(x(u)) - \nabla\phi(z)\| &\leq L\|x(u) - z\| \\ &\leq L\left\|\int_0^u \dot{x}(s)\,ds\right\| \end{aligned}$$

$$\begin{aligned}
&\leq L \int_0^u s \frac{\|\dot{x}(s)\|}{s}\, ds \\
&\leq L M_z(u) \int_0^u s\, ds,
\end{aligned}$$

which results in

$$\|\nabla \phi(x(u)) - \nabla \phi(z)\| \leq \frac{L u^2 M_z(u)}{2} \tag{2.10}$$

Then, from the definition of $I_z(t)$ we deduce that

$$\begin{aligned}
\|I_z(t)\| &\leq \int_0^t u^\alpha \|\nabla \phi(x(u)) - \nabla \phi(z)\|\, du \\
&\leq \frac{L M_z(t)}{2} \int_0^t u^{\alpha+2}\, du \\
&= \frac{L M_z(t) t^{\alpha+3}}{2(\alpha+3)}.
\end{aligned}$$

For the second inequality, we proceed analogously to get

$$\begin{aligned}
\left\|\nabla^2 \phi(x(u)) \dot{x}(u)\right\| &= \left\| \lim_{r \to u} \frac{\nabla \phi(x(r)) - \nabla \phi(x(u))}{r-u} \right\| \\
&\leq \lim_{r \to u} \frac{L}{r-u} \int_u^r \|\dot{x}(s)\|\, ds \\
&\leq \lim_{r \to u} \frac{L M_z(r)}{r-u} \int_u^r s\, ds,
\end{aligned}$$

which yields

$$\left\|\nabla^2 \phi(x(u)) \dot{x}(u)\right\| \leq L u M_z(u). \tag{2.11}$$

Then,

$$\|J_z(t)\| \leq \beta \int_0^t u^\alpha \left\|\nabla^2 \phi(x(u)) \dot{x}(u)\right\| du \leq \beta \int_0^t u^{\alpha+1} L M_z(u)\, du \leq \frac{\beta L M_z(t) t^{\alpha+2}}{\alpha+2},$$

as claimed. □

The dependence of M_z on the initial condition z may be greatly simplified. To this end, set

$$H(t) = 1 - \frac{L\beta t}{(\alpha+2)} - \frac{L t^2}{2(\alpha+3)}. \tag{2.12}$$

The function H is concave, quadratic, does not depend on z, and has exactly one positive zero, given by

$$\tau_1 = -\left(\frac{\alpha+3}{\alpha+2}\right)\beta + \sqrt{\left(\frac{\alpha+3}{\alpha+2}\right)^2 \beta^2 + \frac{2(\alpha+3)}{L}}. \tag{2.13}$$

In particular, H decreases strictly from 1 to 0 on $[0, \tau_1]$.

Lemma 2.3.2. *For every $t \in (0, \tau_1)$,*

$$M_z(t) \leq \frac{\|\nabla \phi(z)\|}{(\alpha+1) H(t)}. \tag{2.14}$$

Proof. If $0 < u \le t$, using (2.7) and (2.8), along with Lemma 2.3.1, we obtain

$$\frac{\|\dot{x}(u)\|}{u} \le \frac{\|I_z(u) + J_z(u)\|}{u^{\alpha+1}} + \frac{\|\nabla\phi(z)\|}{\alpha+1} \le \left[\frac{Lu^2}{2(\alpha+3)} + \frac{L\beta u}{\alpha+2}\right] M_z(u) + \frac{\|\nabla\phi(z)\|}{\alpha+1}. \tag{2.15}$$

Since the right-hand side is nondecreasing in t, we take the supremum for $u \in [0,t]$ to deduce that

$$M_z(t) \le \left[\frac{Lt^2}{2(\alpha+3)} + \frac{L\beta t}{\alpha+2}\right] M_z(t) + \frac{\|\nabla\phi(z)\|}{\alpha+1}.$$

Rearranging the terms, and using the definition of H, given in (2.12), we see that

$$H(t)M_z(t) \le \frac{\|\nabla\phi(z)\|}{(\alpha+1)}.$$

We conclude by observing that H is positive on $(0,\tau_1)$. □

By combining Lemmas 2.3.1 and 2.3.2, and inequalities (2.10) and (2.11), we obtain:

Corollary 2.3.1. *For every $t \in (0,\tau_1)$, we have*

$$\|I_z(t) + J_z(t)\| \le t^{\alpha+1}\left[\frac{1-H(t)}{H(t)}\right]\frac{\|\nabla\phi(z)\|}{(\alpha+1)}$$

$$\left\|\big(\nabla\phi(x(t)) - \nabla\phi(z)\big) + \beta\nabla^2\phi(x(t))\dot{x}(t)\right\| \le \left[\frac{Lt^2}{2} + \beta Lt\right]\frac{\|\nabla\phi(z)\|}{(\alpha+1)H(t)}.$$

We highlight the fact that the bound above depends on z only via the factor $\|\nabla\phi(z)\|$.

2.3.2 Estimates for the restarting time

We begin by finding a lower bound for the restarting time, depending on the parameters α, β and L, but not on the initial condition z.

Lemma 2.3.3. *Let $z \notin \operatorname{argmin}(\phi)$, and let x be the solution of* (DIN-AVD) *with initial conditions $x(0) = z$ and $\dot{x}(0) = 0$. For every $t \in (0,\tau_1)$, we have*

$$\langle \dot{x}(t), \ddot{x}(t)\rangle \ge \frac{t\,\|\nabla\phi(z)\|^2}{(\alpha+1)^2H(t)^2}\left(1 - \frac{(2\alpha+3)\beta Lt}{(\alpha+2)} - \frac{(\alpha+2)Lt^2}{(\alpha+3)}\right).$$

Proof. From (2.7) and (2.8), we know that

$$\dot{x}(t) = -\frac{1}{t^\alpha}\big(I_z(t) + J_z(t)\big) - \frac{t}{\alpha+1}\nabla\phi(z). \tag{2.16}$$

On the other hand,

$$\frac{d}{dt}\left[\frac{1}{t^\alpha}\big(I_z(t) + J_z(t)\big)\right] = -\frac{\alpha}{t^{\alpha+1}}\big(I_z(t) + J_z(t)\big) + \big(\nabla\phi\big(x(t)\big) - \nabla\phi(z)\big) + \beta\nabla^2\phi(x(t))\dot{x}(t).$$

Then,

$$\begin{aligned}\ddot{x}(t) &= \frac{\alpha}{t^{\alpha+1}}\big(I_z(t) + J_z(t)\big) - (\nabla\phi(x(t)) - \nabla\phi(z)) - \beta\nabla^2\phi(x(t))\dot{x}(t)\\ &\quad - \frac{1}{\alpha+1}\nabla\phi(z)\\ &= A(t) - B(t),\end{aligned}$$

where

$$\begin{aligned}A(t) &= \frac{\alpha}{t^{\alpha+1}}\big(I_z(t) + J_z(t)\big) - \frac{1}{\alpha+1}\nabla\phi(z),\\ B(t) &= \big(\nabla\phi(x(t)) - \nabla\phi(z)\big) + \beta\nabla^2\phi(x(t))\dot{x}(t).\end{aligned}$$

With this notation, we have

$$\langle\dot{x}(t), \ddot{x}(t)\rangle = \langle\dot{x}(t), A(t)\rangle - \langle\dot{x}(t), B(t)\rangle \geq \langle\dot{x}(t), A(t)\rangle - \|\dot{x}(t)\|\,\|B(t)\|.$$

For the first term, we do as follows:

$$\begin{aligned}\langle\dot{x}(t), A(t)\rangle &= -\left\langle \frac{1}{t^\alpha}\big(I_z(t) + J_z(t)\big) + \frac{t}{\alpha+1}\nabla\phi(z), \frac{\alpha}{t^{\alpha+1}}\big(I_z(t) + J_z(t)\big) - \frac{1}{\alpha+1}\nabla\phi(z)\right\rangle\\ &\geq \frac{t}{(\alpha+1)^2}\|\nabla\phi(z)\|^2 - \frac{\alpha}{t^{2\alpha+1}}\|I_z(t) + J_z(t)\|^2\\ &\quad - \frac{(\alpha-1)}{t^\alpha(\alpha+1)}\|\nabla\phi(z)\|\,\|I_z(t) + J_z(t)\|\\ &\geq \frac{t}{(\alpha+1)^2}\|\nabla\phi(z)\|^2 - \frac{\alpha t}{(\alpha+1)^2}\left[\frac{1-H(t)}{H(t)}\right]^2\|\nabla\phi(z)\|^2\\ &\quad - \frac{(\alpha-1)t}{(\alpha+1)^2}\left[\frac{1-H(t)}{H(t)}\right]\|\nabla\phi(z)\|^2\\ &= \frac{t\,\|\nabla\phi(z)\|^2}{(\alpha+1)^2}\left(1 - \alpha\left[\frac{1-H(t)}{H(t)}\right]^2 - (\alpha-1)\left[\frac{1-H(t)}{H(t)}\right]\right)\\ &= \frac{t\,\|\nabla\phi(z)\|^2}{(\alpha+1)^2H(t)^2}\Big(H(t)^2 - \alpha\big(1-H(t)\big)^2 - (\alpha-1)H(t)\big(1-H(t)\big)\Big)\\ &= \frac{t\,\|\nabla\phi(z)\|^2}{(\alpha+1)^2H(t)^2}\big((\alpha+1)H(t) - \alpha\big),\end{aligned}$$

where we have used the Cauchy-Schwarz inequality and Corollary 2.3.1. For the second term, we first use (2.16) and observe that

$$\|\dot{x}(t)\| \leq \frac{1}{t^\alpha}\|I_z(t) + J_z(t)\| + \frac{t}{(\alpha+1)}\|\nabla\phi(z)\| \leq \frac{t\,\|\nabla\phi(z)\|}{(\alpha+1)H(t)},$$

and

$$B(t) \leq \left[\frac{Lt^2}{2} + \beta Lt\right]\frac{\|\nabla\phi(z)\|}{(\alpha+1)H(t)},$$

by Corollary 2.3.1. We conclude that

$$\langle\dot{x}(t), \ddot{x}(t)\rangle \geq \frac{t\,\|\nabla\phi(z)\|^2}{(\alpha+1)^2H(t)^2}\left((\alpha+1)H(t) - \alpha - \frac{Lt^2}{2} - \beta Lt\right)$$

$$= \frac{t\,\|\nabla\phi(z)\|^2}{(\alpha+1)^2 H(t)^2}\left(1-\frac{(2\alpha+3)\beta L t}{(\alpha+2)}-\frac{(\alpha+2)Lt^2}{(\alpha+3)}\right),$$

as stated. □

The function G, defined by

$$G(t) = 1-\frac{(2\alpha+3)\beta L t}{(\alpha+2)}-\frac{(\alpha+2)Lt^2}{(\alpha+3)} = (\alpha+1)H(t)-\alpha-\frac{Lt^2}{2}-\beta L t, \tag{2.17}$$

does not depend on the initial condition z. Its unique positive zero is

$$\tau_3 = -\frac{(\alpha+3)(2\alpha+3)}{2(\alpha+2)^2}\beta+\sqrt{\frac{(\alpha+3)^2(2\alpha+3)^2}{4(\alpha+2)^4}\beta^2+\frac{(\alpha+3)}{(\alpha+2)L}}. \tag{2.18}$$

In view of the definition of the restarting time, an immediate consequence of Lemma 2.3.3 is

Corollary 2.3.2. *Let $T_* = \inf\{T(z) : z \notin \operatorname{argmin}(\phi)\}$. Then, $\tau_3 \le T_*$.*

Remark 2.3.1. *If $\beta = 0$, then*

$$\tau_3 = \sqrt{\frac{(\alpha+3)}{(\alpha+2)L}}.$$

The case $\alpha = 3$ and $\beta = 0$ was studied in [83], and the authors provided $\frac{4}{5\sqrt{L}}$ as a lower bound for the restart. The arguments presented here yield a higher bound, since

$$\tau_3 = \sqrt{\frac{6}{5L}} > \frac{1}{\sqrt{L}} > \frac{4}{5\sqrt{L}}.$$

Recall that the function H given in (2.12) decreases from 1 to 0 on $[0,\tau_1]$. Therefore, $H(t) > \frac{1}{2}$ for all $t \in [0,\tau_2)$, where

$$\tau_2 = H^{-1}\left(\tfrac{1}{2}\right) = -\left(\frac{\alpha+3}{\alpha+2}\right)\beta+\sqrt{\left(\frac{\alpha+3}{\alpha+2}\right)^2\beta^2+\frac{\alpha+3}{L}} < \tau_1. \tag{2.19}$$

Evaluating the right-hand side of (2.17), we see that

$$G(\tau_2) = \frac{(1-\alpha)-L\tau_2^2-2\beta L\tau_2}{2} < 0,$$

whence

$$\tau_1 > \tau_2 > \tau_3 > 0. \tag{2.20}$$

These facts will be useful to provide an upper bound for the restarting time.

Proposition 2.3.1. *Let $z \notin \operatorname{argmin}(\phi)$, and let x be the solution of* (DIN-AVD) *with initial conditions $x(0) = z$ and $\dot{x}(0) = 0$. Let ϕ satisfy* (2.5) *with $\mu > 0$. For each $\tau \in (0,\tau_2)\cap(0,T(z)]$, we have*

$$T(z) \le \tau \exp\left[\frac{(\alpha+1)^2}{2\alpha\mu\tau^2\Psi(\tau)}\right], \qquad \text{where} \qquad \Psi(\tau) = \left[2-\frac{1}{H(\tau)}\right]^2.$$

Proof. In view of (2.7) and (2.8), we can use Corollary 2.3.1 to obtain

$$\left\|\dot{x}(\tau)+\frac{\tau}{\alpha+1}\nabla\phi(z)\right\| = \frac{1}{\tau^\alpha}\left\|I(\tau)+J(\tau)\right\| \leq \tau\left[\frac{1}{H(\tau)}-1\right]\frac{\|\nabla\phi(z)\|}{(\alpha+1)}.$$

From the (reverse) triangle inequality and the definition of H, it ensues that

$$\|\dot{x}(\tau)\| \geq \frac{\tau\,\|\nabla\phi(z)\|}{\alpha+1} - \tau\left[\frac{1}{H(\tau)}-1\right]\frac{\|\nabla\phi(z)\|}{(\alpha+1)} = \tau\left[2-\frac{1}{H(\tau)}\right]\frac{\|\nabla\phi(z)\|}{\alpha+1}, \tag{2.21}$$

which is positive, because $\tau \in (0,\tau_2)$. Now, take $t \in [\tau, T(z)]$. Since $\|\dot{x}(t)\|^2$ increases on $\big[0,T(z)\big]$, Remark 2.2.1 gives

$$\frac{d}{dt}\phi\big(x(t)\big) \leq -\frac{\alpha}{t}\|\dot{x}(t)\|^2 \leq -\frac{\alpha}{t}\|\dot{x}(\tau)\|^2 \leq -\frac{1}{t}\left[\frac{\alpha\tau^2\Psi(\tau)\,\|\nabla\phi(z)\|^2}{(\alpha+1)^2}\right].$$

Integrating over $[\tau, T(z)]$, we get

$$\phi\big(x(T(z))\big)-\phi\big(x(\tau)\big) \leq -\left[\frac{\alpha\tau^2\Psi(\tau)\,\|\nabla\phi(z)\|^2}{(\alpha+1)^2}\right]\ln\left[\frac{T(z)}{\tau}\right]. \tag{2.22}$$

It follows that

$$\left[\frac{\alpha\tau^2\Psi(\tau)\,\|\nabla\phi(z)\|^2}{(\alpha+1)^2}\right]\ln\left[\frac{T(z)}{\tau}\right] \leq \phi\big(x(\tau)\big)-\phi\big(x(T(z))\big) \leq \phi(z)-\phi^* \leq \frac{\|\nabla\phi(z)\|^2}{2\mu},$$

in view of (2.5). It suffices to rearrange the terms to conclude. □

Corollary 2.3.3. *Let ϕ satisfy* (2.5) *with $\mu>0$, and let $\tau_* \in (0,\tau_2)\cap(0,T_*]$. Then,*

$$\sup\big\{T(z) : z \notin \operatorname{argmin}(\phi)\big\} \leq \tau_* \exp\left[\frac{(\alpha+1)^2}{2\alpha\mu\tau_*^2\Psi(\tau_*)}\right].$$

2.4 Function value decrease and proof of Theorem 2.2.1

The next result provides the ratio at which the function values have been reduced by the time the trajectory is restarted.

Proposition 2.4.1. *Let $z \notin$ argmin(ϕ), and let x be the solution of* (DIN-AVD) *with initial conditions $x(0)=z$ and $\dot{x}(0)=0$. Let ϕ satisfy* (2.5) *with $\mu>0$. For each $\tau \in (0,\tau_2)\cap(0,T(z)]$, we have*

$$\phi\big(x(t)\big)-\phi^* \leq \left[1-\frac{\alpha\mu\tau^2\Psi(\tau)}{(\alpha+1)^2}\right]\big(\phi(z)-\phi^*\big)$$

for every $t \in [\tau, T(z)]$.

Proof. Take $s \in (0, \tau)$. By combining Remark 2.2.1 with (2.21), we obtain

$$\frac{d}{ds}\phi(x(s)) \leq -\frac{\alpha}{s}\left\|\dot{x}(s)\right\|^2 \leq -\frac{\alpha s \left\|\nabla\phi(z)\right\|^2}{(\alpha+1)^2}\left[2 - \frac{1}{H(s)}\right]^2 \leq -\frac{\alpha s \left\|\nabla\phi(z)\right\|^2}{(\alpha+1)^2}\Psi(\tau)$$

because H decreases on $(0, \tau_1)$, which contains $(0, \tau)$. Integrating on $(0, \tau)$ and using (2.5), we obtain

$$\phi\big(x(\tau)\big) - \phi^* \leq \phi(z) - \phi^* - \frac{\alpha\tau^2\Psi(\tau)\left\|\nabla\phi(z)\right\|^2}{2(\alpha+1)^2} \leq \left[1 - \frac{\alpha\mu\tau^2\Psi(\tau)}{(\alpha+1)^2}\right]\big(\phi(z) - \phi^*\big).$$

To conclude, it suffices to observe that $\phi\big(x(t)\big) \leq \phi\big(x(\tau)\big)$ in view of Remark 2.2.1. □

Remark 2.4.1. *Since Ψ is decreasing in $[0, \tau_2)$, we have $\Psi(t) \geq \Psi(\tau_*) > 0$, whenever $0 \leq t \leq \tau_* < \tau_2$. Moreover, in view of (2.20) and Corollary 2.3.2, we can take $\tau_* = \tau_3$ to obtain a lower bound. If $\beta = 0$, we obtain*

$$\Psi(t) \geq \Psi(\tau_3) = \left[2 - \frac{1}{H(\tau_3)}\right]^2 = \left[2 - \frac{1}{1 - \frac{1}{2(\alpha+2)}}\right]^2 = \left[\frac{2\alpha+2}{2\alpha+3}\right]^2,$$

which is independent of L. As a consequence, the inequality in Proposition 2.4.1 becomes

$$\phi\big(x(t)\big) - \phi^* \leq \left(1 - \frac{4\alpha(\alpha+3)}{(\alpha+2)(2\alpha+3)^2}\frac{\mu}{L}\right)(\phi(x_0) - \phi^*).$$

For $\alpha = 3$, this gives

$$\phi\big(x(t)\big) - \phi^* \leq \left(1 - \frac{8}{45}\frac{\mu}{L}\right)(\phi(x_0) - \phi^*).$$

For this particular case, a similar result, obtained in [83] for strongly convex functions, namely

$$\phi\big(x(t)\big) - \phi^* \leq \left(1 - \frac{3}{25}\left(\frac{67}{71}\right)^2\frac{\mu}{L}\right)(\phi(x_0) - \phi^*).$$

Our constant is approximately 66.37% larger than the one from [83], which implies a greater reduction in the function values each time the trajectory is restarted. On the other hand, if $\beta > 0$, we can still obtain a slightly smaller lower bound, namely $\Psi(\tau_3) > \left(\frac{2\alpha+1}{2\alpha+2}\right)^2$, independent from β and L. The proof is technical and it can be found on Appendix 2.8.

Proof of Theorem 2.2.1

Adopt the notation in Definition 2.2.1, take any $\tau_* \in (0, \tau_2) \cap (0, T_*]$, and set

$$\tau^* = \tau_* \exp\left[\frac{(\alpha+1)^2}{2\alpha\mu\tau_*^2\Psi(\tau_*)}\right], \qquad \text{where} \qquad \Psi(\tau_*) = \left[2 - \frac{1}{H(\tau_*)}\right]^2.$$

In view of Corollaries 2.3.2 and 2.3.3, we have

$$\tau_* \leq T(x_i) \leq \tau^*$$

for all $i \geq 0$ (we assume $x_i \notin \operatorname{argmin}(\phi)$ since the result is trivial otherwise). Given $t > 0$, let m be the largest positive integer such that $m\tau^* \leq t$. By time t, the trajectory will have been restarted at least m times. By Proposition 2.2.1, we know that

$$\phi\big(\chi_{x_0}(t)\big) \leq \phi\big(\chi_{x_0}(m\tau^*)\big) \leq \phi\big(\chi_{x_0}(m\tau_*)\big).$$

We may now apply Proposition 2.4.1 repeatedly to deduce that

$$\phi\big(\chi_{x_0}(t)\big) - \phi^* \leq Q^m\big(\phi(x_0) - \phi^*\big) \qquad \text{where} \qquad Q = \left[1 - \frac{\alpha\mu\tau_*^2\Psi(\tau_*)}{(\alpha+1)^2}\right] < 1.$$

By definition, $(m+1)\tau^* > t$, which entails $m > \frac{t}{\tau^*} - 1$. Since $Q \in (0,1)$, we have

$$Q^m \leq Q^{\frac{t}{\tau^*}-1} = \frac{1}{Q}\exp\left(\frac{\ln(Q)}{\tau^*}\,t\right),$$

and the result is established, with $C = Q^{-1}$ and $K = -\frac{1}{\tau^*}\ln(Q)$. The proof is finished due to the fact that $\phi(u) \leq \phi^* + \frac{L}{2}\|u - u^*\|^2$ for every $u^* \in \operatorname{argmin}(\phi)$. □

The convergence rate given in Theorem 2.2.1, holds for C and K of the form

$$C = C(\tau_*) = \left[1 - \frac{\alpha\mu\tau_*^2\Psi(\tau_*)}{(\alpha+1)^2}\right]^{-1}$$

and

$$\begin{aligned} K = K(\tau_*) &= -\frac{1}{\tau_*}\exp\left[-\frac{(\alpha+1)^2}{2\alpha\mu\tau_*^2\Psi(\tau_*)}\right]\ln\left[1 - \frac{\alpha\mu\tau_*^2\Psi(\tau_*)}{(\alpha+1)^2}\right] \\ &> \frac{\alpha\mu\tau_*\Psi(\tau_*)}{(\alpha+1)^2}\exp\left[-\frac{(\alpha+1)^2}{2\alpha\mu\tau_*^2\Psi(\tau_*)}\right], \end{aligned}$$

for any $\tau_* \in (0,\tau_2) \cap (0, T_*]$. In view of (2.20) and Corollary 2.3.2, $\tau_* = \tau_3$ is a valid choice. On the other hand, the function $K(\cdot)$ vanishes at $\tau \in \{0, \tau_2\}$ and is positive on $(0,\tau_2)$. By continuity, it attains its maximum at some $\hat{\tau}_* \in (0,\tau_2) \cap (0, T_*]$. Therefore, $K(\hat{\tau}_*)$ yields the fastest convergence rate prediction in this framework.

Remark 2.4.2. *It is possible to implement a fixed restart scheme. To this end, we modify Definition 2.2.1 by setting $T_i \equiv \tau$, with any $\tau \in (0,\tau_2) \cap (0,T_*]$, such as $\hat{\tau}_*$ or τ_3, for example. In theory, $\hat{\tau}_*$ gives the same convergence rate as the original restart scheme presented throughout this work. From a practical perspective, though, restarting the dynamics too soon may result in a poorer performance. Therefore, finding larger values of $\hat{\tau}_*$ and τ_3 is crucial to implement a fixed restart (see Remarks 2.3.1 and 2.4.1).*

2.5 Numerical illustration

In this section, we provide a very simple numerical example to illustrate how the convergence is improved by the restarting scheme. A more thorough numerical analysis will be carried out in a forthcoming work.

Example 2.1.2 revisited

We consider the quadratic function $\phi : \mathbb{R}^3 \to \mathbb{R}$, defined in Example 2.1.1 by (2.1), with $\rho = 10$. We set $\alpha = 3.1$ and $\beta = 0.25$, and compute the solutions of (AVD) and (DIN-AVD), starting from $x(1) = x_1 = (1, 1, 1)$ and zero initial velocity, with and without restarting, using the `Python` tool `odeint` from the `scipy` package. Figure 2.3 shows a comparison of the values along the trajectory with and without restarting, first for (AVD), and then for (DIN-AVD). In both cases, the restarted trajectories appear to be more stable and converge faster.

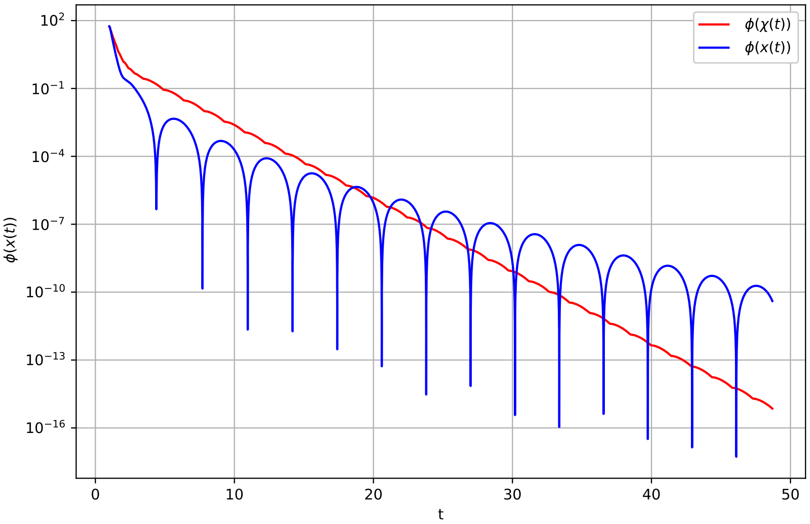


Figure 2.3: Values along the trajectory, with (red) and without (blue) restarting, for (DIN-AVD).

However, one can do better. As mentioned earlier, restarting schemes based on function values are effective from a practical perspective, but show an erratic behavior as the trajectory approaches a minimizer. It seems natural as a heuristic to use the first (or n-th) function-value restart point as a warm start, and then apply speed restarts, for which we have obtained convergence rate guarantees. Although the velocity must be set to zero *after* each restart, there are no constraints on the initial velocity used to compute the warm starting point. The results are shown in Figure 2.4, with initial velocity set to zero and $\dot{x}(1) = -\beta\nabla\phi(x_1)$, respectively.

A linear regression after the first restart provides estimations for the linear convergence rate of the function values along the corresponding trajectories, when modeled as $\phi\big(\chi(t)\big) \sim Ae^{-Bt}$, with $A, B > 0$. The results are displayed in Table 2.1. The absolute value of the exponent B in the linear convergence rate is increased by 34,67% in the case $\dot{x}(1) = 0$, and by 39,86% in the case $\dot{x}(1) = -\beta\nabla\phi(x_1)$. Also, the minimum values for the methods presented in Figure 2.4 can be analyzed. The last and best function values on $[1, 25]$ are displayed on Table 2.2. In all cases, the best value without restart is approximately 10^4 times larger

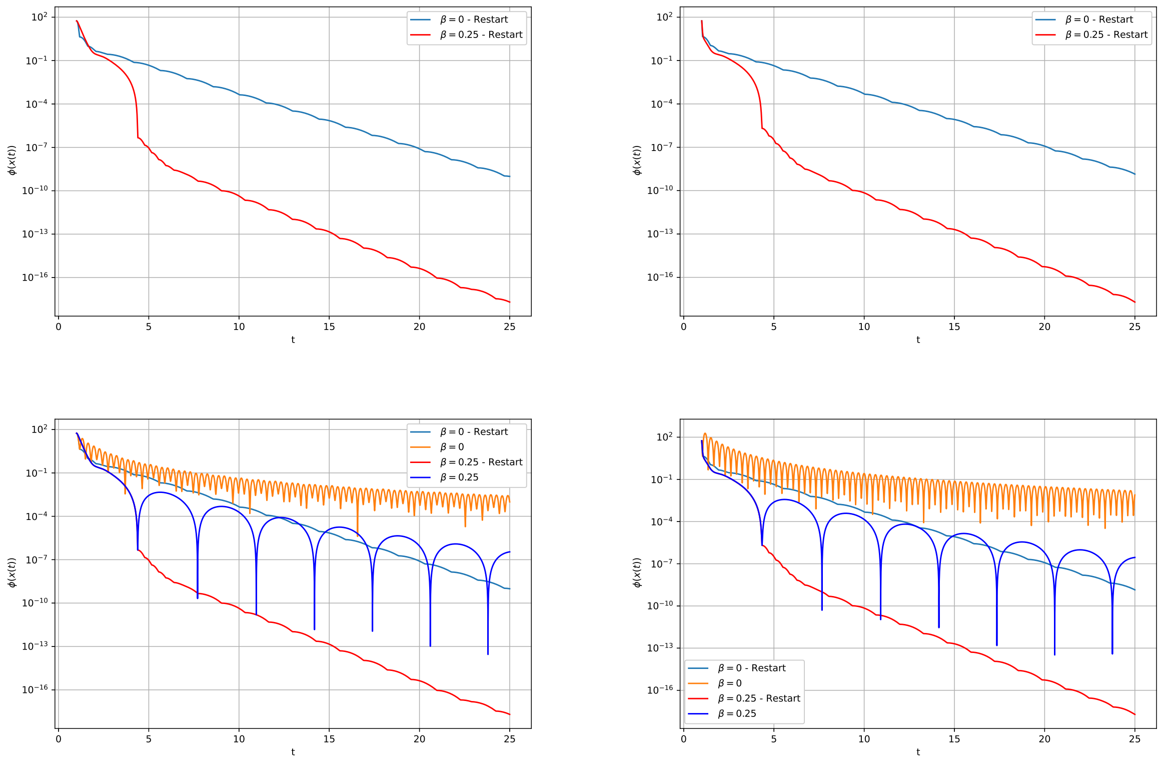


Figure 2.4: Top: Values along the trajectory, with warm start, for (AVD) (blue) and (DIN-AVD) (red), with inicial velocity set to zero (left) and $\dot{x}(1) = -\beta\nabla\phi(x_1)$ (right). Bottom: Includes trajectories without restarting, for reference.

than the one obtained with our policy. We also observe similar final values for the restarted trajectories despite the different initial velocities.

	$\dot{x}(1) = 0$		$\dot{x}(1) = -\beta\nabla\phi(x_1)$	
	$\beta = 0$	$\beta = 0.25$	$\beta = 0$	$\beta = 0.25$
A	3.7545	8.16e-6	3.2051	1.65e-05
B	0.8837	1.1901	0.859	1.2014

Table 2.1: Coefficients in the linear regression, when approximating $\phi\big(\chi(t)\big) \sim Ae^{-Bt}$.

2.6 A first exploration of the algorithmic consequences

Different discretizations of (DIN-AVD) can be used to design implementable algorithms and generate minimizing sequences for ϕ, which hopefully will share the stable behavior we observe in the solutions of (DIN-AVD). Three such algorithms were first proposed in [10], for which we implemented a speed restart scheme, analogue to the one we have used for the solutions

	$\dot{x}(1) = 0$		$\dot{x}(1) = -\beta\nabla\phi(x_1)$	
	$\beta = 0$	$\beta = 0.25$	$\beta = 0$	$\beta = 0.25$
Last value without restart	0.0009	3.4793e-07	0.0079	2.8094e-07
Best value without restart	4.0697e-06	2.8024e-14	3.2770e-05	3.2760e-14
Last (best) value with restart and warm start	9.8118e-10	2.0103e-18	1.3940e-09	1.9452e-18

Table 2.2: Values reached for ϕ at $t = 25$.

of (DIN-AVD). As the dynamics can be rewritten as a first order system, versions of the algorithm using only first order information on the gradient arises naturally as algorithms related to (DIN-AVD) dynamics. Since we obtained very similar results and the numerical analysis of algorithms is not the focus of this research, we describe only the simplest one in detail, Algorithm 1, and present the numerical results for that one. As in [83], a parameter $k_{\min}$ is introduced, to avoid two consecutive restarts to be too close. Notice also, that the restart criteria is also the same as in [83], motivated by the discretization of the continuous speed restart used before.

Algorithm 1: Inertial Gradient Algorithm with Hessian Damping (IGAHD) - Speed Restart version

```
Choose $x_0$, $x_1 \in \mathbb{R}^n$, $N$, $k_{\min}$ and $h > 0$.
for $k = 1 \dots N$ do
    Compute $y_k = x_k + (1 - \frac{\alpha}{k})(x_k - x_{k-1}) - \beta h(\nabla\phi(x_k) - \nabla\phi(x_{k-1}))$,
    and then $x_{k+1} = y_k - h^2\nabla\phi(y_k)$.
    if $\|x_{k+1} - x_k\| < \|x_k - x_{k-1}\|$ and $k \geq k_{\min}$ then
        k=1;
    else
        k=k+1.
end
return $x_N$.
```

Example 2.6.1. *We begin by applying Algorithm 1, as well as the variation with the warm start, to the function $\phi : \mathbb{R}^3 \mapsto \mathbb{R}$ in Examples 2.1.1 and 2.1.2, with the parameters $k_{\min} = 10$, $\beta = h = 1/\sqrt{L}$ and $\alpha = 3.1$. Figure 2.5 shows the evolution of the function values along the iterations. The coefficients in the approximation $\phi(x_k) \sim Ae^{-Bt}$, with $A, B > 0$, obtained for each algorithm, are detailed on Table 2.3. As one would expect, the value of B is similar and that of A is significantly lower. Also, Table 2.4 shows the values obtained along 1000 iterations. The best value without restart is 10^5 times larger than the one obtained with our policy.*

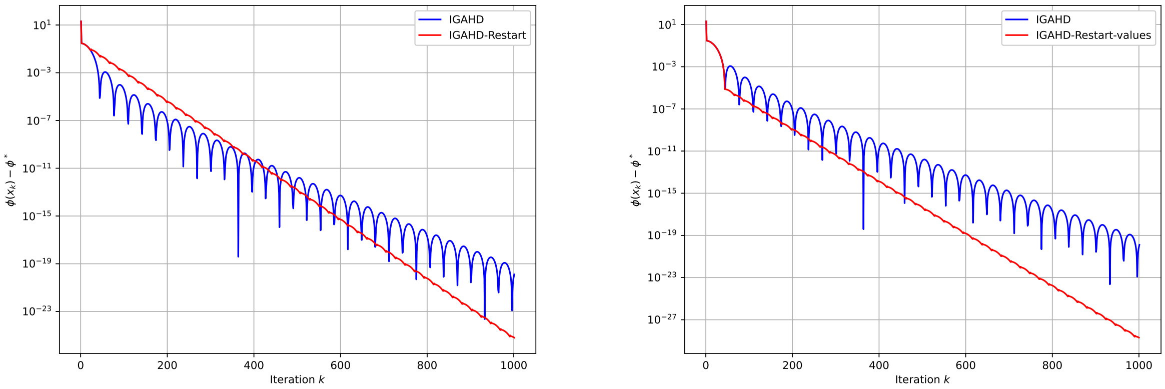


Figure 2.5: Function values along iterations of Algorithm 1 without (left) and with (right) warm start.

	Algorithm 1	Algorithm 1 with warm start
A	0.3722	1.0749e-4
B	0.0571	0.057

Table 2.3: Coefficients in the linear regression for Example 2.6.1.

Last iteration without restart	1.2927e-20
Best iteration without restart	2.2907e-24
Last/best iteration with restart and warm start	2.0206e-29

Table 2.4: Functions values for Example 2.6.1.

Example 2.6.2. *Given a positive definite symmetric matrix A of size $n \times n$, and a vector $b \in \mathbb{R}^n$, define $\phi : \mathbb{R}^n \mapsto \mathbb{R}$ by*

$$\phi(x) = \frac{1}{2} x^T A x + b^T x.$$

For the experiment, $n = 500$, A is randomly generated with eigenvalues in $(0, 1)$, and b is also chosen at random. We first compute L, and set $k_{\min} = 10$, $h = 1/\sqrt{L}$, $\alpha = 3.1$ and $\beta = h$. The initial points $x_0 = x_1$ are generated randomly as well. Figure 2.6 shows the comparison for Algorithm 1 and a variation of it giving a warm start as the one described in the continuous setting. That is, to restart the first time when the function increases instead of decrease, and then performing the speed restart detailed on Algorithm 1. It can be seen, that the restart scheme stabilizes and accelerates the convergence in both cases. The coefficients obtained for each algorithm in the approximation $\phi(x_k) \sim Ae^{-Bt}$, with $A, B > 0$, are presented in Table 2.5. Also, Table 2.6 shows the value gaps obtained along 1800 iterations. The best value without restart is more than 10^4 times larger than the one obtained with restart.

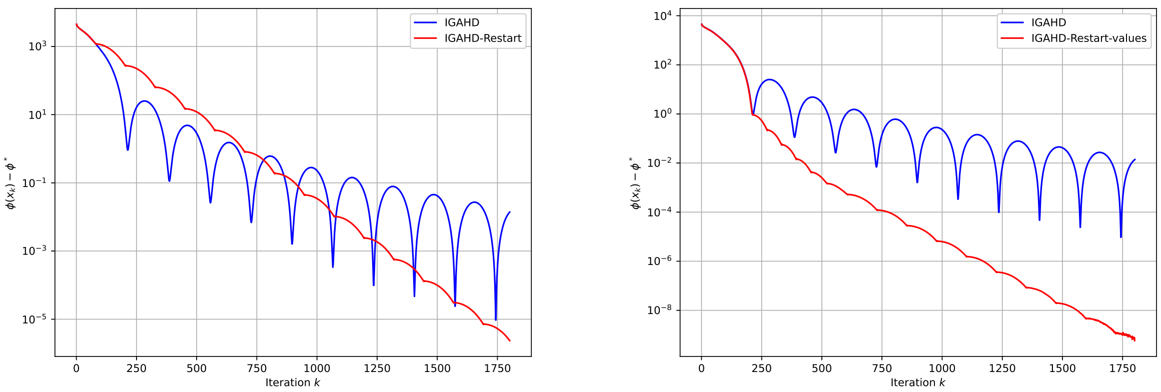


Figure 2.6: Function values along iterations of Algorithm 1 without (left) and with (right) warm start.

	Algorithm 1	Algorithm 1 with warm start
A	3813.01	1.6142
B	0.0117	0.0121

Table 2.5: Coefficients in the linear regression for Example 2.6.2.

Last iteration without restart	0.0139
Best iteration without restart	9.4293e-06
Last/best iteration with restart and warm start	5.8481e-10

Table 2.6: Function values for Example 2.6.2.

2.7 Appendix: Proof of Theorem 2.1.1

Consider the differential equation

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \mathcal{F}\big(x(t)\big)\dot{x}(t) + \mathcal{G}\big(x(t)\big) = 0. \tag{2.23}$$

We assume that γ is continuous and positive, with $\lim_{t\to 0}\gamma(t) = +\infty$, and that $\mathcal{F}$ and $\mathcal{G}$ are (continuous and) sufficiently regular so that the differential equation (2.23), with initial condition $x(\delta) = x_\delta$ and $\dot{x}(\delta) = v_\delta$, has a unique solution defined on $[\delta, T_\infty)$ for some $T_\infty \in (0, \infty]$ and all $\delta > 0$. Let

$$M(\delta, t) := \sup_{s\in[\delta,t]} \big\{\gamma(s)\, \|\dot{x}_\delta(s) - v_0\|\, \big\}. \tag{2.24}$$

We have the following:

Theorem 2.7.1. *Assume there is $T > 0$ such that*

$$\sup_{0<\delta\le t\le T} M(\delta, t) < +\infty. \tag{2.25}$$

Then, the differential equation (2.23), with initial condition $x(0) = x_0$ and $\dot{x}(0) = v_0$, has a solution.

Proof. For $\delta \in (0,T)$, define $x_\delta : [0,T] \to \mathbb{R}^n$ as follows: for $t \in [0,\delta]$, $x_\delta(t) = x_0 + tv_0$; and for $t > \delta$, x_δ is the solution of (2.23) with initial condition $x(\delta) = x_0 + \delta v_0$ and $\dot{x}(\delta) = v_0$. Notice that x_δ is a continuous function such that matches a solution of (2.23) on $[\delta, T]$. From the hypotheses, there exist $c, K > 0$ and such that $\gamma(t) \geq c$ and $M(\delta,t) \leq K$ for all $0 < \delta \leq t \leq T$. Therefore,

$$c\,\|\dot{x}_\delta(s) - v_0\| \leq \gamma(s)\|\dot{x}_\delta(s) - v_0\| \leq M(\delta,t) \leq K$$

whenever $0 < \delta \leq s \leq t \leq T$, so that

$$\|\dot{x}_\delta(s) - v_0\| \leq \frac{K}{c},$$

for all $s \in [0,T]$. As a consequence,

$$\|x_\delta(s) - x_0\| \leq \int_0^s \|\dot{x}_\delta(\tau)\|\,d\tau \leq \|v_0\|\delta + \frac{KT}{c}$$

on $[0,T]$. It follows that (x_δ) is bounded in $H^1(0,T;\mathbb{R}^n)$. By weak sequential compactness and the Rellich–Kondrachov Theorem (see, for instance [23, Theorem 9.16]), there is a sequence (δ_n) converging to zero, such that x_{δ_n} converges uniformly to a continuous function x^*, while $\dot{x}_{\delta_n}$ converges weakly in $L^2(0,T;\mathbb{R}^n)$ to some y^*.

Clearly, $x^*(0) = x_0$. In turn, for $t \in (0,T]$, by the Mean Value Theorem and the definition of M, we have

$$\begin{aligned}\left\|\frac{x^*(t) - x_0}{t} - v_0\right\| &= \lim_{n\to\infty}\left\|\frac{x_{\delta_n}(t) - x_0}{t} - v_0\right\| \\ &= \lim_{n\to\infty}\|\dot{x}_{\delta_n}(c_n) - v_0\| \\ &\leq \frac{\bar{K}}{\min\limits_{s\in(0,t]}\gamma(s)},\end{aligned}$$

which tends to zero as $t \to 0$. It remains to prove that x^* satisfies (2.23). To this end, take any $t_0 \in (0,T)$, and observe that $\delta_n < t_0$ for all sufficiently large n. Therefore, x_{δ_n} satisfies (2.23) on $[t_0,T)$ for all such n. Multiplying by

$$\Gamma(t) := \exp\left(\int_{t_0}^t \gamma(s)\,ds\right),$$

we deduce that

$$\Gamma(t)\dot{x}_{\delta_n}(t) - \Gamma(t_0)\dot{x}_{\delta_n}(t_0) + \int_{t_0}^t \Gamma(s)\,\mathcal{F}\big(x_{\delta_n}(s)\big)\dot{x}_{\delta_n}(s)\,ds + \int_{t_0}^t \Gamma(s)\,\mathcal{G}\big(x_{\delta_n}(s)\big)\,ds = 0.$$

By taking yet another subsequence if necessary, we may assume that $\dot{x}_{\delta_n}(t_0)$ converges to some v^*. From the uniform convergence of x_{δ_n} to x^* on $[0,T]$, and the weak convergence of $\dot{x}_{\delta_n}$ to y^* in $L^2(0,T;\mathbb{R}^n)$, it ensues that

$$\Gamma(t)y^*(t) - \Gamma(t_0)\dot{v}^* + \int_{t_0}^t \Gamma(s)\,\mathcal{F}\big(x^*(s)\big)y^*(s)\,ds + \int_{t_0}^t \Gamma(s)\,\mathcal{G}\big(x^*(s)\big)\,ds = 0$$

for all $t \in (t_0,T)$. As a consequence, x^* is continuously differentiable, $\dot{x}^* = y$, and x^* satisfies (2.23). □

Corollary 2.7.1. *Equation* (DIN-AVD) *has at least one solution.*

Proof. According to Theorem 2.7.1, for the existence, it suffices to show that the expression $M(\delta, t)$, defined in (2.24), is bounded for $0 < \delta \leq t \leq T$, for some $T > 0$. Mimicking the proof of Lemma 2.3.2, we show that

$$H(t)M(\delta, t) \leq \frac{\|\nabla\phi(x_0)\|}{\alpha+1}, \qquad \text{with} \qquad H(t) = 1 - \frac{\beta L t}{\alpha+2} - \frac{Lt^2}{2(\alpha+3)}.$$

The only positive zero of H is τ_1, given by (2.13), and H is decreasing on $(0, \tau_1)$. Hence, if $T < \tau_1$, then

$$\sup_{0<\delta\leq t\leq T} M(\delta, t) \leq \frac{\|\nabla\phi(x_0)\|}{(\alpha+1)H(T)} < +\infty,$$

as claimed. □

Proposition 2.7.1. *Equation* (DIN-AVD)*, with initial condition* $x(0) = x_0$ *and* $\dot{x}(0) = 0$*, has at most one solution in a neighborhood of* $t = 0$*.*

Proof. Let x and y satisfy (DIN-AVD) with the same initial state and null initial velocity. We define

$$\tilde{M}(t) = \sup_{u\in[0,t)} \{\|\dot{x}(u) - \dot{y}(u)\|\},$$

and proceed as in the proof of Lemma 2.3.1, to obtain

$$\|\nabla\phi(x(t)) - \nabla\phi(y(t))\| \leq Lt\tilde{M}(t) \tag{2.26}$$

As x and y satisfy (DIN-AVD), we integrate by parts to obtain

$$\begin{aligned}
t^\alpha(\dot{x}(t) - \dot{y}(t)) &= -\int_0^t u^\alpha \left(\nabla\phi(x(u)) - \nabla\phi(y(u))\right) du \\
&\quad - \beta\int_0^t u^\alpha \left(\nabla^2\phi(x(u))\dot{x}(u) - \nabla^2\phi(y(u))\dot{y}(u)\right) du \\
&= -\int_0^t u^\alpha \left(\nabla\phi(x(u)) - \nabla\phi(y(u))\right) du \\
&\quad - \beta\int_0^t u^\alpha \frac{d}{du}\left(\nabla\phi(x(u)) - \nabla\phi(y(u))\right) du \\
&= -\int_0^t u^\alpha \left(\nabla\phi(x(u)) - \nabla\phi(y(u))\right) du \\
&\quad - \beta t^\alpha \left(\nabla\phi(x(t)) - \nabla\phi(y(t))\right) \\
&\quad + \alpha\int_0^t u^{\alpha-1} \left(\nabla\phi(x(u)) - \nabla\phi(y(u))\right) du.
\end{aligned}$$

Using (2.26), and the fact that $\tilde{M}(t)$ is increasing, we get

$$t^\alpha \|\dot{x}(t) - \dot{y}(t)\| \leq \int_0^t Lu^{\alpha+1}\tilde{M}(u)\, du + \beta L t^{\alpha+1}\tilde{M}(t) + \alpha\beta\int_0^t Lu^\alpha \tilde{M}(u)\, du$$

$$\leq \frac{1}{\alpha+2} L\tilde{M}(t)t^{\alpha+2} + \frac{2\alpha+1}{\alpha+1}\beta L\tilde{M}(t)t^{\alpha+1}.$$

Then,

$$\|\dot{x}(t) - \dot{y}(t)\| \leq \frac{1}{\alpha+2} L\tilde{M}(T)T^2 + \frac{2\alpha+1}{\alpha+1}\beta L\tilde{M}(T)T,$$

whenever $0 < t \leq T$. Taking supremum, we conclude that

$$Q(t)\tilde{M}(T) \leq 0 \qquad \text{with} \qquad Q(t) = 1 - \frac{2\alpha+1}{\alpha+1}\beta Lt - \frac{1}{\alpha+2}Lt^2,$$

for all $T > 0$. Since $Q(T) > 0$ in a neighborhood of 0, it follows that $\tilde{M}$ must vanish there, whence x and y must coincide. □

2.8 Appendix: A bound for $\Psi(\tau_3)$

Since $H(t) > 1/2$ and decreasing on $(0, \tau_2)$, $2 - H(t)^{-1}$ is positive and decreasing there. Whence Ψ is decreasing, and $\Psi(t) \geq \Psi(\tau)$, whenever $0 \leq t \leq \tau < \tau_2$. In particular, $\Psi(t) \geq \Psi(\tau_3)$ for every $t \in [0, \tau_3]$. It remains to compute (or find a lower bound for) $\Psi(\tau_3)$; or yet equivalently for $H(\tau_3)$. Consider

$$\Psi(t) = \left(2 - \frac{1}{H(t)}\right)^2,$$

and τ_3 as the positive root of $G(t)$ given by (2.18). Let $z = L\beta^2$ and consider the function

$$P(z) = 2(-\sqrt{z}\sqrt{K_1 z + K_2} + K_3 z + K_4),$$

with

$$\begin{aligned} K_1 &= (\alpha+3)^2(2\alpha+3)^2 \\ K_2 &= 4(\alpha+2)^3(\alpha+3) \\ K_3 &= (\alpha+3)(2\alpha+3) \\ K_4 &= 2(\alpha+1)(\alpha+2)^3. \end{aligned}$$

Then, $\Psi(\tau_3)$ can be written as

$$\Psi(\tau_3) = \Delta(z) = 4\left(\frac{P(z)}{P(z) + 2K_5}\right)^2,$$

with $K_5 = 2(\alpha+2)^4$. The aim is to find a lower bound for $\Psi(\tau_3)$. Then, we will split the proof in three steps.

$P(z)$ is positive.

Notice that we are only interested in the case where $z > 0$. Let us proceed by contradiction, assume that there is a $z > 0$ such that $P(z)$ is negative. That is,

$$K_3 z + K_4 < \sqrt{z}\sqrt{K_1 z + K_2}.$$

As both sides are positive, we can take the square and noticing that $K_3^2 = K_1$ yields

$$z(2K_3K_4 - K_2) + K_4^2 < 0.$$

That is,

$$4z(\alpha+2)^3(\alpha+3)\left[(2\alpha+3)(\alpha+1)-1\right] + 4(\alpha+1)^2(\alpha+2)^6 < 0,$$

which is clearly a contradiction.

Derivative $P'(z)$.

Computing the derivative, it yields

$$P'(z) = \frac{2K_3\sqrt{z}\sqrt{K_1z+K_2} - 2K_1z - K_2}{\sqrt{z}\sqrt{K_1z+K_2}}.$$

Notice that the sign of the derivative is given by the numerator. Proceeding as in the previous step, we get that the $P'(z)$ is negative for every $z > 0$.

Limit of $P(z)$.

$$\begin{aligned}
\lim_{z\to\infty} P(z) &= 2K_4 + 2\lim_{z\to\infty} -\sqrt{z}\sqrt{K_1z+K_2} + K_3z \\
&= 2K_4 + 2\lim_{z\to\infty} \frac{K_3^2z^2 - z(K_1z+K_2)}{\sqrt{z}\sqrt{K_1z+K_2} + K_3z} \\
&= 2K_4 - 2\lim_{z\to\infty} \frac{zK_2}{\sqrt{z}\sqrt{K_1z+K_2} + K_3z} \\
&= 2K_4 - 2\lim_{z\to\infty} \frac{K_2}{\sqrt{K_1 + \dfrac{K_2}{z}} + K_3} \\
&= 2K_4 - \frac{K_2}{K_3} \\
&= 4(\alpha+2)^4\frac{(2\alpha+1)}{(2\alpha+3)}
\end{aligned}$$

Using the previous results, we can perform an analysis on the function $\Delta(z)$. Notice that, the derivative is given by

$$\Delta'(z) = 8\left(\frac{P(z)}{P(z)+2K_5}\right)\frac{2K_5P'(z)}{(P(z)+2K_5)^2},$$

and as $P(z)$ is positive and $P'(z)$ is negative, we get that $\Delta'(z) < 0$ and then, $\Delta(z)$ is a positive and decreasing function. Then, for every value of $z > 0$, $\Delta(z)$ is greater than the limit towards ∞. Computing the limit,

$$\lim_{z\to\infty}\Delta(z) = \lim_{z\to\infty} 4\left(\frac{P(z)}{P(z)+2K_5}\right)^2$$

$$= \lim_{z\to\infty} 4 \left(\frac{4(\alpha+2)^4 \dfrac{(2\alpha+1)}{(2\alpha+3)}}{4(\alpha+2)^4 \dfrac{(2\alpha+1)}{(2\alpha+3)} + 4(\alpha+2)^4} \right)^2$$

$$= \frac{1}{4} \frac{(2\alpha+1)^2}{(\alpha+1)^2}$$

Then, we have proved that

$$\Psi(\tau_3) > \frac{1}{4} \frac{(2\alpha+1)^2}{(\alpha+1)^2},$$

for every value of L and β. In the case where $\alpha = 3$, the bound gives $\left(\frac{7}{8}\right)^2$.

Chapter 3

Inertial Krasnoselskii-Mann iterations

3.1 Introduction

Krasnoselskii-Mann (KM) iterations are at the core of numerical methods used in optimization, fixed point theory and variational analysis, since they include many fundamental splitting algorithms whose convergence can be analyzed in a unified manner. These include the forward-backward algorithm [57, 73] to approximate a zero of the sum of two maximally monotone operators, and its various particular instances: on the one hand, we have the gradient projection algorithm [48, 53], the gradient method [27] and the proximal point algorithm [61, 79, 24, 49], to cite some abstract methods, as well as the Iterative Shrinkage-Thresholding Algorithm (ISTA) [36, 33], to speak more concretely. KM iterations also encompass other splitting methods like Douglas-Rachford [42], primal-dual methods [30, 8, 31, 88, 34] and the three-operator splitting [37].

As it was showed by Nesterov in [67], the inclusion of inertia on the gradient method improves the convergence, so the main goal of this part is to study the effect of inertia in an more general operator setting. To our knowledge, the first extensions beyond the optimization setting was developed in [5], followed by [59, 58] some years later. The main drawback of their analysis is that they require an implicit hypothesis on the sequence generated by the algorithm (the summability of a certain series) to ensure its convergence. In [5], however, this difficulty is overcome, in some special cases and for small values of the inertial parameters. These ideas were also used in [22], and then improved in [41], by adapting the inertial factors to the relaxation ones (see below). A similar principle had been used in [9], whose analysis was based on [14]. Nonasymptotic convergence rates for the residuals have been given in [81, 51]. Strong and linear convergence can be found in [82], for strictly contractive *forward-projection* operators. Other extensions have been considered in [40, 32, 65, 39]. See also [38] for a more thorough account of KM iterations, with and without inertia. Interest in this type of methods increased remarkably in the past decade in view of theoretical advances in the convergence theory for the *Fast Iterative Shrinkage-Thresholding Algorithm* (FISTA) [20], obtained in [29, 12, 13].

In this chapter inertial Krasnoselskii-Mann iterations will be studied in their general form

$$\left\{\begin{array}{rcl} y_k &=& x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} &=& (1-\lambda_k)y_k + \lambda_k T_k(y_k), \end{array}\right. \tag{3.1}$$

where (T_k) is a family of operators defined on a real Hilbert space $\mathcal{H}$, and the positive sequences (α_k) and (λ_k) are the *inertial* and *relaxation* (or *averaging*) parameters, respectively.

The aim of this chapter is to provide convergence results for the iterations (3.1), in the weak and strong sense, depending on the assumptions over the family of operators T_k. Numerical illustrations are provided where the effect of inertia is observed by being implemented over two existing fixed-point algorithms.

3.2 Vanishing residuals and weak convergence

An operator $T : \mathcal{H} \to \mathcal{H}$ is *quasi-nonexpansive* if $\operatorname{Fix}(T) \neq \emptyset$ and $\|Ty - p\| \leq \|y - p\|$ for all $y \in \mathcal{H}$ and $p \in \operatorname{Fix}(T)$. This implies, in particular, that

$$2\langle y - p, Ty - y\rangle \leq -\|Ty - y\|^2 \tag{3.2}$$

for all $y \in \mathcal{H}$ and $p \in \operatorname{Fix}(T)$.

In this section, we consider a family (T_k) of quasi-nonexpansive operators on $\mathcal{H}$, with $F := \bigcap_{k\geq 1} \operatorname{Fix}(T_k) \neq \emptyset$, along with a sequence (x_k, y_k) satisfying (3.1), where (α_k) is a nondecreasing sequence[1] in $[0, 1)$, and (λ_k) is a sequence in $(0, 1)$ such that $\inf_{k\geq 1} \lambda_k > 0$.

To simplify the notation, given $p \in F$, we set

$$\begin{cases} \nu_k &= \left(\lambda_k^{-1} - 1\right) \\ \delta_k &= \nu_{k-1}(1 - \alpha_{k-1})\|x_k - x_{k-1}\|^2, \\ \Delta_k(p) &= \|x_k - p\|^2 - \|x_{k-1} - p\|^2, \quad \Delta_1(p) = 0 \\ C_k(p) &= \|x_k - p\|^2 - \alpha_{k-1}\|x_{k-1} - p\|^2 + \delta_k, \quad C_1(p) = \|x_1 - p\|^2, \\ \omega_k &= \|x_k - 2x_{k-1} + x_{k-2}\|^2. \end{cases} \tag{3.3}$$

At different points, and in order to simplify the computations, we shall make use of a basic property of the norm in $\mathcal{H}$: for every $x, y \in \mathcal{H}$ and $\alpha \in [0, 1]$, we have

$$\|\alpha x + (1 - \alpha)y\|^2 = \alpha\,\|x\|^2 + (1 - \alpha)\,\|y\|^2 - \alpha(1 - \alpha)\,\|x - y\|^2\,. \tag{3.4}$$

The following auxiliary result will be useful in the sequel:

Lemma 3.2.1. *Let (T_k) be a family of quasi-nonexpansive operators on $\mathcal{H}$, with $F := \bigcap_{k\geq 1} \operatorname{Fix}(T_k) \neq \emptyset$, and let (x_k, y_k) satisfy (3.1). For each $k \geq 1$ and $p \in F$, we have*

$$\Delta_{k+1}(p) + \delta_{k+1} + \nu_k\alpha_k\omega_{k+1} \leq \alpha_k\Delta_k(p) + \big[\alpha_k(1 + \alpha_k) + \nu_k\alpha_k(1 - \alpha_k)\big]\|x_k - x_{k-1}\|^2. \tag{3.5}$$

Proof. Take $p \in F$. From (3.1), it follows that

$$\begin{aligned} \|x_{k+1} - p\|^2 &= \|y_k - p\|^2 + \lambda_k^2\|y_k - T_ky_k\|^2 + 2\lambda_k\langle y_k - p, T_ky_k - y_k\rangle \\ &\leq \|y_k - p\|^2 - \lambda_k(1 - \lambda_k)\|y_k - T_ky_k\|^2, \end{aligned} \tag{3.6}$$

[1] This is just to simplify the proof and is sufficiently general for practical purposes.

where the inequality is given by (3.2). From

$$\begin{aligned}\|y_k - p\|^2 &= \|x_k - p + \alpha_k(x_k - x_{k-1})\|^2 \\ &= \|x_k - p\|^2 + \alpha_k^2\|x_k - x_{k-1}\|^2 + 2\alpha_k\langle x_k - p, x_k - x_{k-1}\rangle\end{aligned}$$

and

$$2\alpha_k\langle x_k - p, x_k - x_{k-1}\rangle = \alpha_k\|x_k - p\|^2 + \alpha_k\|x_k - x_{k-1}\|^2 - \alpha_k\|x_{k-1} - p\|^2,$$

we deduce that

$$\|y_k - p\|^2 = (1+\alpha_k)\|x_k - p\|^2 + \alpha_k(1+\alpha_k)\|x_k - x_{k-1}\|^2 - \alpha_k\|x_{k-1} - p\|^2. \tag{3.7}$$

By combining expressions (3.6) and (3.7), we obtain

$$\begin{aligned}\|x_{k+1} - p\|^2 \le (1+\alpha_k)\|x_k - p\|^2 + \alpha_k(1+\alpha_k)\|x_k - x_{k-1}\|^2 - \alpha_k\|x_{k-1} - p\|^2 \\ - \lambda_k(1-\lambda_k)\|y_k - T_k y_k\|^2.\end{aligned}$$

Recalling from (3.3) that $\Delta_k(p) = \|x_k - p\|^2 - \|x_{k-1} - p\|^2$, we rewrite the latter as

$$\Delta_{k+1}(p) \le \alpha_k\Delta_k(p) + \alpha_k(1+\alpha_k)\|x_k - x_{k-1}\|^2 - \lambda_k(1-\lambda_k)\|y_k - T_k y_k\|^2. \tag{3.8}$$

In turn,

$$\lambda_k^2\|y_k - T_k y_k\|^2 = \|x_{k+1} - x_k\|^2 + \alpha_k^2\|x_k - x_{k-1}\|^2 - 2\alpha_k\langle x_{k+1} - x_k, x_k - x_{k-1}\rangle, \tag{3.9}$$

and

$$-2\alpha_k\langle x_{k+1} - x_k, x_k - x_{k-1}\rangle = \alpha_k\omega_{k+1} - \alpha_k\|x_{k+1} - x_k\|^2 - \alpha_k\|x_k - x_{k-1}\|^2,$$

together give

$$\lambda_k^2\|y_k - T_k y_k\|^2 = (1-\alpha_k)\|x_{k+1} - x_k\|^2 - \alpha_k(1-\alpha_k)\|x_k - x_{k-1}\|^2 + \alpha_k\omega_{k+1}. \tag{3.10}$$

By multiplying the latter by $\nu_k = (1-\lambda_k)/\lambda_k$, and using the definition of δ_k in (3.3), we rewrite this as

$$\delta_{k+1} + \nu_k\alpha_k\omega_{k+1} = \nu_k\alpha_k(1-\alpha_k)\|x_k - x_{k-1}\|^2 + \lambda_k(1-\lambda_k)\|y_k - T_k y_k\|^2. \tag{3.11}$$

Summing (3.8) and (3.11), we obtain (3.5). □

We are now in a position to show that the sequence (x_n) remains anchored to the set F, while both the residuals $\|y_k - T_k y_k\|$ and the speed $\|x_k - x_{k-1}\|$ tend to 0. We shall make some assumptions on the parameter sequences (α_k) and (λ_k).

Hypothesis A. *There is k_0 such that*

$$\alpha_k(1+\alpha_k) + (\lambda_k^{-1} - 1)\alpha_k(1-\alpha_k) - (\lambda_{k-1}^{-1} - 1)(1-\alpha_{k-1}) \le 0,$$

for all $k \ge k_0$.

A reinforced version with strict inequality is given by:

Hypothesis B.

$$\limsup_{k\to\infty}\left[\alpha_k(1+\alpha_k)+(\lambda_k^{-1}-1)\alpha_k(1-\alpha_k)-(\lambda_{k-1}^{-1}-1)(1-\alpha_{k-1})\right]<0.$$

Remark 3.2.1. *With Hypothesis A or B, there exist $\varepsilon\geq 0$ and $k_0\geq 1$ such that*

$$\alpha_k(1+\alpha_k)+(\lambda_k^{-1}-1)\alpha_k(1-\alpha_k)\leq(\lambda_{k-1}^{-1}-1)(1-\alpha_{k-1})-\varepsilon \tag{3.12}$$

for all $k\geq k_0$ (if Hypothesis B holds, then $\varepsilon>0$; otherwise, $\varepsilon=0$). Also, under Hypothesis B, $\alpha:=\sup_{k\geq 1}\alpha_k<1$ and $\lambda:=\inf_{k\geq 1}\lambda_k>0$.

Theorem 3.2.1. *Let (T_k) be a family of quasi-nonexpansive operators on $\mathcal{H}$, and let (x_k,y_k) satisfy (3.1). Take $p\in F=\bigcap_{k\geq 1}\mathrm{Fix}(T_k)$.*

i) Assume that there is $k_0\geq 1$ such that

$$\alpha_k(1+\alpha_k)+(\lambda_k^{-1}-1)\alpha_k(1-\alpha_k)-(\lambda_{k-1}^{-1}-1)(1-\alpha_{k-1})\leq 0 \tag{3.13}$$

for all $k\geq k_0$. Then, the sequence $\big(C_k(p)\big)_{k\geq k_0}$ is nonincreasing and nonnegative, thus $\lim\limits_{k\to\infty}C_k(p)$ exists.

ii) If Hypothesis B holds, the series $\sum\limits_{k\geq 2}\omega_k$, $\sum\limits_{k\geq 1}\|x_k-x_{k-1}\|^2$, $\sum\limits_{k\geq 1}\delta_k$ and $\sum\limits_{k\geq 1}\|y_k-T_ky_k\|^2$ are convergent, and there is a constant $M>0$, depending only on (α_k) and (λ_k), such that

$$\min_{1\leq k\leq n}\|y_k-T_ky_k\|^2\leq\frac{M\,\mathrm{dist}(x_1,F)^2}{n}. \tag{3.14}$$

Moreover, for each $p\in F$, $\lim\limits_{k\to\infty}\|x_k-p\|$ exists.

Proof. Without any loss of generality, we may assume that (3.12) holds with $k_0=1$. Take any $p\in F$, and combine (3.12) with (3.5), to obtain

$$\begin{aligned}\Delta_{k+1}(p)+\delta_{k+1}+\nu_k\alpha_k\omega_k &\leq \alpha_k\Delta_k(p)+\left[\nu_{k-1}(1-\alpha_{k-1})-\varepsilon\right]\|x_k-x_{k-1}\|^2\\ &= \alpha_k\Delta_k(p)+\delta_k-\varepsilon\|x_k-x_{k-1}\|^2.\end{aligned} \tag{3.15}$$

On the one hand, (3.15) immediately gives

$$\Delta_{k+1}(p)\leq\alpha_k\Delta_k(p)+\delta_k. \tag{3.16}$$

On the other, since (α_k) is nondecreasing, we have

$$\begin{aligned}C_{k+1}(p)-C_k(p)&=\Delta_{k+1}(p)-\big(\alpha_k\|x_k-p\|^2-\alpha_{k-1}\|x_{k-1}-p\|^2\big)+\delta_{k+1}-\delta_k\\ &\leq\Delta_{k+1}(p)+\delta_{k+1}-\alpha_k\Delta_k(p)-\delta_k.\end{aligned}$$

Therefore, (3.15) implies

$$C_{k+1}(p)+\nu_k\alpha_k\|x_{k+1}-2x_k+x_{k-1}\|^2+\varepsilon\|x_k-x_{k-1}\|^2\leq C_k(p). \tag{3.17}$$

It ensues that $\big(C_k(p)\big)$ is nonincreasing. To show that it is nonnegative, suppose that $C_{k_1}(p)<0$ for some $k_1\geq 1$. Since $\big(C_k(p)\big)$ is nonincreasing,

$$\|x_k-p\|^2-\alpha_{k-1}\|x_{k-1}-p\|^2\leq C_k(p)\leq C_{k_1}(p)<0$$

for all $k \geq k_1$. If follows that $\|x_k - p\|^2 \leq \|x_{k-1} - p\|^2 + C_{k_1}(p)$, and so

$$0 \leq \|x_k - p\|^2 \leq \|x_{k-1} - p\|^2 + C_{k_1}(p) \leq \cdots \leq \|x_{k_1} - p\|^2 + (k - k_1)C_{k_1}(p)$$

for all $k \geq k_1$, which is impossible. As a consequence $\big(C_k(p)\big)$ is nonnegative, and $\lim_{k\to\infty} C_k(p)$ exists.

For ii), inequality (3.12) holds with $\varepsilon > 0$. The summability of the first two series follows from (3.17). In particular,

$$\varepsilon \sum_{k\geq 1} \|x_k - x_{k-1}\|^2 \leq C_1(p) = \|x_1 - p\|^2. \tag{3.18}$$

The third one is a consequence of the second one, since $\lambda := \inf_{k\geq 1} \lambda_k > 0$. For the last one, use (3.10) to write

$$\lambda_k^2 \|y_k - T_k y_k\|^2 \leq (1+\alpha)\|x_{k+1} - x_k\|^2 + \alpha(1+\alpha)\|x_k - x_{k-1}\|^2.$$

In view of (3.18), this gives the summability of the fourth series, with

$$n \min_{1\leq k\leq n} \|y_k - T_k y_k\|^2 \leq \sum_{k\geq 1} \|y_k - T_k y_k\|^2 \leq \frac{(1+\alpha)^2}{\varepsilon\lambda^2}\|x_1 - p\|^2.$$

Since this holds for each $p \in F$, we obtain (3.14) with $M = \frac{(1+\alpha)^2}{\varepsilon\lambda^2}$. Now, denoting the positive part of $d \in \mathbb{R}$ by $[d]_+$, we obtain from (3.16) that

$$(1-\alpha)\big[\Delta_{k+1}(p)\big]_+ + \alpha\big[\Delta_{k+1}(p)\big]_+ \leq \alpha\big[\Delta_k(p)\big]_+ + \delta_k.$$

Summing for $k \geq 1$, we obtain

$$(1-\alpha)\sum_{k\geq 1} \big[\Delta_{k+1}(p)\big]_+ \leq \alpha\big[\Delta_1(p)\big]_+ + \sum_{k\geq 1}\delta_k = \sum_{k\geq 1}\delta_k < \infty.$$

By writing $h_k = \|x_k - p\|^2 - \sum_{j=1}^k \big[\Delta_j(p)\big]_+$, we get $h_{k+1} - h_k = \Delta_{k+1}(p) - \big[\Delta_{k+1}(p)\big]_+ \leq 0$, from which we conclude that $\lim_{k\to\infty} \|x_k - p\| = \lim_{k\to\infty} h_k$ exists.

□

Remark 3.2.2. *Hypotheses A and B are closely related, but different, from the hypotheses used in [9] for forward-backward iterations. In the non-inertial case* $\alpha = 0$*, Hypothesis A is just* $\limsup_{k\to\infty} \lambda_k < 1$*. On the other hand, since* (α_k) *is nondecreasing and bounded, we have* $\alpha_k \to \alpha \in [0,1]$*. If* $\lambda_k \to \lambda$*, then Hypothesis B is reduced to*

$$\lambda(1 - \alpha + 2\alpha^2) < (1-\alpha)^2. \tag{3.19}$$

For each $\alpha \in [0,1)$*, there is* $\lambda_\alpha > 0$ *such that* (3.19) *holds for all* $\lambda < \lambda_\alpha$*.*

In order to prove the weak convergence of the sequences generated by Algorithm (3.1), we shall use the following nonautonomous extension of the concept of demiclosedness.

The family of operators $(I - T_k)$ is *asymptotically demiclosed at* 0 if for every sequence (z_k) such that $z_k \rightharpoonup z$ and $z_k - T_k z_k \to 0$, we must have $z \in F = \bigcap_{k\geq 1} \mathrm{Fix}(T_k)$.

Of course, if $T : \mathcal{H} \to \mathcal{H}$ is nonexpansive and $T_k \equiv T$, then $I - T_k$ is asymptotically demiclosed at 0. We shall discuss other examples in the next section.

Theorem 3.2.2. *Let (T_k) be a family of quasi-nonexpansive operators on $\mathcal{H}$, with $F = \bigcap_{k\geq 1}\operatorname{Fix}(T_k) \neq \emptyset$. Let (x_k, y_k) satisfy (3.1), and assume Hypothesis B holds. If $(I - T_k)$ is asymptotically demiclosed at 0, then both x_k and y_k converge weakly, as $k \to \infty$, to a point in F.*

Proof. Recall that $\lim\limits_{k\to\infty} \|y_k - T_k y_k\| = \lim\limits_{k\to\infty} \|x_k - x_{k-1}\| = 0$, by part ii) of Theorem 3.2.1. From (3.1), we deduce that (y_k) and (x_k) have the same (weak and strong) limit points. Suppose $x_{n_k} \rightharpoonup x$. Then, $y_{n_k} \rightharpoonup x$ as well. Since $y_{n_k} - T_k y_{n_k} \to 0$, the asymptotic demiclosedness implies $x \in F$. Opial's Lemma [72] (see, for instance, [74, Lemma 5.2]) yields the conclusion. □

3.3 Strong and linear convergence

We now focus on the strong convergence of the sequences generated by (3.1), and their convergence rate. As before, we assume that (α_k) is nondecreasing but we do not assume, in principle, that $\inf_{k\geq 1} \lambda_k > 0$.

Given $q \in (0,1)$, an operator $T : \mathcal{H} \to \mathcal{H}$ is *q-quasi-contractive* if $\operatorname{Fix}(T) \neq \emptyset$ and $\|Ty - p\| \leq q\|y - p\|$ for all $y \in \mathcal{H}$ and $p \in \operatorname{Fix}(T)$. If T is q-quasi-contractive, then $\operatorname{Fix}(T) = \{p^*\}$. Given $\lambda, q \in (0,1)$ and $\xi \in [0,1]$, we define

$$Q(\lambda, q, \xi) := \xi\big(1 - \lambda + \lambda q^2\big) + (1 - \xi)(1 - \lambda + \lambda q)^2 = (1 - \lambda + \lambda q)^2 + \xi\lambda(1 - \lambda)(1 - q)^2. \tag{3.20}$$

Notice that $Q(\lambda, q, \xi) \in (0,1)$, and that it decreases as λ increases, or as either q or ξ decreases. The quantity $Q(\lambda, q, \xi)$ will play a crucial role in the linear convergence rate of the sequences satisfying (3.1). The inclusion of the auxiliary parameter ξ will also allow us to establish convergence rates, with and without inertia, in a unified manner (see the discussion in Subsection 3.3.3).

The following result establishes a bound on the distance to a solution after performing a standard KM step:

Lemma 3.3.1. *Let $T : \mathcal{H} \to \mathcal{H}$ be q-quasi-contractive with fixed point p^*, and let $x, y \in \mathcal{H}$ and $\lambda > 0$ be such that $x = (1 - \lambda)y + \lambda Ty$. Then, for each $\xi \in [0,1]$, we have*

$$\|x - p^*\|^2 \leq Q(\lambda, q, \xi)\|y - p^*\|^2 - \xi\lambda(1 - \lambda)\|Ty - y\|^2. \tag{3.21}$$

Proof. Notice that

$$\|x - p^*\| = \|(1 - \lambda)(y - p^*) + \lambda(Ty - p^*)\|.$$

Then, using (3.4), we get

$$\begin{aligned} \|x - p^*\|^2 &= (1 - \lambda)\|y - p^*\|^2 + \lambda\|Ty - p^*\|^2 - \lambda(1 - \lambda)\|Ty - y\|^2 \\ &\leq \big(1 - \lambda + \lambda q^2\big)\|y - p^*\|^2 - \lambda(1 - \lambda)\|Ty - y\|^2. \end{aligned} \tag{3.22}$$

On the other hand, we have

$$\|x - p^*\| \leq (1 - \lambda)\|y - p^*\| + \lambda\|Ty - p^*\| \leq (1 - \lambda + \lambda q)\|y - p^*\|. \tag{3.23}$$

Then, inequality (3.21) is just a convex combination of (3.22) and the square of (3.23). □

3.3.1 Convergence analysis

We now turn to the convergence of the sequences verifying (3.1). To simplify the notation, for each $k \in \mathbb{N}$, we set

$$\tilde{C}_k(p) = \|x_k - p^*\|^2 - \alpha_{k-1}\|x_{k-1} - p^*\|^2 + \xi\delta_k \quad \text{with} \quad \tilde{C}_1(p^*) = \|x_1 - p^*\|^2.$$

We have the following:

Proposition 3.3.1. *Let (T_k) be a sequence of operators on $\mathcal{H}$, such that* $\operatorname{Fix}(T_k) \equiv \{p^*\}$ *and T_k is q_k-quasi-contractive for each $k \in \mathbb{N}$. Let (x_k, y_k) satisfy* (3.1)*, and let $\xi \in [0,1]$. Write $Q_k = Q(\lambda_k, q_k, \xi)$, where Q is defined in* (3.20)*. For each $k \in \mathbb{N}$, we have*

$$\begin{aligned}\|x_{k+1} - p^*\|^2 + \xi\delta_{k+1} \leq Q_k & \left[(1+\alpha_k)\|x_k - p^*\|^2 - \alpha_k\|x_{k-1} - p^*\|^2\right] \\ & + \left[Q_k\alpha_k(1+\alpha_k) + \xi\nu_k\alpha_k(1-\alpha_k)\right]\|x_k - x_{k-1}\|^2. \end{aligned} \tag{3.24}$$

If, moreover,

$$Q_k\alpha_k(1+\alpha_k) + \xi\nu_k\alpha_k(1-\alpha_k) - \xi Q_k\nu_{k-1}(1-\alpha_{k-1}) \leq 0 \tag{3.25}$$

for all $k \in \mathbb{N}$, then

$$\tilde{C}_{k+1}(p^*) \leq \left[\prod_{j=1}^{k} Q_j\right] \|x_1 - p^*\|^2 \tag{3.26}$$

and

$$\|x_{k+1} - p^*\|^2 \leq \left[\alpha^k + \sum_{j=1}^{k} \alpha^{k-j}\left[\prod_{i=1}^{j} Q_i\right]\right] \|x_1 - p^*\|^2. \tag{3.27}$$

Proof. We use (3.1) and (3.21) to obtain

$$\|x_{k+1} - p^*\|^2 \leq Q_k\|y_k - p^*\|^2 - \xi\lambda_k(1-\lambda_k)\|y_k - T_k y_k\|^2.$$

Now, by (3.7), we deduce that

$$\begin{aligned}\|x_{k+1} - p^*\|^2 \leq Q_k & \left[(1+\alpha_k)\|x_k - p^*\|^2 + \alpha_k(1+\alpha_k)\|x_k - x_{k-1}\|^2 - \alpha_k\|x_{k-1} - p^*\|^2\right] \\ & - \xi\lambda_k(1-\lambda_k)\|y_k - T_k y_k\|^2.\end{aligned}$$

On the other hand, from (3.11), we get

$$\xi\delta_{k+1} \leq \xi\nu_k\alpha_k(1-\alpha_k)\|x_k - x_{k-1}\|^2 + \xi\lambda_k(1-\lambda_k)\|y_k - T_k y_k\|^2,$$

and the last two inequalities together imply (3.24). For the second part, inequalities (3.24) and (3.25) together give

$$\|x_{k+1} - p^*\|^2 + \xi\delta_{k+1} \leq Q_k\left[(1+\alpha_k)\|x_k - p^*\|^2 - \alpha_k\|x_{k-1} - p^*\|^2\right] + \xi Q_k\delta_k.$$

Subtracting $\alpha_k\|x_k - p^*\|^2$, we are left with

$$\begin{aligned}\tilde{C}_{k+1}(p^*) &\leq \big(Q_k(1+\alpha_k) - \alpha_k\big)\|x_k - p^*\|^2 - \alpha_k Q_k\|x_{k-1} - p^*\|^2 + \xi Q_k\delta_k \\ &\leq Q_k\|x_k - p^*\|^2 - Q_k\alpha_{k-1}\|x_{k-1} - p^*\|^2 + \xi Q_k\delta_k\end{aligned}$$

$$= \quad Q_k\tilde{C}_k(p^*),$$

where the second inequality comes from α_k being nondecreasing and $Q_k \leq 1$. This gives (3.26), recalling that $\tilde{C}_1(p^*) = \|x_1 - p^*\|^2$. Now, since $\|x_{k+1} - p^*\|^2 - \alpha_k\|x_k - p^*\|^2 \leq \tilde{C}_{k+1}(p^*)$, we have

$$\begin{aligned}\|x_{k+1} - p^*\|^2 &\leq \alpha_k\|x_k - p^*\|^2 + \left[\prod_{j=1}^{k} Q_j\right]\|x_1 - p^*\|^2 \\ &\leq \alpha\|x_k - p^*\|^2 + \left[\prod_{j=1}^{k} Q_j\right]\|x_1 - p^*\|^2,\end{aligned}$$

which we then iterate to obtain (3.27). □

The preceding estimations allow us to establish the main result of this section, namely:

Theorem 3.3.1. *Let (T_k) be a sequence of operators on $\mathcal{H}$, such that* $\mathrm{Fix}(T_k) \equiv \{p^*\}$ *and T_k is q_k-quasi-contractive for each $k \in \mathbb{N}$. Let (x_k, y_k) satisfy* (3.1)*, and let $\xi \in [0,1]$. Write $Q_k = Q(\lambda_k, q_k, \xi)$, and assume that* (3.25) *holds for all $k \in \mathbb{N}$. We have the following:*

i) If $\sum_{k=1}^{\infty} \lambda_k(1 - q_k^2) = \infty$, then x_k converges strongly to p^*, as $k \to \infty$.

ii) If $\lambda_k \geq \lambda > 0$ and $q_k \leq q < 1$ for all $k \in \mathbb{N}$, then x_k converges linearly to p^*, as $k \to \infty$. More precisely,

$$\|x_k - p^*\|^2 \leq \left[\frac{Q(\lambda, q, \xi)^{k+1} - \alpha^{k+1}}{Q(\lambda, q, \xi) - \alpha}\right]\|x_1 - p^*\|^2 = \mathcal{O}\left(Q(\lambda, q, \xi)^k\right). \tag{3.28}$$

Proof. For part i), write $p_k = \lambda_k(1 - q_k^2)$, and observe that $Q_k \leq 1 - p_k$, because Q increases with ξ. It ensues that

$$\prod_{k=1}^{K} Q_k \leq \prod_{k=1}^{K}(1 - p_k) = \exp\left[\sum_{k=1}^{K}\ln(1 - p_k)\right] \leq \exp\left[-\sum_{k=1}^{K} p_k\right]$$

since $\ln(1 - z) \leq -z$. If $\sum_{k=1}^{\infty} \lambda_k(1 - q_k^2) = \infty$, then $\prod_{k=1}^{\infty} Q_k = 0$. By (3.26), $\lim_{k\to\infty} \tilde{C}_k(p^*) = 0$. As in the proof of Theorem 3.2.1, we can show that the sum of the first two terms in $\tilde{C}_k(p^*)$, namely $\|x_k - p^*\|^2 - \alpha_{k-1}\|x_{k-1} - p^*\|^2$, is nonnegative. Therefore, $\lim_{k\to\infty}[\|x_k - p^*\|^2 - \alpha_{k-1}\|x_{k-1} - p^*\|^2] = 0$. If $\alpha_k \equiv 0$, the conclusion is straightforward. Otherwise, given any $\varepsilon > 0$, there is $K \in \mathbb{N}$ such that

$$\|x_k - p^*\|^2 \leq \alpha\|x_{k-1} - p^*\|^2 + \varepsilon$$

for all $k \geq K$, since α_k is nondecreasing. This implies

$$\|x_k - p^*\|^2 \leq \alpha^{k-K}\|x_K - p^*\|^2 + \varepsilon(1 - \alpha)^{-1},$$

so that $\limsup_{k\to\infty} \|x_k - p^*\| \leq \varepsilon(1 - \alpha)^{-1}$, and the conclusion follows.

For ii), we know that $Q(\lambda_k, q_k, \xi) \leq Q(\lambda, q, \xi)$, because Q increases either if λ decreases, and also if q increases. Gathering the common factors in the second and third terms on the left-hand side of inequality (3.25), we deduce that $Q \geq \alpha$ (strictly if $\alpha > 0$). Using (3.27), and observing that the case $Q(\lambda, q, \xi) = \alpha$ is incompatible with inequality (3.25), we deduce that

$$\begin{aligned}\|x_{k+1} - p^*\|^2 &\leq \alpha^k \left[\sum_{j=0}^{k} \left(\frac{Q(\lambda, q, \xi)}{\alpha} \right)^j \right] \|x_1 - p^*\|^2 \\ &= \left[\frac{\alpha^{k+1} - Q(\lambda, q, \xi)^{k+1}}{\alpha - Q(\lambda, q, \xi)} \right] \|x_1 - p^*\|^2,\end{aligned}$$

as claimed. □

3.3.2 Behavior with and without inertia

In the non-inertial case $\alpha_k \equiv 0$, (3.25) holds if either $\xi = 0$ or $\lambda_k \leq 1$ for all k, as in Hypothesis A. This is less restrictive than Hypothesis B (see Remark 3.2.2). To simplify the explanation, suppose $q_k \equiv q \in (0, 1)$. The best convergence rate is

$$\|x_k - p^*\| = \mathcal{O}\big(q^k\big),$$

obtained from Theorem 3.3.1 with $\lambda_k \equiv 1$ and $\xi = 0$. If $\alpha_k > 0$ for at least one k, the case $\xi = 0$ is ruled out, and

$$q^2 \leq (1 - \lambda_k + \lambda_k q)^2 = Q(\lambda_k, q, 0) \leq Q(\lambda_k, q, \xi) \leq Q(\lambda_k, q, 1) = 1 - \lambda_k + \lambda_k q^2.$$

All inequalities are strict if $\lambda_k \in (0, 1)$. This suggests that there may be operators for which the inertial step actually deteriorate the convergence, so inertial steps should be handled with caution and this can be seen as an argument *against* the use of inertia. Actually, it is possible to find a wide variety of behaviors, even for some of the simplest operators, as shown by the following case study:

Example 3.3.1. *Let $\lambda_k \equiv \lambda \in (0, 1)$ and $\alpha_k \equiv \alpha \in [0, 1)$. Take $q \in (0, 1]$, and consider the operator $T : \mathbb{R} \to \mathbb{R}$, defined by $Ty = -qy$, whose unique fixed point is the origin.*
If $\alpha = 0$, for each $k \geq 0$, we have $x_{k+1} = Lx_k$, where we have written $L = 1 - \lambda(1 + q)$. Iterating from $x_0 = 1$, we obtain $|x_k| = |L|^k$. If $\lambda(1 + q) = 1$, convergence occurs in one iteration.
Now, let $\alpha \in (0, 1)$, so that (3.1) reads

$$x_{k+1} = L\big(x_k + \alpha(x_k - x_{k-1})\big). \tag{3.29}$$

Here, we take $x_1 = x_0 = 1$. We can rewrite (3.29) in matrix form as

$$X_{k+1} = MX_k,$$

where

$$M = \begin{pmatrix} (1 + \alpha)L & -\alpha L \\ 1 & 0 \end{pmatrix}, \quad X_k = \begin{pmatrix} x_k \\ x_{k-1} \end{pmatrix}.$$

As before, convergence occurs in one step if $L = 0$. The eigenvalues of M are

$$\mu_{\pm} = \frac{(1 + \alpha)L \pm \sqrt{(1 + \alpha)^2 L^2 - 4\alpha L}}{2}.$$

Let us consider the case $L > 0$ first. If $(1+\alpha)^2L^2 < 4\alpha L$ (which is $\lambda(1+q) > (1-\alpha)^2/(1+\alpha)^2$), the eigenvalues are complex conjugates, both with modulus $|\mu_\pm| = \sqrt{\alpha L} < 1$. Now, $\sqrt{\alpha L} < L$ if, and only if, $L > \alpha$, which means that $\lambda(1+q) < 1-\alpha$. Since $|x_k| = \mathcal{O}(|\mu_\pm|)$, the inertial iterations converge strictly faster than the noninertial ones if

$$\frac{(1-\alpha)^2}{(1+\alpha)^2} < \lambda(1+q) < 1-\alpha.$$

If $L = \alpha$, the convergence rate is the same. Else, if $(1+\alpha)^2L^2 \geq 4\alpha L$, then M has two real eigenvalues (counting multiplicities), with $0 < \mu_- \leq \mu_+$. But since $L \in (0,1)$ implies $-L < -L^2$, we always have

$$\mu_+ < \frac{(1+\alpha)L + \sqrt{(1+\alpha)^2L^2 - 4\alpha L^2}}{2} = \frac{(1+\alpha)L + L\sqrt{(1-\alpha)^2}}{2} = L < 1.$$

Therefore, the inertial iterations also converge strictly faster if

$$0 < \lambda(1+q) \leq \frac{(1-\alpha)^2}{(1+\alpha)^2}.$$

When $L < 0$ ($\lambda(1+q) > 1$), the matrix M will always have two real eigenvalues, one of each sign. It is easy to verify that $|\mu_+| < |\mu_-|$, which implies that $|\mu_-|$ determines the convergence (the initial condition is not an eigenvector of M, so both eigenvalues intervene). But

$$\begin{aligned}\mu_- &= -\frac{(1+\alpha)|L| + \sqrt{(1+\alpha)^2L^2 + 4\alpha|L|}}{2}\\ &< -\frac{(1+\alpha)|L| + \sqrt{(1+\alpha)^2L^2}}{2}\\ &= -|L|\\ &= L.\end{aligned}$$

In this case, the inertial algorithm performs worse than the noninertial one. Moreover, the inertial iterations do not converge if $\mu_- \leq -1$, which is equivalent to

$$\lambda(1+q) \geq \frac{2(1+\alpha)}{1+2\alpha}.$$

A few comments are in order:

- *For $0 < \lambda(1+q) < 1-\alpha$, the inertial iterations converge at a strictly faster linear rate than the noninertial ones, even in the noncontracting case $q = 1$.*
- *At the transition point $\lambda(1+q) = 1-\alpha$ the convergence rate is the same.*
- *In the interval $1-\alpha < \lambda(1+q) < \frac{2(1+\alpha)}{1+2\alpha}$, the inertial step is counterproductive and noninertial iterations perform better, except for the singular value $\lambda(1+q) = 1$, where both converge in one iteration. In both cases, the closer $\lambda(1+q)$ is to 1, the faster the convergence.*

- *If* $\lambda(1+q) \geq \frac{2(1+\alpha)}{1+2\alpha}$, *the inertial iterations do not converge, while the noninertial ones do. Notice that, picking* λ *and* α *satisfying* (3.19) *can be read as picking* $\lambda < S(\alpha)$, *with* $S(\alpha) = \frac{(1-\alpha)^2}{1-\alpha+2\alpha^2}$. *Calling* $P(\alpha) = \frac{1+\alpha}{1+2\alpha}$, *it is easy to see that*

$$\lambda < S(\alpha) < P(\alpha) \leq \frac{2}{1+q} P(\alpha), \ \forall q \in (0,1].$$

Then $\lambda(1+q) < \frac{2(1+\alpha)}{1+2\alpha}$ *for all* $q \in (0,1]$. *Therefore, this last case is incompatible with Hypotheses* (A) *or* (B).

Now, the convergence rate results given by Theorem 3.3.1 correspond to worst-case scenarios, which certainly must include cases like the one discussed in Example 3.3.1. However, this situation need not be representative of other concrete instances found in practice, in which inertia improves either the theoretical convergence rate guarantees (see Subsection 3.4.2 below, and the commented references), or the actual behavior when the algorithm is implemented. In fact, the numerical tests reported below show noticeable improvements in the performance of the selected algorithms, upon adding the inertial substep.

3.3.3 Some insights into inequality (3.25)

To fix the ideas, we comment on some special cases of inequality (3.25), especially with constant parameters:

1. In the limiting case $q_k \equiv 1$, we have $Q_k \equiv 1$. With constant parameters $\lambda_k \equiv \lambda$, $\alpha_k \equiv \alpha$, (3.25) becomes
$$\lambda\alpha(1+\alpha) - \xi(1-\lambda)(1-\alpha)^2 \leq 0.$$
If
$$\frac{\alpha\lambda(1+\alpha)}{(1-\lambda)(1-\alpha)^2} \leq 1, \tag{3.30}$$
then, there is $\xi_{\alpha,\lambda,1} \in (0,1)$ such that (3.25) holds for all $\xi \in [\xi_{\alpha,\lambda,1}, 1]$. If $\xi = 1$, it is precisely the constant case in Hypothesis A (see (3.19) for a more direct comparison).
2. Keeping $\lambda_k \equiv \lambda \in (0,1)$, $\alpha_k \equiv \alpha \in (0,1)$, and fixing $\xi = 1$, let us take $q_k \equiv q \in (0,1)$. In this case, condition (3.25) is equivalent to
$$\Psi(\lambda) := (1+\alpha^2)(1-q^2)\lambda^2 - \big(2\alpha^2 + (1-\alpha)(2-q^2)\big)\lambda + (1-\alpha)^2 \geq 0. \tag{3.31}$$
Observe that $\Psi(0) = (1-\alpha)^2 > 0$, while $\Psi(1) = -\alpha q^2(1+\alpha) < 0$. Since Ψ is quadratic, the equation $\Psi(\lambda) = 0$ has exactly one root in $(0,1)$, which we denote by $\lambda_{\alpha,q}$. It follows that, for each $(\alpha, q) \in [0,1) \times (0,1)$, inequality (3.31) holds for all $\lambda \leq \lambda_{\alpha,q}$. The values of $\lambda_{\alpha,q}$ on $[0,1) \times (0,1)$ are depicted in Figure 3.1. Once a value for the inertial parameter α has been selected, the best theoretical convergence rate is
$$Q(\lambda_{\alpha,q}, q, 1) = 1 - \lambda_{\alpha,q}(1-q^2).$$
On the other hand, using the formula for the roots of a quadratic equation and some algebraic manipulations, we deduce that
$$\left[\frac{2\alpha^2 + (1-\alpha)}{2\alpha^2 + (1-\alpha)(2-q^2)}\right] \lambda_{\alpha,1} \leq \lambda_{\alpha,q} \leq \lambda_{\alpha,1}$$

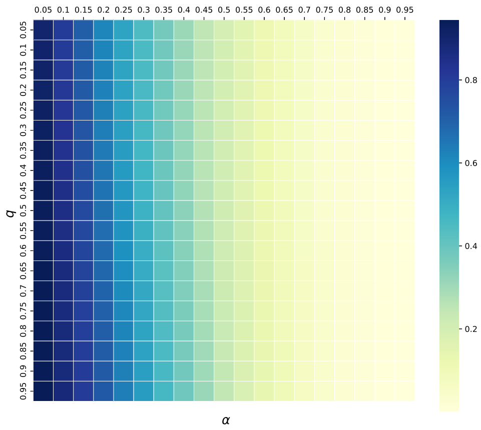


Figure 3.1: Values of $\lambda_{\alpha,q}$.

for every $(\alpha, q) \in [0,1) \times (0,1)$. Therefore, $\lambda_{\alpha,q} \to \lambda_{\alpha,1}$ as $q \to 1$, and there is no discontinuity as the contractive character is lost.

The case $\xi \in (0,1)$ is more involved. Lower values of ξ make the constant Q smaller, but may also restrict the possible values for α and λ, in view of inequality (3.25). In the fully general case, if α, λ and q satisfy

$$\left[\frac{\alpha\lambda(1+\alpha)}{(1-\alpha)(1-\lambda)}\right]\left[\frac{1-\lambda+\lambda q^2}{1-\lambda+\lambda q^2-\alpha}\right] < 1,$$

then, there is $\xi_{\alpha,\lambda,q} \in (0,1)$ such that (3.25) holds for all $\xi \in [\xi_{\alpha,\lambda,q}, 1]$. As $q \to 1$, we recover (3.30) as a limit case.

3.4 Examples

A broad class of algorithms to solve optimization and differential inclusion problems can be stated as fixed point iterations. Some well known techniques are introduced, all of them align with the framework presented in this chapter, that is, all of them preserve convergence after the inclusion of inertia.

3.4.1 Averaged Operators

An operator $T : \mathcal{H} \to \mathcal{H}$ is γ-averaged if there is a nonexpansive operator $R : \mathcal{H} \to \mathcal{H}$ such that $T = (1-\gamma)I + \gamma R$. In this case, $\text{Fix}(T) = \text{Fix}(R)$.

Let $R : \mathcal{H} \to \mathcal{H}$ be nonexpansive and let (γ_k) be a sequence in $(0,1)$. Setting $T_k = (1-\gamma_k)I + \gamma_k R$, (3.1) can be rewritten as

$$\left\{\begin{array}{rcl} y_k & = & x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} & = & (1-\gamma_k\lambda_k)y_k + \gamma_k\lambda_k R(y_k), \end{array}\right. \tag{3.32}$$

and Hypothesis B becomes

$$\limsup_{k\to\infty} \left[\alpha_k(1+\alpha_k) + \big((\gamma_k\lambda_k)^{-1}-1\big)\alpha_k(1-\alpha_k) - \big((\gamma_{k-1}\lambda_{k-1})^{-1}-1\big)(1-\alpha_{k-1})\right] < 0.$$

If $\gamma_k\lambda_k \to \eta > 0$, this is

$$\eta(1-\alpha+2\alpha^2) < (1-\alpha)^2. \tag{3.33}$$

It is not necessary to implement the algorithm using the operator R explicitly. However, the interval for the relaxation parameters is enlarged, and it may be convenient to over-relax. We shall come back to this point in the numerical illustrations.

3.4.2 Euler Iterations and Gradient Descent

An operator B is *β-cocoercive* with $\beta > 0$ if $\langle Bx-By, x-y\rangle \geq \beta \left\|Bx - By\right\|^2$ for all $x, y \in \mathcal{H}$.

Let $B : \mathcal{H} \to \mathcal{H}$ be cocoercive with constant β, and let (ρ_k) be a sequence in $(0, 2\beta)$. For each $k \geq 1$, set

$$T_k = I - \rho_k B.$$

Then, T_k is nonexpansive (thus quasi-nonexpansive) and $(\rho_k/2\beta)$-averaged. If $\rho_- := \inf_{k\geq 1}\rho_k > 0$, the family $(I - T_k)$ is asymptotically demiclosed. If $\lambda_k\rho_k \to \sigma$, Hypothesis B becomes

$$\sigma(1-\alpha+2\alpha^2) < 2\beta(1-\alpha)^2.$$

Now, let $f : \mathcal{H} \to \mathcal{H}$ be convex and differentiable, and assume ∇f is Lipschitz-continuous with constant L. Then, $B = \nabla f$ is cocoercive with constant $\beta = 1/L$. If, moreover, f is strongly convex with parameter μ and $\rho_k \leq 2/(L+\mu)$, then T_k is q_k-quasi-contractive with

$$q_k = 1 - \frac{2\mu L\rho_k}{L+\mu} \leq 1 - \frac{2\mu L\rho_-}{L+\mu} =: q.$$

Therefore, (T_k) is q-quasi-contractive. Considering the non-inertial case ($\alpha_k \equiv 0$), $\lambda_k \equiv 1$ and the fixed-sted choice $\rho_k = 2/(\mu+L)$, the algorithm exhibits a rate of convergence

$$f(x_k) - f^* \leq \frac{L}{2}\left(\frac{Q-1}{Q+1}\right)^{2k} \left\|x_0 - x^*\right\|^2,$$

where $Q = L/\mu$ is the *condition number* [70, Theorem 2.1.15]. Introducing the inertial term, and using

$$\rho_k = 1/L \qquad \text{and} \qquad \alpha_k \equiv \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right),$$

it turns into [70, Constant Step scheme, III], which has a rate of convergence of

$$f(x_k) - f^* \leq \min\left\{\left(1-\sqrt{\frac{\mu}{L}}\right)^k, \frac{4L}{(2\sqrt{L}+k\sqrt{\mu})^2}\right\}\left(f(x_0) - f^* + \frac{\mu}{2}\left\|x_0 - x^*\right\|^2\right).$$

Here, Hypothesis B can be written as

$$\lambda < \frac{2Q}{1-\sqrt{Q}+2Q},$$

which gives the condition for the convergence of Nesterov's constant step scheme with constant relaxation λ.

3.4.3 Proximal and Forward-Backward Methods

Let $M : \mathcal{H} \to 2^{\mathcal{H}}$ be maximally monotone and let (ρ_k) be a positive sequence. The *proximal method* consists in iterating

$$z_{k+1} = (I + \rho_k M)^{-1} z_k, \tag{3.34}$$

for $k \geq 1$. The operator $T_k = J_{\rho_k M} := (I + \rho_k M)^{-1}$ is nonexpansive, $\frac{1}{2}$-averaged, and $Z = \bigcap_{k\geq 1} \mathrm{Fix}(T_k) = M^{-1}0$. If $\lambda_k \to \lambda$, Hypothesis A is reduced to

$$\lambda(1-\alpha+2\alpha^2) < 2(1-\alpha)^2.$$

As before, the family $(I - T_k)$ is asymptotically demiclosed at 0 if $\inf_{k\geq 1} \rho_k > 0$. To see this, let (z_k) be a sequence in $\mathcal{H}$ such that $z_k \rightharpoonup z$ and $z_k - T_k z_k \to 0$. We must show that $0 \in Mz$. By the definition of T_k, we have

$$\frac{1}{\rho_k}(z_k - T_k z_k) \in M(T_k z_k).$$

The left-hand side converges strongly to zero, while $T_k z_k \rightharpoonup z$. We conclude by the weak-strong closedness of the graph of M.

Let $A : \mathcal{H} \to 2^{\mathcal{H}}$ be maximally monotone, let $B : \mathcal{H} \to \mathcal{H}$ be cocoercive with parameter β, and let (ρ_k) be a sequence in $(0, 2\beta)$. For each $k \geq 1$, set

$$T_k = (I + \rho_k A)^{-1}(I - \rho_k B).$$

Then, T_k is γ_k-averaged with $\gamma_k = 2\beta(4\beta - \rho_k)^{-1}$. If $\rho_k \to \rho$ and $\lambda_k \to \lambda$, then Hypothesis B is equivalent to

$$\lambda(1-\alpha+2\alpha^2) < \left(2 - \frac{\rho}{2\beta}\right)(1-\alpha)^2.$$

As in the proximal case, the family $(I - T_k)$ is asymptotically demiclosed at 0 if $\inf_{k\geq 1} \rho_k > 0$.

3.4.4 Douglas-Rachford and primal-dual splitting

Let $A, B : \mathcal{H} \to 2^{\mathcal{H}}$ be maximally monotone, and let (r_k) be a positive sequence. The *Douglas-Rachford* splitting method consists in iterating $z_{k+1} = T_{r_k} z_k$, for $k \geq 1$, where

$$T_r = J_{rA} \circ \big(2J_{rB} - I\big) + \big(I - J_{rB}\big) = \frac{1}{2}\big(I + (2J_{rA} - I) \circ (2J_{rB} - I)\big). \tag{3.35}$$

The second expression shows that T_r is averaged. Using the weak-strong closedness of the graphs of A and B, and a little algebra, one proves that the family $\big(I - T_{r_k}\big)$ is asymptotically demiclosed if $\inf_{k\geq 0} r_k > 0$. Finally, observe that $\operatorname{Zer}(A+B) = J_{rB}\operatorname{Fix}(T_r)$.

More generally, let X and Y be Hilbert spaces, and consider the *primal problem*, which is to find $\hat{x} \in X$ such that

$$0 \in A\hat{x} + L^*BL\hat{x},$$

where $A : X \to 2^X$ and $B : Y \to 2^Y$ are maximally monotone operators, and $L : X \to Y$ is linear and bounded. The *dual problem* is to find $\hat{y} \in Y$ such that

$$0 \in B^{-1}\hat{y} - LA^{-1}(-L^*\hat{y}).$$

The primal and dual solutions, namely $\hat{x}$ and $\hat{y}$, are linked by the inclusions

$$-L^*\hat{y} \in A\hat{x} \qquad \text{and} \qquad L\hat{x} \in B^{-1}\hat{y}.$$

Remark 3.4.1. *Let $f : X \to \mathbb{R} \cup \{+\infty\}$ and $g : Y \to \mathbb{R} \cup \{+\infty\}$ be closed and convex, and set $A = \partial f$ and $B = \partial g$. The inclusions above are the optimality conditions for the primal and dual (in the sense of Fenchel-Rockafellar) optimization problems*

$$\min_{x\in X}\{f(x) + g(Lx)\} \qquad \textit{and} \qquad \min_{y\in Y}\{g^*(y) + f^*(-L^*y)\}, \tag{3.36}$$

respectively. Douglas-Rachford splitting applied to $A = \partial g^*$ and $B = \partial\big(f^* \circ (-L^*)\big)$ yields the alternating direction method of multipliers (see [46]).

In order to find a primal-dual pair, the *primal-dual* splitting algorithm (see [30]) iterates:

$$\left\{\begin{array}{rcl} x_{k+1} & = & J_{\tau A}\big(x_k - \tau L^* y_k\big) \\ y_{k+1} & = & J_{\sigma B^{-1}}\big(y_k + \sigma L(2x_{k+1} - x_k)\big), \end{array}\right. \tag{3.37}$$

with $\tau\sigma\|L\|^2 \leq 1$. The algorithm can be expressed as $(x_{k+1}, y_{k+1}) = T(x_k, y_k)$, where $T : X \times Y \to X \times Y$ is a 1/2-averaged operator (see [18, Remark 4.34]).

An inertial version of the primal-dual iterations is given by

$$\left\{\begin{array}{l} (y_k, v_k) = (x_k, u_k) + \alpha_k\left[(x_k, u_k) - (x_{k-1}, u_{k-1})\right] \\ p_{k+1} = J_{\tau A}(y_k - \tau L^* v_k) \\ q_{k+1} = J_{\sigma B^{-1}}(v_k + \sigma L(2p_{k+1} - y_k)) \\ (x_{k+1}, u_{k+1}) = (1 - \lambda_k)(y_k, v_k) + \lambda_k(p_{k+1}, q_{k+1}), \end{array}\right. \tag{3.38}$$

with appropriate sequences α_k and λ_k.

In [25], the authors propose the *Split Douglas-Rachford* algorithm

$$\left\{\begin{array}{rcl} v_k & = & \Sigma\big(I - J_{\Sigma^{-1}B}\big)\big(Lx_k + \Sigma^{-1}y_k\big) \\ x_{k+1} & = & J_{\Upsilon A}\big(x_k - \Upsilon L^* v_k\big) \\ y_{k+1} & = & \Sigma L(x_{k+1} - x_k) + v_k, \end{array}\right. \tag{3.39}$$

where Υ and Σ are elliptic linear operators that induce an *ad-hoc* metric and account for preconditioning.

3.4.5 Three Operator Splitting

Given three maximally monotone operators A, B, C defined on the Hilbert space H, we wish to find $\hat{x} \in H$ such that

$$0 \in A\hat{x} + B\hat{x} + C\hat{x}. \tag{3.40}$$

If C is β-cocoercive, the three-operator splitting method [37] generates a sequence (z_k) by

$$\begin{cases} x_k^B = J_{\rho B}(z_k) \\ x_k^A = J_{\rho A}(2x_k^B - z_k - \rho C x_k^B) \\ z_{k+1} = z_k + \lambda_k(x_k^A - x_k^B) \end{cases} \tag{3.41}$$

starting from a point $z_0 \in H$. Here $\rho \in (0, 2\beta)$, $\lambda_k \in (0, 1/\gamma)$ and

$$\gamma = \frac{2\beta}{4\beta - \rho}. \tag{3.42}$$

This recurrence is generated by iterating the γ-averaged operator

$$T = I - J_{\rho B} + J_{\rho A} \circ (2J_{\rho B} - I - \rho C \circ J_{\rho B}),$$

and we have $\operatorname{Zer}(A + B + C) = J_{\rho B}(\operatorname{Fix} T)$. Also, it gives the forward-backward method if $B = 0$ and the Douglas-Rachford method if $C = 0$. An inertial version is given by

$$\begin{cases} u_k = z_k + \alpha_k(z_k - z_{k-1}) \\ x_k^B = J_{\rho B}(u_k) \\ x_k^A = J_{\rho A}(2x_k^B - u_k - \rho C x_k^B) \\ z_{k+1} = u_k + \lambda_k(x_k^A - x_k^B), \end{cases} \tag{3.43}$$

for appropriate choices of α_k, λ_k. One particular instance is given by the optimization problem

$$\min f(x) + g(x) + h(Lx), \tag{3.44}$$

where f, g, h are closed and convex, h has a $(1/\beta)$-Lipschitz-continuous gradient, and L is a bounded linear mapping.

Remark 3.4.2. *Notice that $\gamma > 1$, and then the feasible set for λ_k is greater than $(0, 1)$, allowing to consider the non-relaxed case $\lambda_k \equiv 1$.*

Remark 3.4.3. *There are two recent works that extend the results proposed by Davis and Yin, extending the feasible interval for the step sizes ρ from $(0, 2\beta)$ to $(0, 4\beta)$. In [7], they prove the convergence of algorithm (3.41) not using the fact that the operator T is averaged. Considering the same hypotheses over the three operators, $\rho \in (0, 4\beta)$ and $(\lambda_k)_k$ such that $\lambda_k \in (0, 2-\rho/(2\beta)]$ and $\sum \lambda_k(2-\frac{\rho}{2\beta}-\lambda_k) = \infty$[2], they prove weak convergence of the iterations defined by the Davis and Yin splitting algorithm (3.41).*

The same result is proved in [35] using a different strategy. If $\rho \in (0, 4\beta)$ and the relaxation parameter $\lambda \in (0, 2 - \frac{\rho}{2\beta})$, then the operator

$$(1 - \lambda)I + \lambda T = \lambda J_{\rho B} \circ (2J_{\rho A} - I - \rho C \circ J_{\rho A}) + I - \lambda J_{\rho A},$$

is averaged with parameter $\gamma = 2\lambda\beta/(4\beta - \rho)$.

This result also applies for the forward-backward method.

[2] Notice that not all of the values of λ_k are allowed to be equal to $2 - \rho/(2\beta)$, because that case is incompatible with the second hypothesis.

3.5 Numerical Illustrations

In this section, we test the performance of the algorithm given by iterations (3.1) in two of the settings described in Section 3.4. More precisely, we apply an inertial primal-dual splitting method to solve a TV-based denoising problem, and an inertial three-operator splitting algorithm to in-paint a corrupted image.

3.5.1 Primal-Dual Splitting and TV-based Denoising

The algorithm will be tested in an image processing framework. Consider the problem

$$\min_{x\in\mathbb{R}^{N_1\times N_2}} F^{TV}(x) := \frac{1}{2}\left\|Rx-b\right\|^2 + w\left\|\nabla x\right\|_1, \tag{3.45}$$

where $x \in \mathbb{R}^{N_1\times N_2}$ is an image to recover from a noisy observation $b \in \mathbb{R}^{M_1\times M_2}$, $R : \mathbb{R}^{N_1\times N_2} \to \mathbb{R}^{M_1\times M_2}$ is a blur operator, w is a positive parameter, and $\nabla : x \mapsto \nabla x = (D_1x, D_2x)$ is the classical discrete gradient, whose adjoint ∇^* is the discrete divergence. A formulation for the gradient and divergence operators can be seen on [28]. In these experiments, R will be a Gaussian blur of size 9×9, standard deviation 4 and relative boundary conditions (see [50] for details on the construction of the operator), and $w = 10^{-4}$. Considering the original image $\bar{x}$ in Figure 3.3a composed by 256×256 pixels, the observation b is generated as $b = R\bar{x} + e$, where e is an additive zero-mean white Gaussian noise with standard deviation 10^{-3} (Figure 3.3b).

Setting $f = 0$, $g : (u, v^1, v^2) \mapsto \frac{1}{2}\left\|u-b\right\|^2 + w\left\|v^1\right\|_1 + w\left\|v^2\right\|_1$ and $L : x \mapsto (Rx, D_1x, D_2x)$, the problem (3.45) can be formulated as (3.36), and solved via (3.38). Since

$$\operatorname{prox}_{\sigma g^*} : \begin{pmatrix} u \\ v^1 \\ v^2 \end{pmatrix} \mapsto \begin{pmatrix} \dfrac{u-\sigma b}{\sigma+1} \\ v^1 - \sigma\operatorname{prox}_{\frac{w}{\sigma}\|\cdot\|_1}\left(\dfrac{v^1}{\sigma}\right) \\ v^2 - \sigma\operatorname{prox}_{\frac{w}{\sigma}\|\cdot\|_1}\left(\dfrac{v^2}{\sigma}\right) \end{pmatrix} \tag{3.46}$$

we are lead to Algorithm 2.

For a stopping criterion, we consider the relative error

$$\mathcal{R}(x_{k+1}, x_k) \mapsto \frac{\|x_{k+1}-x_k\|}{\|x_k\|}. \tag{3.47}$$

Since the involved operator is 1/2-averaged (see [26]), we may set $\lambda_k \equiv \lambda \in (0,2)$, as explained in Section 3.4.1.

The algorithm is tested for 17 combinations of τ, σ satisfying the critical condition $\tau\sigma\left\|L\right\|^2 = 1$ (according to [25], this tends to yield the best performance). The number $\|L\|$ is computed using an adaptation of [77, Algorithm 12].

Algorithm 2:

```
Choose $x_0, x_1 \in \mathbb{R}^{N_1 \times N_2}$, $u_0, u_1 \in \mathbb{R}^{m_1 \times m_2}$, $v_0^1, v_1^1, v_0^2, v_1^2 \in \mathbb{R}^{N_1 \times N_2}$, $(\lambda_k)_{k\in\mathbb{N}}$ and $(\alpha_k)_{k\in\mathbb{N}}$ such that
  hypotheses of Theorem 3.2.1 are fulfilled, $\tau$ and $\sigma$ such that $\tau\sigma \|L\|^2 \leq 1$, $\varepsilon > 0$ and $r_0 > \varepsilon$ ;
while $r_k > \varepsilon$ do
    $(\bar{x}_k, \bar{u}_k, \bar{v}_k^1, \bar{v}_k^2) = (x_k, u_k, v_k^1, v_k^2) + \alpha_k[(x_k, u_k, v_k^1, v_k^2) - (x_{k-1}, u_{k-1}, v_{k-1}^1, v_{k-1}^2)]$;
    $p_{k+1} = \bar{x}_k - \tau R^* \bar{u}_k - \tau D_1^* \bar{v}_k^1 - \tau D_2^* \bar{v}_k^2$;
    $q_{k+1} = (\bar{u}_k + \sigma R(2p_{k+1} - \bar{x}_k) - \sigma b)/(\sigma + 1)$;
    $w_{k+1}^1 = \bar{v}_k^1 + \sigma D_1(2p_{k+1} - \bar{x}_k) - \sigma \operatorname{prox}_{w\|\cdot\|_1/\sigma}(\bar{v}_k^1/\sigma + D_1(2p_{k+1} - \bar{x}_k))$;
    $w_{k+1}^2 = \bar{v}_k^2 + \sigma D_2(2p_{k+1} - \bar{x}_k) - \sigma \operatorname{prox}_{w\|\cdot\|_1/\sigma}(\bar{v}_k^2/\sigma + D_2(2p_{k+1} - \bar{x}_k))$;
    $(x_{k+1}, u_{k+1}, v_{k+1}^1, v_{k+1}^2) = (1 - \lambda_k)(\bar{x}_k, \bar{u}_k, \bar{v}_k^1, \bar{v}_k^2) + \lambda_k(p_{k+1}, q_{k+1}, w_{k+1}^1, w_{k+1}^2)$ ;
    $r_k = \mathcal{R}((x_{k+1}, u_{k+1}, v_{k+1}^1, v_{k+1}^2), (x_k, u_k, v_k^1, v_k^2))$ ;
end
return $(x_{k+1}, u_{k+1}, v_{k+1}^1, v_{k+1}^2)$
```

Comparison in terms of the parameters τ and σ. In a first stage, we compare the performance of the primal-dual splitting algorithm given by (3.37) (that is, Algorithm 2 with $\alpha_k \equiv 0$), and its inertial counterpart (3.38), with $\lambda_k \equiv 1$. The sequence $(\alpha_k)_{k\in\mathbb{N}}$ is

$$\alpha_k = \alpha \left(1 - \frac{1}{k^2}\right), \tag{3.48}$$

with $\alpha = 1/(3 + 0.0001)$ (condition (3.33) with $\eta = \lambda/2$ gives the constraint $\alpha < 1/3$). Table 3.1 shows the execution time, number of iterations, and the value for the objective value reached, using a tolerance $\varepsilon = 10^{-5}$. These results are depicted graphically, along with the percentage of reduction, in Figure 3.2. The recovered images are collected in Figures 3.3c and 3.3d.

			Original algorithm			Inertial algorithm		
Case	τ	σ	Time	Iters.	$F^{TV}(x)$	Time	Iters.	$F^{TV}(x)$
1	0.0004	282.8427	72.59	1565	7.30	55.11	1095	7.13
2	0.0010	122.6475	115.66	2437	2.84	86.97	1741	2.66
3	0.0024	53.183	110.16	2330	1.35	83.98	1672	1.27
4	0.0054	23.0614	98.28	2077	0.7566	72.33	1446	0.7341
5	0.0125	10	94.80	2015	0.4624	69.59	1394	0.4537
6	0.0288	4.3362	105.19	2253	0.2975	77.83	1562	0.2928
7	0.0665	1.8803	122.23	2593	0.2107	89.83	1773	0.2091
8	0.1533	0.8153	156.34	3248	0.1592	112.09	2184	0.1589
9	0.3536	0.3536	140.91	2922	0.1428	101.69	1956	0.1427
10	0.8153	0.1533	139.50	2856	0.1350	98.97	1908	0.1350
11	1.8803	0.0665	151.08	3123	0.1312	107.72	2084	0.1312
12	4.3362	0.0288	108.08	2249	0.1303	78.03	1503	0.1303
13	10	0.0125	60.28	1238	0.1301	42.78	833	0.1301
14	23.0614	0.0054	47.61	983	0.1302	35.70	693	0.1302
15	53.1830	0.0024	70.78	1466	0.1302	54.61	1065	0.1302
16	122.6475	0.0010	119.22	2471	0.1302	89.91	1762	0.1302
17	282.8427	0.0004	179.22	3767	0.1302	150.52	2999	0.1302

Table 3.1: Execution time, number of iterations and final function value for the original primal-dual algorithm and the inertial version, with tolerance $\varepsilon = 10^{-5}$.

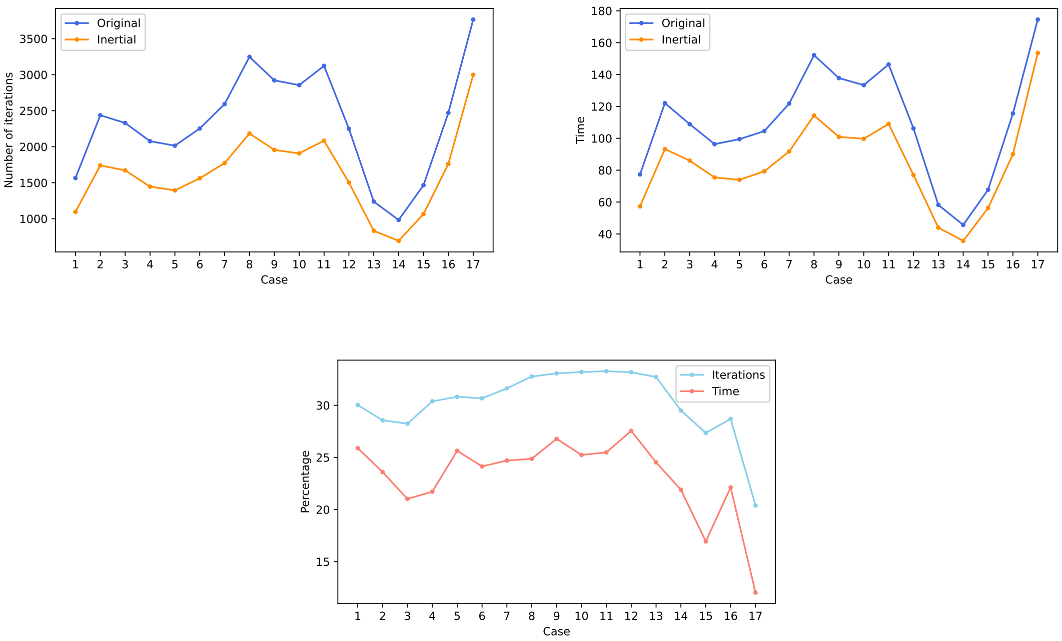
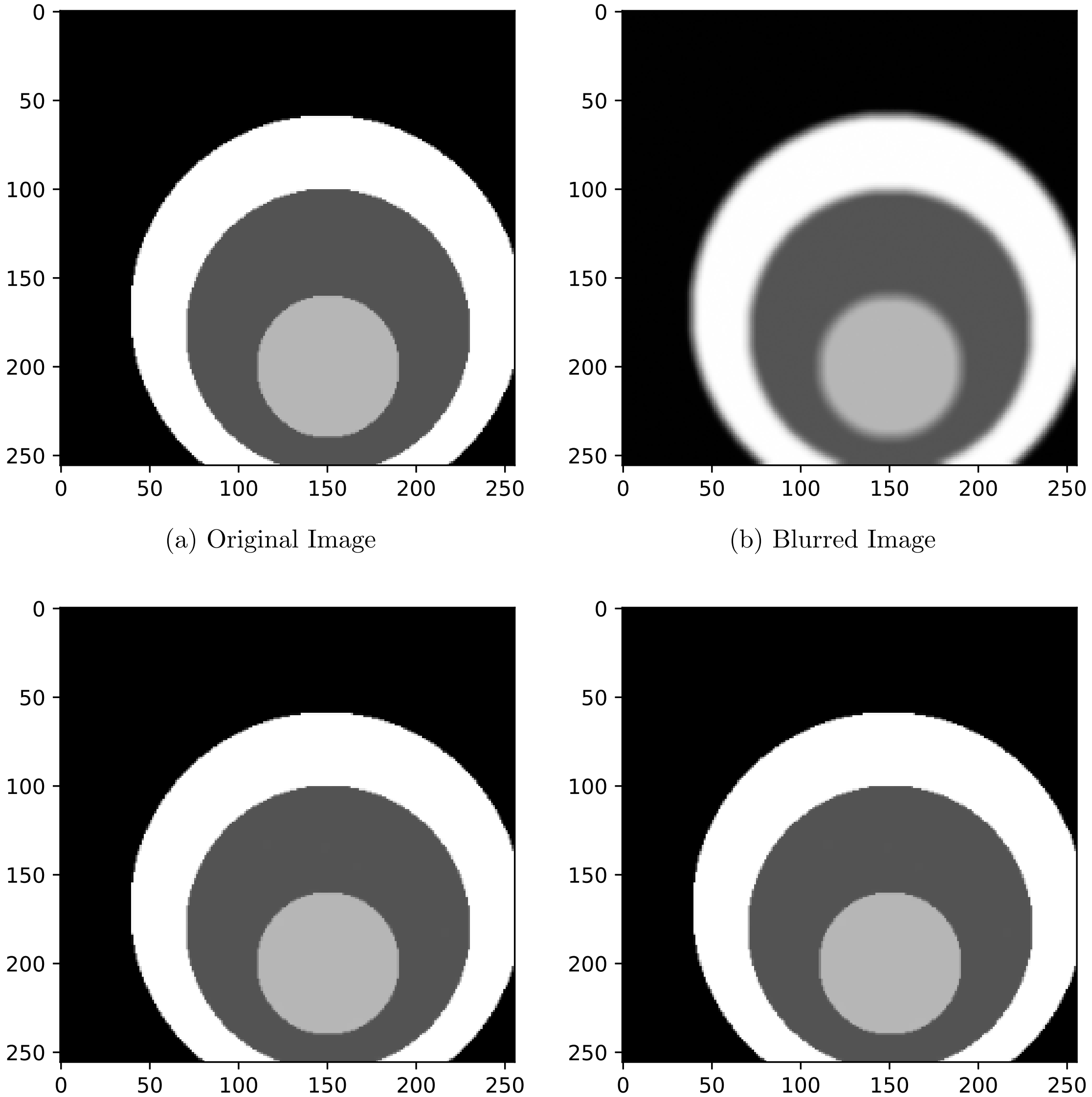


Figure 3.2: Number of iterations (top left), execution time (top right), and percentage of reduction (bottom), from Table 3.1.

Comparison in terms of the relaxation parameter λ. For both algorithms, case 14 showed the best performance in terms of iterations and execution time. We now assess the performance of the inertial algorithm with different values for $\lambda_k \equiv \lambda \in (0, 2)$, and the corresponding inertial parameters fulfilling condition (3.33). The results are shown in Tables 3.2 and 3.3, along with the value of α used in (3.48). A graphic depiction is shown as heatmaps in Figure 3.4. Larger values of the relaxation parameter λ resulted in an improvement in the performance of both algorithms, but limit the impact of inertia, as it reduces the feasible range for the limit α. A more thorough study on the selection of these parameters is the object of a forthcoming article.

		Original algorithm			Inertial algorithm		
λ	α	Time	Iterations	$F^{TV}(x)$	Time	Iterations	$F^{TV}(x)$
0.2	0.6534	119.16	2592	0.1303	49.23	992	0.1304
0.4	0.5425	74.44	1589	0.1302	40.45	799	0.1303
0.6	0.4619	62.28	1341	0.1302	39.06	773	0.1302
0.8	0.3943	54.05	1146	0.1302	33.94	730	0.1302
1.0	0.3333	46.12	983	0.1302	34.47	693	0.1302
1.2	0.2748	41.16	861	0.1301	35.17	684	0.1302
1.4	0.1352	38.22	771	0.1301	34.45	675	0.1301
1.6	0.0967	33.89	718	0.1301	33.59	655	0.1301
1.8	0.0535	32.28	679	0.1301	32.62	657	0.1301

Table 3.2: Execution time, number of iterations and final function value for the original primal-dual algorithm and the inertial version (case 14), with tolerance $\varepsilon = 10^{-5}$.



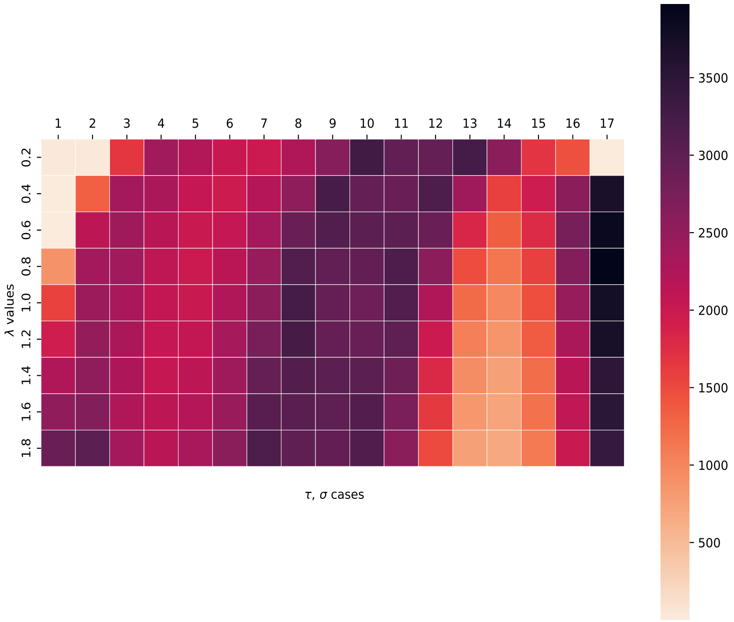
(a) Original Image

(b) Blurred Image

(c) Recovered without inertia

(d) Recovered with inertia

Figure 3.3: Original, blurred and recovered images. Lowest recovered value $F^{TV}(x) = 0.1301$ (case 13, both methods).



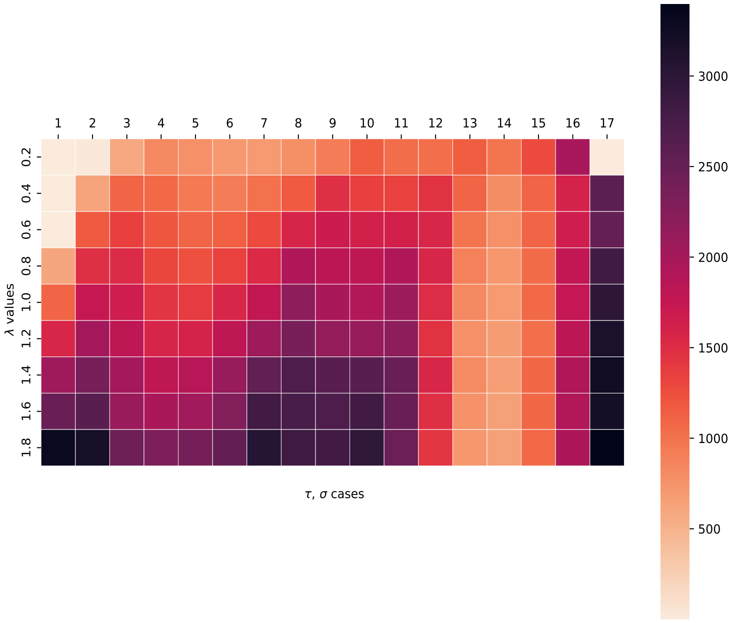


Figure 3.4: Average number of iterations performed by the original (top) and inertial (bottom) algorithms, with tolerance $\varepsilon = 10^{-5}$, for each value of λ, and each case of τ and σ, from Table 3.2.

λ	α	% Iterations reduction	% Time reduction
0.2	0.6534	61.73	58.69
0.4	0.5425	49.72	45.66
0.6	0.4619	42.36	37.28
0.8	0.3943	36.30	37.21
1.0	0.3333	29.50	25.26
1.2	0.2748	20.56	14.55
1.4	0.1352	12.45	9.86
1.6	0.0967	8.77	0.89
1.8	0.0535	3.24	-1.05

Table 3.3: Reduction percentage for the original primal-dual algorithm and the inertial version (case 14), with tolerance $\varepsilon = 10^{-5}$.

Finally, Figure 3.5 shows the evolution of the function values, the distance to the limit and the residuals, all in logarithmic scale, for case 14. The figure also includes the plot of $k\left\|z_k - Tz_k\right\|^2$. Theorem 3.2.1 states that the residuals show an non-asymptotic rate given by (3.14), so we can conjecture an asymptotic rate of $o(1/k)$.

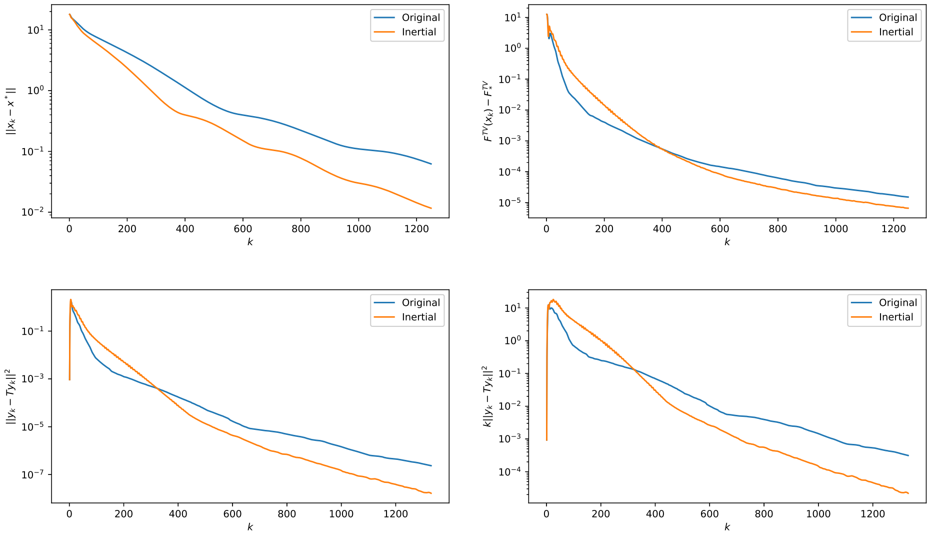


Figure 3.5: Evolution to the distance to the computed solution (top left), objective function values (top right), residuals $\left\|z_k - Tz_k\right\|^2$ (bottom left) and $k\left\|z_k - Tz_k\right\|^2$ (bottom right), for case 14.

3.5.2 Three-Operator Splitting and Image In-painting

Suppose that Z is a color image represented as a 3-D tensor where $Z(:,:,1), Z(:,:,2), Z(:,:,3)$ are the red, green and blue channels, respectively. Consider a damaged image Y, with randomly erased pixels, represented by the white color. The positions of the erased pixels are known. Denote $\mathcal{A}$ the linear operator that selects the set of correct entries of Z (and so $\mathcal{A}^*$ is the *zero upsampling* operator). The objective is to recover the image, by filling the erased pixels. Following [37] we consider the following formulation of the in-panting problem:

$$\min_{Z\in\mathcal{H}} F(Z) := \frac{1}{2}\left\|\mathcal{A}(Z-Y)\right\|^2 + w\left\|Z_{(1)}\right\|_* + w\left\|Z_{(2)}\right\|_*, \tag{3.49}$$

where $\mathcal{H}$ is the set of 3-D tensors, $Z_{(1)}$ is the matrix $[Z(:,:,1)\, Z(:,:,2)\, Z(:,:,3)]$, $Z_{(2)}$ is the matrix $[Z(:,:,1)^T\, Z(:,:,2)^T\, Z(:,:,3)^T]^T$, $\|\cdot\|_*$ denotes the matrix nuclear norm and w is a penalty parameter, which we take equal to 1 here, for simplicity. This problem fits in the context of (3.44), with $f(Z) = g(Z) = \|Z\|_*$ and $h(Z) = \frac{1}{2}\|Z-Y\|_2^2$. In this case, the operator $\nabla(h\circ\mathcal{A})$ is cocoercive with constant 1. With the error function $\mathcal{R}$ defined in (3.47), the iterations defined by (3.43) lead to Algorithm 3.

Algorithm 3:

```
Choose $Z_0, Z_1 \in \mathbb{R}^{m\times n}$, $(\lambda_k)_{k\in\mathbb{N}}$ and $(\alpha_k)_{k\in\mathbb{N}}$ such that hypotheses of Theorem 3.2.1 are fulfilled,
  $\rho \in (0,2)$, $\varepsilon > 0$ and $r_0 > \varepsilon$ ;
while $r_k > \varepsilon$ do
    $U_k = Z_k + \alpha_k(Z_k - Z_{k-1})$;
    $X_k^g = \mathrm{prox}_{\rho g}(U_k)$;
    $Z_{k+\frac{1}{2}} = 2X_k^g - U_k - \rho\mathcal{A}^*\nabla h(\mathcal{A}X_k^g)$;
    $Z_{k+1} = U_k + \lambda_k(\mathrm{prox}_{\rho f}(Z_{k+\frac{1}{2}}) - X_k^g)$;
    $r_{k+1} = \mathcal{R}(Z_{k+1}, Z_k)$
end
Return $Z_{n+1}, X_n^g$;
```

As in the previous section, Algorithm 3 will be tested in the case $\alpha_k \equiv 0$ (the algorithm studied in [37]) and, for the inertial version,

$$\alpha_k = \left(1 - \frac{1}{k}\right)\alpha, \tag{3.50}$$

where α satisfies the condition (3.33). The corresponding algorithms will be referred to as original and inertial, respectively. Algorithm (3) returns both the value of Z_k and X_k^g, since the latter represents the image solution of the problem. Throughout this section, the initial points are both set to zero.

Comparison in terms of the number of erased pixels. Between 10000 and 250000 pixels are randomly erased from the image in Figure 3.10a to obtain the one in Figure 3.10b. We compare the number of iterations and execution time needed by both methods with step size $\rho = 1$ and $\lambda_k \equiv 1$, for a tolerance of 10^{-3}. The results are shown in Figure 3.6. The reduction stands between 12% and 22% in most cases, and the improvement seems to increase with the number of erased pixels.

Comparison in terms of the step size. Both algorithms are tested for the same image with 250000 randomly erased pixels for different values of the step size ρ. As $\beta = 1$, Remark 3.4.3 allow us to choose $\rho \in (0,4)$ and $\lambda_k \equiv \lambda \in (0, 2-\rho/(2\beta))$. Then for each ρ we use $\lambda = 1 - \rho/(4\beta)$. For the inertial version, the constant α in (3.50) is adapted accordingly. The results are reported in Table 3.4 and depicted graphically in Figure 3.7. The percentage of reduction is noticeably higher for lower values of ρ (always above 35% when $\rho \leq 2$). This is to be expected, since larger values of ρ require lower values of α, which limits the effect of inertia.

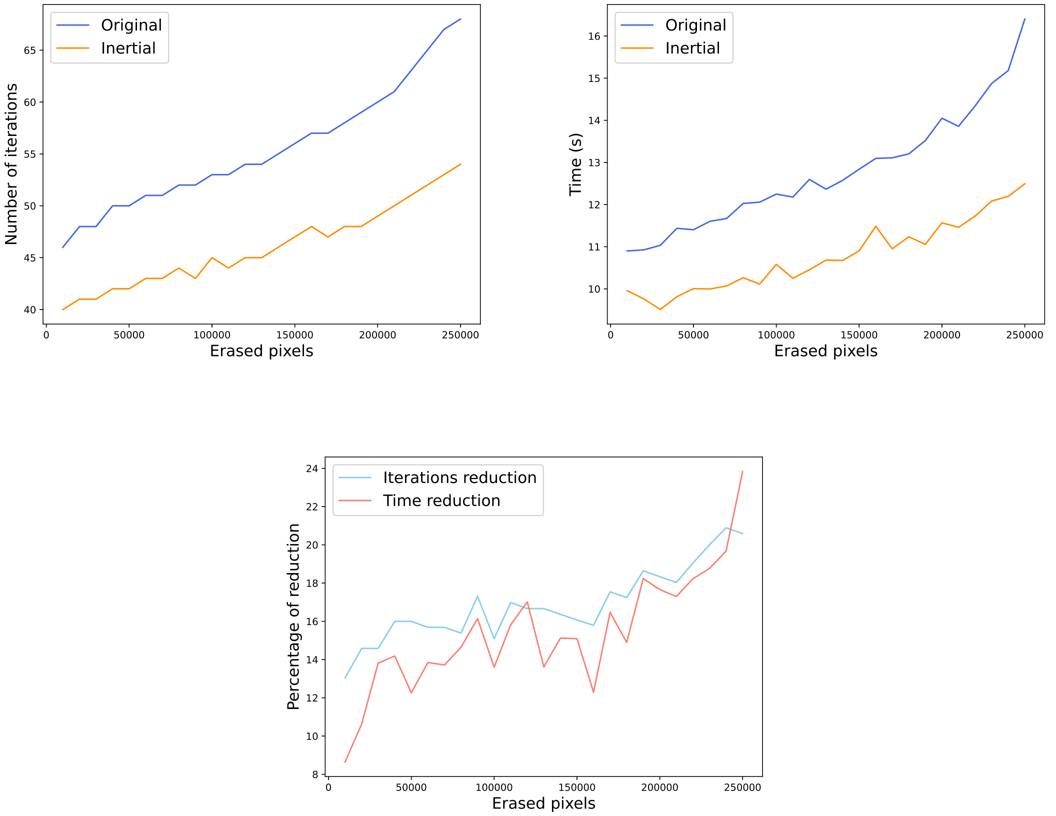


Figure 3.6: Number of iterations (top left), execution time (top right) and percentage of reduction (bottom) in terms of the number of erased pixels, with step size $\rho = 1$ and relaxation parameter $\lambda_k \equiv 1$, for a tolerance of 10^{-3}.

Comparison in terms of the relaxation parameter. Finally, we fix the value $\rho = 1$, and compare the performance of the two methods for different values of the relaxation parameter λ, which, as before, limit the possible range for the inertial parameter α in view of condition (3.33). The results are presented in Table 3.5, and shown graphically in Figure 3.8. As with the step size, the reduction is greater for lower values of λ, which is consistent with the loss of the inertial character imposed by condition (3.33). Nevertheless, observe that over-relaxing with $\lambda = 1.2$ or $\lambda = 1.4$ gives better results (both in number of iterations and execution time) than keeping λ in a neighborhood of 1.

The evolution of the function values, the distance to the limit and the residuals are shown (in logarithmic scale) in Figure 3.9 for 250000 erased pixels, using $\rho = 1$ and $\lambda_k \equiv 1$. As in the previous example, the sequence $k \left\|z_k - Tz_k\right\|^2$ tends to zero, allowing us to conjecture again an asymptotic rate of $o(1/k)$. Finally, Figure 3.10 shows the original, corrupted (with 250000 erased pixels) and recovered images.[3]

[3] For the sake of a fair visual comparison, we follow the implementation used in [37], as described in https://damek.github.io/ThreeOperators.html, which differs slightly from the description given in Section 3.4.5 in that it contains a Bregman update.

	Original algorithm		Inertial algorithm	
ρ	Time (s)	Iterations	Time (s)	Iterations
0.2	128.77	295	67.29	158
0.4	62.03	166	17.70	88
0.6	23.90	121	12.54	63
0.8	19.85	99	10.61	52
1.0	18.22	88	9.56	48
1.2	16.62	83	10.03	50
1.4	16.77	83	10.48	52
1.6	17.64	87	11.41	56
1.8	18.92	93	12.25	60
2.0	21.45	100	13.54	65
2.2	22.42	108	15.12	71
2.4	24.36	117	16.36	78
2.6	27.36	129	20.21	86
2.8	33.87	144	20.84	96
3.0	35.10	163	24.09	109
3.2	40.82	190	28.12	126
3.4	50.19	230	33.43	152
3.6	68.55	302	43.52	197
3.8	108.05	486	67.46	307

Table 3.4: Execution time and number of iterations in terms of the step size ρ.

	Original algorithm		Inertial algorithm	
λ	Time (s)	Iterations	Time (s)	Iterations
0.6	24.47	108	13.57	56
0.7	21.28	94	11.65	51
0.8	18.67	83	12.64	55
0.9	16.94	75	12.76	56
1.0	15.52	69	12.76	56
1.1	14.28	63	12.51	55
1.2	13.35	59	12.53	54
1.3	12.52	55	11.90	52
1.4	12.04	52	11.71	51

Table 3.5: Execution time and number of iterations for different values of λ.

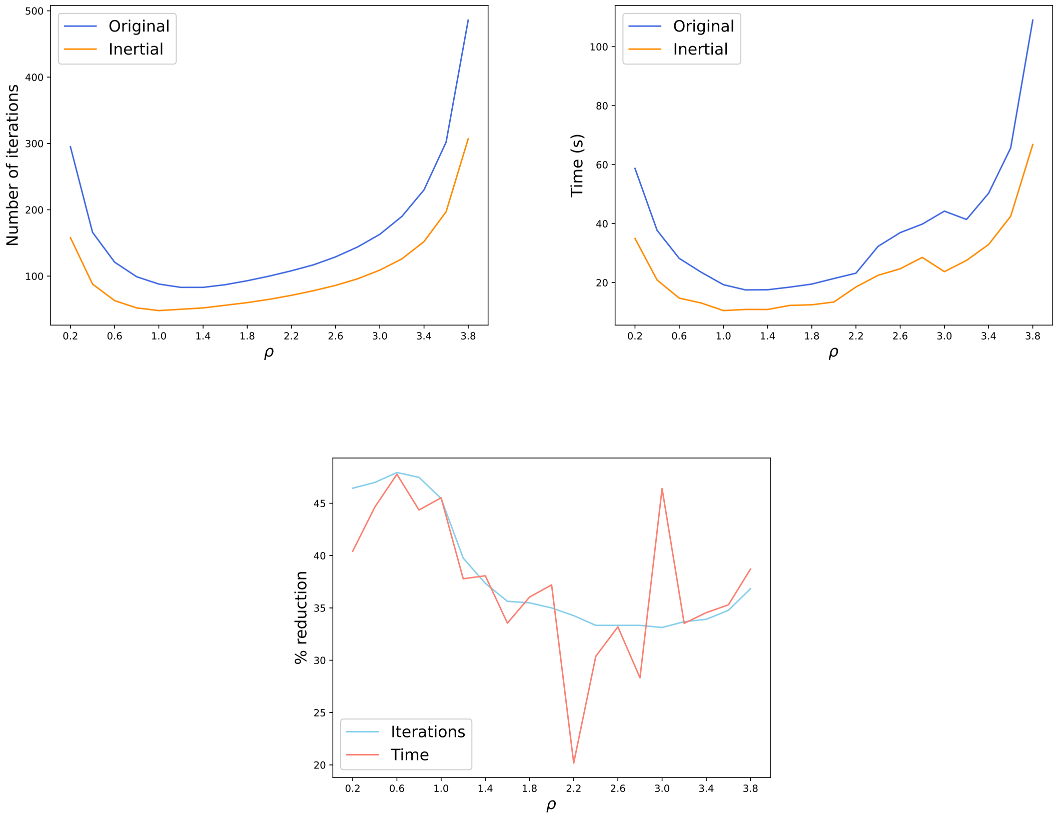


Figure 3.7: Number of iterations (top left), execution time (top right) and percentage of reduction (bottom) in terms of the step size ρ.

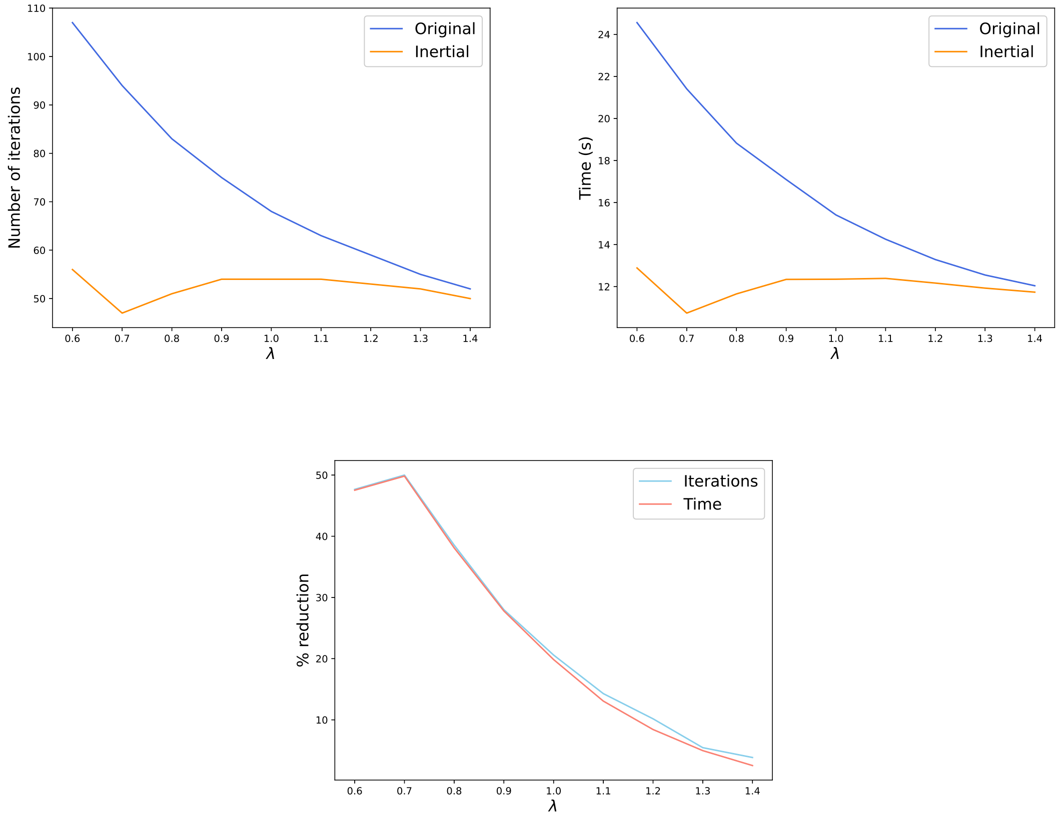


Figure 3.8: Number of iterations (top left), execution time (top right) and percentage of reduction (bottom) in terms of the relaxation parameter λ.

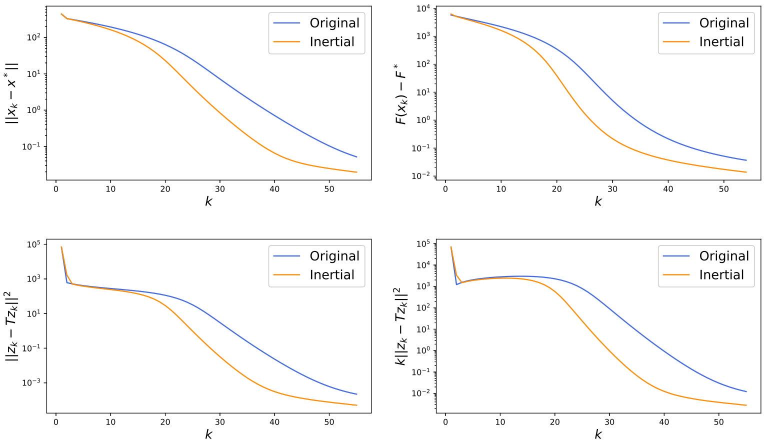
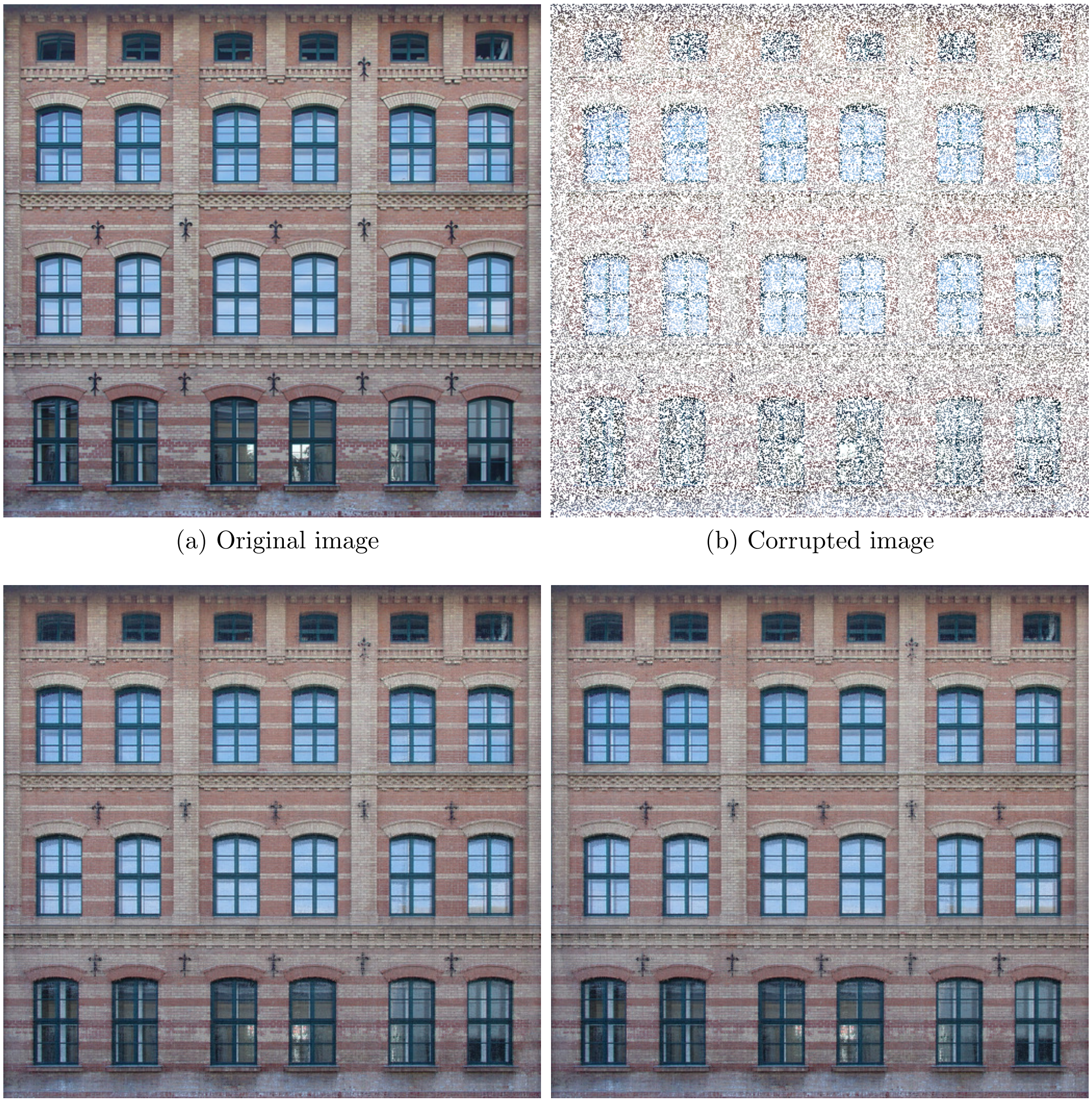


Figure 3.9: Evolution to the distance to the computed solution (top left), objective function values (top right), residuals $\|z_k - Tz_k\|^2$ (bottom left) and $k\|z_k - Tz_k\|^2$ (bottom right), for 250000 erased pixels using $\rho = 1$ and $\lambda_k \equiv 1$.



(a) Original image

(b) Corrupted image

(c) Recovered without inertia

(d) Recovered with inertia

Figure 3.10: Original image (a), corrupted image with 250000 randomly erased pixels (b), images recovered without inertia (c), and with inertia (d).

3.6 Performance analysis for KM iterations

On [43], Drori and Teboulle studied rates of convergence for first-order algorithms in a novel manner: the worst-case behavior of a certain algorithm is an optimization problem, and they refer to it as Performance Estimation Problem (PEP).

Consider a fixed value $n \in \mathbb{N}$, and a first order algorithm $\mathcal{A}$ over a class of functions $\mathcal{F}$ defined over $\mathbb{R}^d$. The algorithm iterates from a starting point $x_0 \in \mathbb{R}^d$, generating a finite sequence of points $x_k \in \mathbb{R}^d$, with $k = 1, \ldots, n$, aiming to minimize the values of an objective function $f \in \mathcal{F}$. We refer to $\mathcal{A}$ as a first order algorithm because at each step, it depends only on the previous steps, their function values and their gradients, that is,

$$x_{k+1} = \mathcal{A}(x_0, \ldots, x_k, f(x_0), \ldots, f(x_k), \nabla f(x_0), \ldots, \nabla f(x_k)), \quad k = 1, \ldots, n.$$

Considering a minimizer x^* of f, the problem of the worst case convergence for the function values can be stated as follows: to maximize the value of $f(x_n) - f(x^*)$ subject to the constraints that the sequence x_k must be generated by the algorithm $\mathcal{A}$, and that f is a function over the class $\mathcal{F}$. Then, the PEP has the following structure:

$$\begin{cases} \max & f(x_{n+1}) - f(x^*) \\ \text{s.t.} & x_{k+1} \text{ is generated by } \mathcal{A}, \text{ for } k = 1, \ldots, n, \\ & f \in \mathcal{F}, \\ & x^* \text{is a minimizer of } f, \\ & \|x_0 - x^*\| \leq R, \\ & x_0, \ldots, x_n, x^* \in \mathbb{R}^d. \end{cases} \tag{3.51}$$

The condition $\|x_0 - x^*\| \leq R$, with $R > 0$, is a common practice, and it states that the distance from the starting point of the algorithm to the minimizer is bounded. At a first glance, problem (3.51) seems too general to solve, but on [43], the problem is stated for one of the most common first order algorithms: the gradient method. In that case, $\mathcal{F}$ is the class of convex and differentiable functions with L-Lipschitz gradient. Using properties of this class of functions and duality arguments, the PEP for the gradient method can be stated as a semidefinite program (SDP), and find an optimal solution, which matches the known tight rate for the algorithm.

The procedure to state the infinite dimensional problem (3.51) into a SDP relies on using the concept of *interpolation*. Taylor, Hendrickx and Glineur have made significant contributions through two research articles that extend the formulation of Drori and Teboulle, and formalize the interpolation in the context of the PEP. In [86], they provide conditions of interpolation of first order methods for smooth strongly convex functions. In [85] they extend their analysis to composite convex functions and a larger class of first order algorithms for nonsmooth functions, such as the proximal method and its variations.

Ryu, Taylor, Bergeling and Gisselson in [80] set a PEP to find the tightest contraction factor for a nonexpansive operator. That is, to find the greater positive ρ such that $\|Tx - Ty\| \leq \rho \|x - y\|$. They focus mainly on operators defining splitting algorithms such as Davis-Yin or Douglas-Rachford. In the case of Douglas-Rachford algorithm, they find explicit tight bounds.

The main idea behind PEP, that is, setting the problem of finding tight rates for algorithms has inspired research in various domains. For example, Mourcer, Taylor and Bach in

[64] extend the PEP to continuous dynamics involving convex functions, aiming to estimate the Lyapunov functions, even in a stochastic context. See also [87, 84] where they estimate Lyapunov functions for first order methods.

In the following, results of an ongoing joint work with Juan Pablo Contreras[4] are presented. Building on the concepts discussed previously, we aim to set a PEP for inertial KM iterations. As of our current knowledge, there are no results of tight rates for the inertial iterations presented on this chapter. Consequently, PEP emerges as a promising strategy to address this challenge. We focus on maximizing the residual of the iterations $\|y_n - Ty_n\|$, for fixed parameters α, λ and a fixed nonexpansive operator T. Just as the Drori and Teboulle approach, we show that PEP for inertial KM iterations can be written as a SDP. Although this is an ongoing research by the time of writing of the thesis, it serves as both motivation for future work and provides valuable insights into approaching unresolved questions.

3.6.1 PEP for inertial KM iterations

Consider the Krasnoselskii-Mann (KM) iterations:

$$x_{k+1} = (1-\lambda)x_k + \lambda T(x_k), \tag{3.52}$$

where $T : \mathcal{H} \to \mathcal{H}$ is a nonexpansive mapping and fixed averaging parameter $\lambda \in (0,1)$. If $x_* \in \operatorname{Fix}(T)$ and $x_0 \in \mathcal{H}$, an optimal error bound for the residuals can be stated (see [54]), which is

$$\|x_k - T(x_k)\|^2 \leq \frac{1}{k+1}\left(\frac{k}{k+1}\right)^k \frac{\|x_0 - x_*\|^2}{\lambda(1-\lambda)}. \tag{3.53}$$

The previous rate is proved to be tight, and the result is achieved by using semidefinite programming. This strategy, and the Performance Estimation Problem (PEP) described before motivate us to perform a numerical analysis for the rate of the residuals in the inertial KM iterations with fixed parameters: given x_0, $x_1 \in \mathcal{H}$, $\alpha \in [0,1)$, $\lambda \in (0,1)$,

$$\begin{cases} y_k &= x_k + \alpha(x_k - x_{k-1}) \\ x_{k+1} &= (1-\lambda)y_k + \lambda T y_k. \end{cases} \tag{3.54}$$

If λ, α satisfy

$$\lambda < \frac{(1-\alpha)^2}{(1-\alpha+2\alpha^2)}, \tag{3.55}$$

then Theorem 3.2.2 gives that both x_k and y_k converge weakly to a point in $\operatorname{Fix}(T)$. Given parameters α and λ, the tight rate of convergence corresponds to the worst case of convergence for an operator T. In other words, to find the operator T which gives the slowest speed of convergence. Therefore, we aim to maximize the value for the residuals, under the constraints of nonexpansiveness of the operator, the existence of fixed points, and that the sequence is generated by KM iterations. That is, the problem, in a general form, can be stated as

$$\begin{cases} \max\limits_{x_0, x_1, x_*} & \|y_n - Ty_n\|^2 \\ \text{s.t.} & x_k,\ y_k \text{ are generated as inertial KM iterations,} \\ & T \text{ is a nonexpansive mapping,} \\ & T \text{ admits a fixed point } x^* \end{cases}$$

[4] Postdoctoral Researcher, Universidad Católica del Norte, Antofagasta, Chile.

for a fixed $n \in \mathbb{N}$, and vectors $x_k, y_k \in \mathbb{R}^d$, for $k = 0, \ldots, n$. In the following, we state the objective function and each one of the constraints in a convenient form to tackle the problem using a numerical solver. The constraint of x_k and y_k being generated as KM iterations will be used to rewrite the other expressions only in terms of x_k.

- **Objective function.** Using (3.54), observe that
$$\begin{aligned}\|y_n - Ty_n\|^2 &= \left\| \frac{1}{\lambda}(y_n - x_{n+1}) \right\|^2 \\ &= \left\| \frac{1}{\lambda} x_{n+1} - \frac{(\alpha+1)}{\lambda} x_n + \frac{\alpha}{\lambda} x_{n-1} \right\|^2\end{aligned}$$

- **Constraint 1: T must be nonexpansive.** That is, for each $k, j = 1, \ldots, n$, with $k \neq j$,
$$\|Ty_k - Ty_j\|^2 \leq \|y_k - y_j\|^2 .$$
Notice that (3.54) allows to rewrite the previous expression only in terms of x_k, x_j, using
$$Ty_k = \frac{x_{k+1}}{\lambda} - \frac{(1-\lambda)}{\lambda}(1+\alpha)x_k + \frac{\alpha}{\lambda}(1-\lambda)x_{k-1}, \tag{3.56}$$
and
$$y_k = x_k(1+\alpha) - \alpha x_{k-1}. \tag{3.57}$$

- **Constraint 2: T admits a fixed point x^*.** We will generate a sequence x_k, $k = 0, \ldots, n+1$ and an extra point, x_{n+2} that will serve as fixed point. Then, the fixed-point condition will be modeled as follows: for every $k = 1, \ldots, n$,
$$\|Ty_k - x_{n+2}\|^2 \leq \|y_k - x_{n+2}\|^2 .$$
Using (3.56) and (3.57), this expression can be rewritten in terms of x_k.

Using the notation $\nu = (1-\lambda)/\lambda$, the problem can be formulated as follows:

$$\begin{cases} \max\limits_{x_0, x_1, \ldots, x_{n+1}} \left\| \frac{1}{\lambda} x_{n+1} - \frac{(\alpha+1)}{\lambda} x_n + \frac{\alpha}{\lambda} x_{n-1} \right\|^2 \\ \qquad \text{s.t. } \left\| \frac{x_{k+1}}{\lambda} - \nu(1+\alpha)x_k + \alpha\nu x_{k-1} - \frac{x_{j+1}}{\lambda} + \nu(1+\alpha)x_j - \alpha\nu x_{j-1} \right\|^2 \\ \qquad \leq \|x_k(1+\alpha) - \alpha x_{k-1} - x_j(1+\alpha) + \alpha x_{j-1}\|^2 , \\ \qquad\quad \forall\, k,\, j = 1 \ldots n,\, k \neq j, \\ \qquad \left\| \frac{x_{k+1}}{\lambda} - \nu(1+\alpha)x_k + \alpha\nu x_{k-1} - x_{n+2} \right\|^2 \\ \qquad \leq \|x_k(1+\alpha) - \alpha x_{k-1} - x_{n+2}\|^2 , \forall\, k, j = 1, \ldots, n,\ k \neq j, \\ \qquad \|x_0 - x_*\|^2 \leq 1, \\ \qquad \|x_1 - x_*\|^2 \leq 1. \end{cases} \tag{3.58}$$

As it was mentioned in the introductory PEP (3.51), a constrain of boundedness on the fixed point is added. The optimization problem consists in an objective function and constraints written as squared norms depending only on the sequence x_k and the parameters α, λ. Then, the squared norms can be expressed as inner products, and the problem can be formulated in terms of $\langle x_k, x_j\rangle$, with k, $j = 0, \dots, n+2$. Considering this formulation, we can set as the variables of the optimization problem the value of each inner product $\langle x_k, x_j\rangle$. That is, the variables can be represented in a lower triangular matrix

$$\mathbf{X} = \begin{pmatrix} \langle x_0, x_0\rangle & \dots & 0 \\ \vdots & \ddots & \vdots \\ \langle x_0, x_{n+2}\rangle & & \langle x_{n+2}, x_{n+2}\rangle \end{pmatrix}.$$

To write the objective function, let us define a cost vector

$$c = \left[\lambda^{-1}, -(\alpha+1)\lambda^{-1}, \alpha\lambda^{-1}\right],$$

where c_i, $i \in \{1,2,3\}$ states for each one of the components. Then, let us define the lower triangular matrix $C \in \mathbb{M}_{n+3}(\mathbb{R})$ as

$$C_{ij} = \begin{cases} c_i c_j & \text{for}\, i,\ j \in \{n-1, n, n+1\},\ i \geq j, \\ 0 & \text{in any other case.} \end{cases}$$

Therefore, the objective function of the problem (3.58) can be written as $\mathbf{tr}(C\mathbf{X})$, where $\mathbf{tr}$ stands for the trace operator. Notice that the problem has n^2+2 constraints, each of them can be written as $\mathbf{tr}(A_i\mathbf{X}) \leq b_i$, $i = 1, \dots, n^2+2$. Then, for a fixed value of $n \in \mathbb{N}$, problem (3.58) reduces to

$$\begin{cases} \max\limits_{\mathbf{X}\in\mathcal{L}_{n+3}(\mathbb{R})} & \mathbf{tr}(C\mathbf{X}) \\ \text{s.t.} & \mathbf{tr}(A_i\mathbf{X}) \leq b_i,\ i = 1, \dots, n^2+2, \end{cases} \tag{3.59}$$

where $\mathcal{L}_{n+3}(\mathbb{R})$ stands for the set of the lower triangular matrices of size $n+3$. Problem (3.59) is a semi definite program (SDP), and we will study the solutions using the optimization solver `cvxpy` in `python`.

3.6.2 The non-inertial case

First we solve (3.58) in the non-inertial setting, that is, $\alpha = 0$, for $n = 1\dots, N$. Figure 3.11 shows the optimal value of the problem for every n, with $N = 25$, starting from $n = 3$ to avoid an erratic behavior in the first iterations. The simulation shows that the best rate of convergence is achieved in the case $\lambda = 1/2$, which is consistent with the best value of λ on the rate (3.53). The solution obtained by PEP matches Lieder's tight rate (3.53) Moreover, which is depicted in Figure 3.12 for $\lambda = 1/2$. The results obtained by the previous model align with established theoretical findings, offering a motivation to conjecture about unknown rates, particularly the case of inertial iterations.

3.6.3 The inertial case

For the inertial KM iterations, currently there are no examples of tight bounds for the residuals. Then, the numerical experiments described in the following hopefully will provide

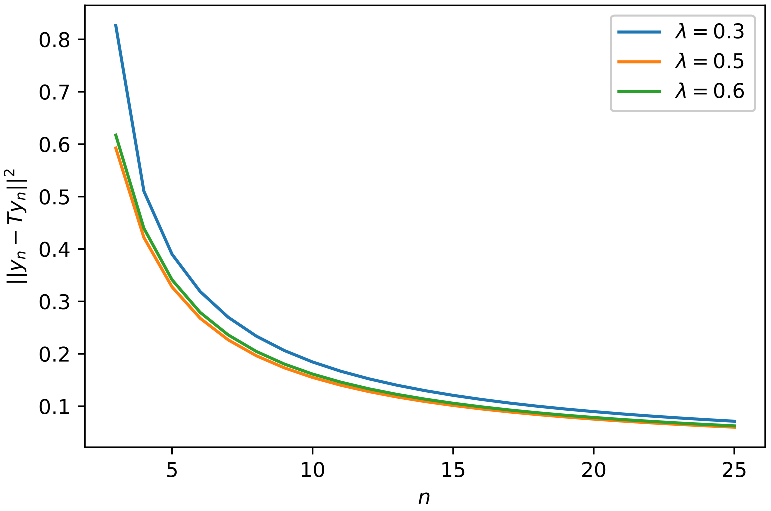


Figure 3.11: Solution of the PEP (3.58), non-inertial case.

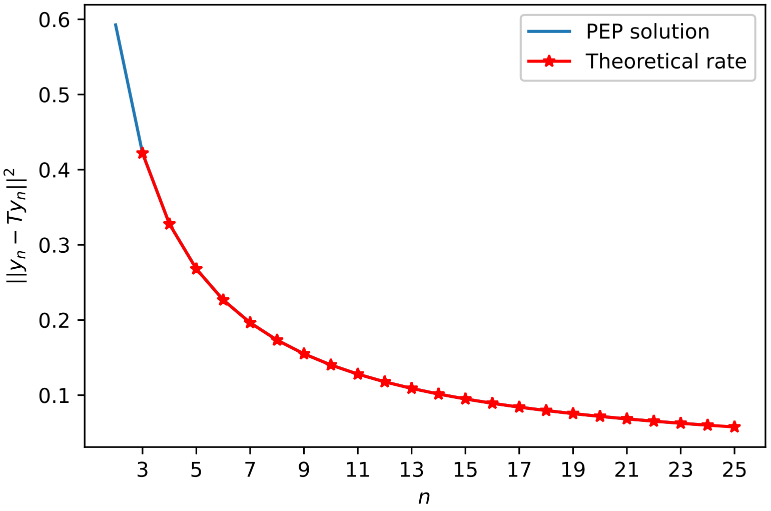


Figure 3.12: Comparison of PEP solution, $\lambda = 1/2$ and rate (3.53).

a better insight to obtain theoretical results in an upcoming research. First, an overall result: notice that condition (3.55) in the case $\lambda = 1/2$ gives $\alpha < 1/3$. Figure 3.13 shows the results of the experiment in the case $\lambda = 1/2$ and feasible values of $\alpha < 1/3$. Here, the best case is achieved using $\alpha = 0$, that is, not using inertia. Let us recall that in the strict contractive case, the rate (3.28) is provided for the iterations, and although there is no information whether its tightness, it also suggests not implementing inertia. Then, both results leads us to conjecture that KM iterations are a extremely general case for the use of inertia.

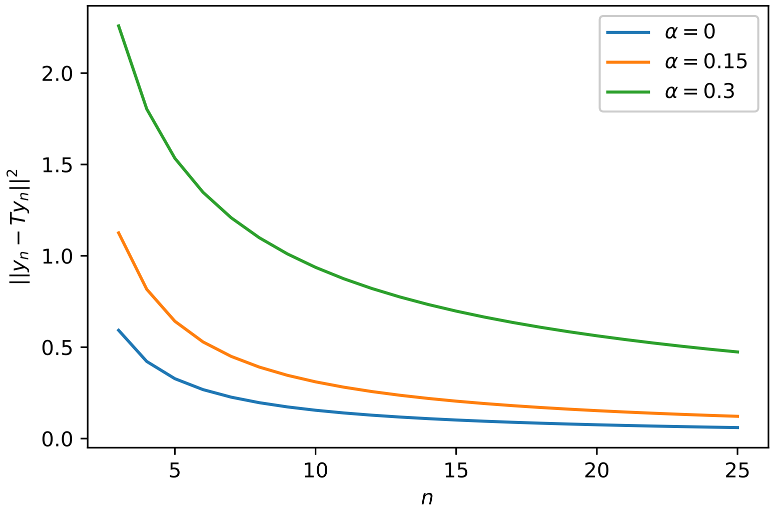


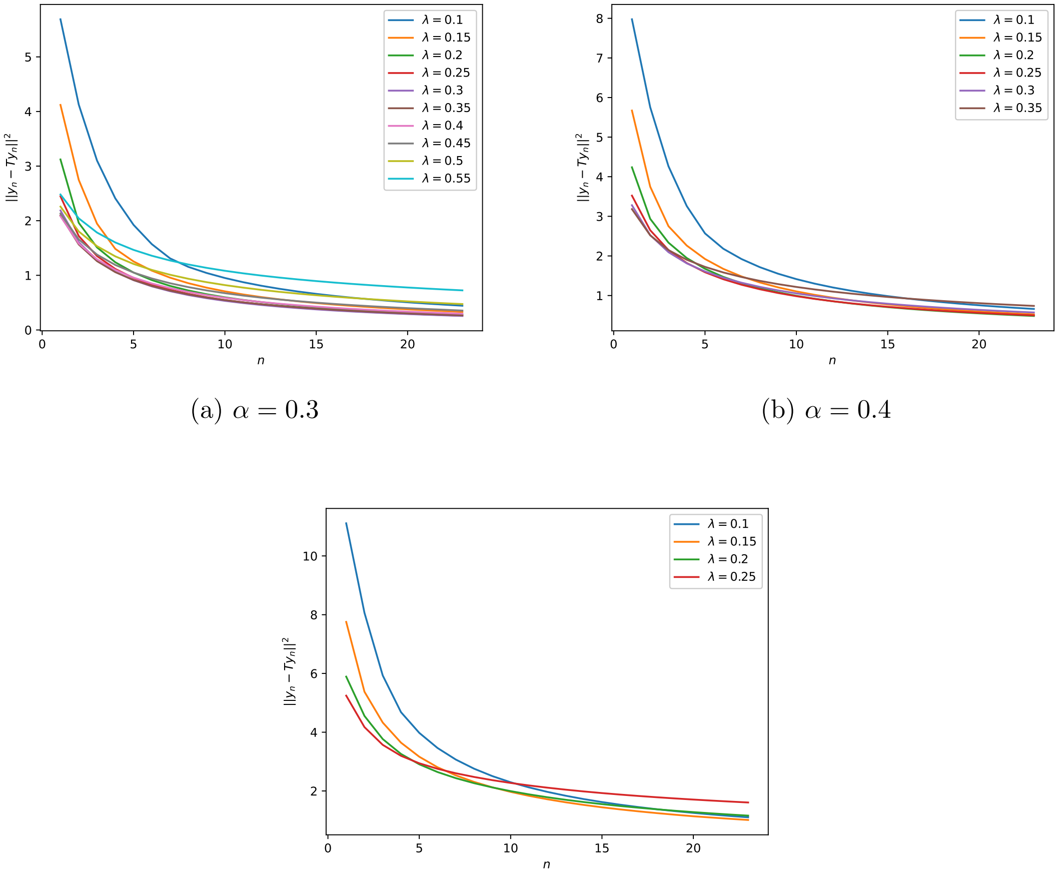
Figure 3.13: PEP solution, $\lambda = 1/2$ and different inertial values.

Nevertheless, the study of inertia under this numerical experiment may be revealing for future research. Then, the next experiment considers solutions of the PEP only in the inertial setting. For a given value of $\alpha \in (0, 1)$, we must pick $\lambda \in (0, \lambda(\alpha))$, with

$$\lambda(\alpha) = \frac{(1-\alpha)^2}{(1-\alpha+2\alpha^2)},$$

to fulfill condition (3.55). The function $\lambda(\alpha)$ is decreasing on $[0, 1]$, with $\lambda(0) = 1$ and $\lambda(1) = 0$, which implies that as larger the value of α is, the smaller is the interval of feasible values of λ. The experiment will be performed as follows: for different values of $\alpha \in (0, 1)$, problem (3.58) will be solved using several values of $\lambda \in (0, \lambda(\alpha))$. Figure 3.14 shows the solutions of (3.58) for different combinations of α, λ.

Considering the results of the previous experiment, a natural question is given a value of α, what is the best option of λ in the feasible set to achieve the best rate of convergence? As a criteria for picking the best value of λ, we will choose the one that attains the minimum value at the last iteration. For a fixed value of α, let us call $\lambda_* = \lambda(\alpha)$ using (3.55), and then we will pick the best λ on the feasible set $(0, \lambda_*)$. Figure 3.15 shows the results obtained and they are detailed on Table 3.6. There, it can be seen that the ratio $\lambda_{\text{best}}/\lambda_*$ remains constant equal to 1/2. Thus, this experiment lead us to the following conjecture: given an inertial coefficient α, the best value λ to choose is the middle point of the interval $(0, \lambda_*)$. Notice that this result is consistent for the best value of λ in the theoretical rate (3.53).



(a) $\alpha = 0.3$

(b) $\alpha = 0.4$

(c) $\alpha = 0.5$

Figure 3.14: PEP solution in the inertial case.

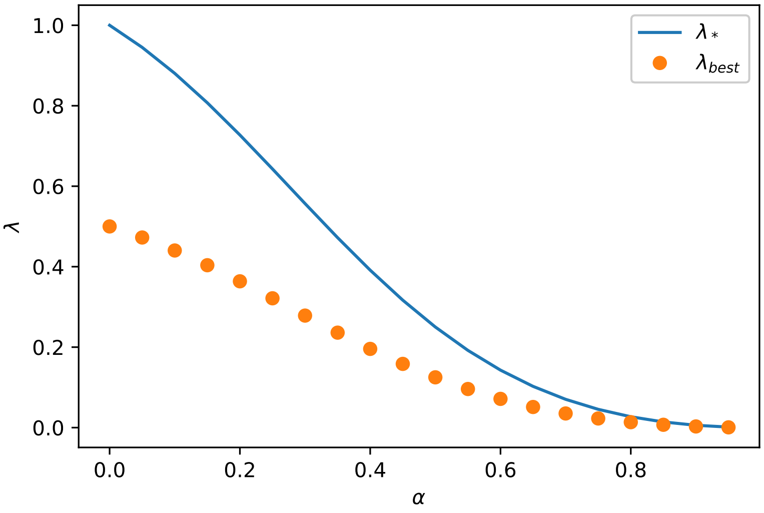


Figure 3.15: Best value of λ for every value of $\alpha \in [0, 1]$.

α	λ_*	λ_{best}	$\lambda_{\text{best}}/\lambda_*$
0	1.0000	0.5000	0.5
0.05	0.9450	0.4725	0.5
0.10	0.8804	0.4402	0.5
0.15	0.8073	0.4036	0.5
0.20	0.7273	0.3636	0.5
0.25	0.6429	0.3214	0.5
0.30	0.5568	0.2784	0.5
0.35	0.4721	0.2360	0.5
0.40	0.3913	0.1957	0.5
0.45	0.3168	0.1584	0.5
0.50	0.2500	0.1250	0.5
0.55	0.1919	0.0960	0.5
0.60	0.1429	0.0714	0.5
0.65	0.1025	0.0513	0.5
0.70	0.0703	0.0352	0.5
0.75	0.0455	0.0227	0.5
0.80	0.0270	0.0135	0.5
0.85	0.0141	0.0071	0.5
0.90	0.0058	0.0029	0.5
0.95	0.0013	0.0007	0.5

Table 3.6: Best value of λ for every value of $\alpha \in [0, 1)$.

3.6.4 Discussion

This chapter presents important results for the convergence of inertial Krasnosleskii-Mann iterations. In the strict contractive case, a rate of linear convergence is provided. While this rate may imply that the optimal performance of the iterations is achieved by not implementing inertia, the numerical instances tested show the opposite: both the primal-dual splitting algorithm and the Davis and Yin method boost their convergence when inertia is used.

A natural question that arises from this part of the research is to explain this apparent contradiction. From our point of view, two key aspects require consideration. First, the tightness of the rate found. Secondly, the class of operators studied. It is worth noting that although the linear convergence is guaranteed for any contractive operator, our tests are performed over two algorithms defined by firmly nonexpansive operators. Then, a possible answer to the discrepancy between theoretical and practical behavior, is that maybe we are studying a broader class of operators than necessary.

The PEP presented before matches the theoretical rate for residuals in the noninertial case, suggesting that is a reliable rate estimating tool. However, also gives a slower convergence by implementing inertia. This lead us to conjecture that, in effect, we may be studying inertia in a too large context. This motivate us to focus in the study of inertia in maybe averaged or firmly nonexpansive operators.

Chapter 4

Conclusions and perspectives

The content of this Thesis summarizes the work carried out since the second half of 2020, divided into two main themes, which led to the elaboration of two scientific articles, and some preliminary results of an ongoing research.

First part: Restart

The first part is devoted to propose a restart scheme for a continuous dynamics involving the Hessian of a convex function, which can be seen as an extension of the continuous restart scheme setting for Nesterov's accelerated method, proposed by Su, Boyd and Candès on 2014. Our contribution is to provide a convergence result for the restart scheme, along with a rate of convergence, and an existence theorem for the solutions of the dynamics. Numerical simulations are provided for the restarted trajectories, showing the acceleration and the mitigation of the oscillations of the values of the function towards its minimizer. Also, a first glance to the algorithmic consequences is studied, applying the restart scheme to first order algorithms with a gradient correction term. This research led us to the following questions that hopefully will be addressed in a future work:

- In practice, it can be seen that the bound found for the restart scheme is not tight, that is, the function values converge faster. Although it is the same problem exhibited by the bound found by Su, Boyd and Candès, we conjecture that it can be improved by using a different approach on the upper bound of the restart time. Also, it can be observed that in practice, the speed restart time is small, leading to an almost immediate restart. By the time of writing of this thesis, a new speed-restart criteria is being studied, aiming to restart after the speed-criteria presented here.
- For discrete inertial settings, just as Nesterov's algorithm, a fixed restart rate of convergence is easy to compute using the linear convergence rates in the strongly convex case. There are no existing results of linear rates for algorithms with gradient correction at the time of writing this thesis. One interesting perspective is to study rates for this kind of algorithm in the strongly convex case.
- For the discrete setting, that is, the algorithms with a gradient correction term, we aim to establish a convergence result by implementing a restart routine. This is not straightforward, as there are not even rates for Nesterov's algorithm using the discrete

speed criteria showed in Algorithm 1. Existing literature primarily provides convergence rates for restart schemes for Nesterov's Algorithm or FISTA, where the restart criteria typically depends on the strong convexity parameter (see for example [69, 68, 89, 66, 44]). As this parameter in practice is not easy to know beforehand, there are works on restarted schemes that use approximations of it to define the restart [16, 55]. For automatic restart criteria, we refer to the work of Alamo et al., [3, 2], where linear convergence is achieved using criteria defined by the reduction of the function values or the composite gradient mapping values. An interesting challenge is how to implement some of these methods in algorithms with gradient corrections.

- The results of the convergence for the discrete algorithms, such as the ones presented on [10, 11] ask to the value of β to be significantly small, leading the algorithm to approach Nesterov's method. In practice, it can be noticed an improvement on the velocity of the convergence for a larger range of values of β. Thus, we presume that the feasible interval should be larger, or another kind of algorithms derived from the second order dynamic will allow more freedom for β.

Second part: Inertial KM iterations

The second part of the Thesis focuses on the study of the inclusion of an inertial term on KM iterations. Although there are several related works on this subject, our contribution is to provide a unified proof for the weak convergence using a general pair of sequences α_k and λ_k, along with a new strong convergence result. Also, we provide two inertial settings for existing fixed point algorithms that outperforms the original versions. The study of this topic raised the following questions:

- The given bound for the strong convergence implies that the optimal results are obtained not using inertia. Also, the PEP results support this hypothesis. In contrast, the numerical simulations show that the inclusion of inertia can accelerate the convergence, both in time and iterations. This makes us conjecture that we are addressing the problem of the inclusion of the inertia in a very broad class of iterations. By considering a wide class of operators, for example we are dealing with isometries, such as rotations, that usually do not show fast convergence properties. Thus, we should focus on a more restricted class of operators, or particular instances of the nonexpansive ones, such as firmly nonexpansive or averaged operators.
- If the rate provided is close to reality, that is, if there exists some kind of nonexpansive operator for which inertial KM iterations converges slower than the non-inertial, then we should be able to find an example that will provide us a lower bound on the rate of convergence.
- Numerical illustrations show an interesting behavior in the case of the over-relaxation for averaged operators, that is, to use a sequence λ_k (or a fixed value of λ for practical purposes) which is greater than one. We aim to provide a more analytical study of this phenomenon.

Bibliography

[1] Samir Adly, Hedy Attouch, and Van Nam Vo. Asymptotic behavior of Newton-like inertial dynamics involving the sum of potential and nonpotential terms. *Fixed Point Theory and Algorithms for Sciences and Engineering*, 2021(1):17, 2021.

[2] Teodoro Alamo, Pablo Krupa, and Daniel Limon. Gradient based restart fista. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3936–3941. IEEE, 2019.

[3] Teodoro Alamo, Daniel Limon, and Pablo Krupa. Restart FISTA with global linear convergence. In *2019 18th European Control Conference (ECC)*, pages 1969–1974. IEEE, 2019.

[4] Felipe Álvarez. On the minimizing property of a second order dissipative system in Hilbert spaces. *SIAM Journal on Control and Optimization*, 38, 2001.

[5] Felipe Álvarez and Hedy Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Analysis*, 9(1):3–11, 2001.

[6] Felipe Álvarez and J.M. Pérez. A dynamical system associated with Newton's method for parametric approximations of convex minimization problems. *Applied Mathematics and Optimization*, 38(2):193–217, 1998.

[7] Francisco J Aragón-Artacho and David Torregrosa-Belén. A direct proof of convergence of Davis–Yin splitting algorithm allowing larger stepsizes. *Set-Valued and Variational Analysis*, 30(3):1011–1029, 2022.

[8] Hedy Attouch, Luis M. Briceno-Arias, and Patrick L. Combettes. A parallel splitting method for coupled monotone inclusions. *SIAM Journal on Control and Optimization*, 48(5):3246–3270, 2010.

[9] Hedy Attouch and Alexandre Cabot. Convergence of a relaxed inertial forward–backward algorithm for structured monotone inclusions. *Applied Mathematics & Optimization*, 80(3):547–598, 2019.

[10] Hedy Attouch, Zaki Chbani, Jalal Fadili, and Hassan Riahi. First-order optimization algorithms via inertial systems with Hessian driven damping. *Mathematical Programming*, pages 1–43, 2020.

[11] Hedy Attouch, Zaki Chbani, Jalal Fadili, and Hassan Riahi. Convergence of iterates for first-order optimization algorithms with inertia and Hessian driven damping. *Optimization*, 72(5):1199–1238, 2023.

[12] Hedy Attouch, Zaki Chbani, Juan Peypouquet, and Patrick Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168:123–175, 2018.

[13] Hedy Attouch and Juan Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

[14] Hedy Attouch and Juan Peypouquet. Convergence of inertial dynamics and proximal algorithms governed by maximally monotone operators. *Mathematical Programming*, 174:391–432, 2019.

[15] Hedy Attouch, Juan Peypouquet, and Patrick Redont. Fast convex optimization via inertial dynamics with Hessian driven damping. *Journal of Differential Equations*, 261(10):5734–5783, 2016.

[16] Jean-François Aujol, Charles H. Dossal, Hippolyte Labarrière, and Aude Rondepierre. FISTA restart using an automatic estimation of the growth parameter. HAL-03153525v4, 2022.

[17] Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta mathematicae*, 3(1):133–181, 1922.

[18] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham, second edition, 2017.

[19] Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[20] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[21] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[22] Radu Ioan Boţ, Ernö Robert Csetnek, and Christopher Hendrich. Inertial Douglas–Rachford splitting for monotone inclusion problems. *Applied Mathematics and Computation*, 256:472–487, 2015.

[23] Haim Brézis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York, NY, 2011.

[24] Haim Brézis and Pierre Louis Lions. Produits infinis de résolvantes. *Israel Journal of Mathematics*, 29(4):329–345, 1978.

[25] Luis Briceño-Arias and Fernando Roldán. Primal-dual splittings as fixed point iterations in the range of linear operators. *Journal of Global Optimization*, 2022.

[26] Luis M. Briceño-Arias. Forward-Douglas–Rachford splitting and forward-partial inverse method for solving monotone inclusions. *Optimization*, 64(5):1239–1261, 2015.

[27] Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

[28] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.

[29] Antonin Chambolle and Charles Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.

[30] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[31] Patrick L. Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5-6):475–504, 2004.

[32] Patrick L Combettes and Lilian E Glaudin. Quasi-nonexpansive iterations on the affine hull of orbits: from mann's mean value algorithm to inertial methods. *SIAM Journal on Optimization*, 27(4):2356–2380, 2017.

[33] Patrick L. Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale modeling & simulation*, 4(4):1168–1200, 2005.

[34] Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of optimization theory and applications*, 158(2):460–479, 2013.

[35] Minh N Dao and Hung M Phan. An adaptive splitting algorithm for the sum of two generalized monotone operators and one cocoercive operator. *Fixed Point Theory and Algorithms for Sciences and Engineering*, 2021(1):1–19, 2021.

[36] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

[37] Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-valued and variational analysis*, 25(4):829–858, 2017.

[38] Qiao-Li Dong, Yeol Je Cho, Songnian He, Panos M. Pardalos, and Themistocles M. Rassias. *The Krasnosel'skii-Mann iterative method: recent progress and applications.* Springer, 2022.

[39] Qiao-Li Dong, Yeol Je Cho, and Themistocles M. Rassias. General inertial Mann algorithms and their convergence analysis for nonexpansive mappings. *Applications of Nonlinear Analysis*, pages 175–191, 2018.

[40] Qiao-Li Dong and Han-bo Yuan. Accelerated Mann and CQ algorithms for finding a fixed point of a nonexpansive mapping. *Fixed Point Theory and Applications*, 2015(1):1–12, 2015.

[41] Yunda Dong. New inertial factors of the Krasnoselskii-Mann iteration. *Set-valued and variational analysis*, 29:145–161, 2021.

[42] Jim Douglas, Jr. and H. H. Rachford, Jr. On the numerical solution of heat conduction problems in two and three space variables. *Trans. Amer. Math. Soc.*, 82:421–439, 1956.

[43] Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

[44] Olivier Fercoq and Zheng Qu. Restarting accelerated gradient methods with a rough strong convexity estimate. *arXiv preprint arXiv:1609.07358*, 2016.

[45] Ignacio Fierro, Juan José Maulén, and Juan Peypouquet. Inertial Krasnoselskii-Mann iterations. *arXiv preprint arXiv:2210.03791*, 2022.

[46] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & mathematics with applications*, 2(1):17–40, 1976.

[47] Pontus Giselsson and Stephen Boyd. Monotonicity and restart in fast gradient methods. In *53rd IEEE Conference on Decision and Control*, pages 5058–5063. IEEE, 2014.

[48] Alan A. Goldstein. Convex programming in Hilbert space. *Bulletin of the American Mathematical Society*, 70(5):709–710, 1964.

[49] Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM journal on control and optimization*, 29(2):403–419, 1991.

[50] Per Christian Hansen, James G. Nagy, and Dianne P. O'Leary. *Deblurring images: Matrices, spectra, and filtering*, volume 3 of *Fundamentals of Algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.

[51] Olaniyi S. Iyiola and Yekini Shehu. New convergence results for inertial Krasnoselskii–Mann iterations in Hilbert spaces with applications. *Results in Mathematics*, 76(2):1–25, 2021.

[52] Mark Aleksandrovich Krasnosel'skii. Two comments on the method of successive approximations. *Usp. Math. Nauk*, 10:123–127, 1955.

[53] Evgeny S. Levitin and Boris T. Polyak. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 6(5):1–50, 1966.

[54] Felix Lieder. *Projection based methods for conic linear programming—optimal first order complexities and norm constrained quasi newton methods*. PhD thesis, Heinrich-Heine-Universität Düsseldorf, 2018.

[55] Qihang Lin and Lin Xiao. An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. In *International Conference on Machine Learning*, pages 73–81. PMLR, 2014.

[56] Qihang Lin and Lin Xiao. An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. *Computational Optimization and Applications*, 60(3):633–674, 2015.

[57] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[58] Dirk A. Lorenz and Thomas Pock. An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2):311–325, 2014.

[59] Paul-Emile Maingé. Convergence theorems for inertial KM-type algorithms. *Journal of Computational and Applied Mathematics*, 219(1):223–236, 2008.

[60] W. Robert Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4(3):506–510, 1953.

[61] Bernard Martinet. Regularisation, d'inéquations variationelles par approximations succesives. *Revue Française d'informatique et de Recherche operationelle*, 1970.

[62] Juan José Maulén and Juan Peypouquet. A speed restart scheme for a dynamics with Hessian driven damping. *Journal of Optimization Theory and Applications*, 2023.

[63] Ramzi May. Asymptotic for a second-order evolution equation with convex potential andvanishing damping term. *Turkish Journal of Mathematics*, 41(3):681–685, 2017.

[64] Céline Moucer, Adrien Taylor, and Francis Bach. A systematic approach to Lyapunov analyses of continuous-time models in convex optimization. *SIAM Journal on Optimization*, 33(3):1558–1586, 2023.

[65] Abdellatif Moudafi. A reflected inertial Krasnoselskii-type algorithm for Lipschitz pseudo-contractive mappings. *Bulletin of the Iranian Mathematical Society*, 44:1109–1115, 2018.

[66] Ion Necoara, Yurii Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

[67] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

[68] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[69] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

[70] Yurii Nesterov. *Lectures on convex optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, 2018.

[71] Brendan O'Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

[72] Zdzisław Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591–597, 1967.

[73] Gregory B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979.

[74] Juan Peypouquet. *Convex optimization in normed spaces*. SpringerBriefs in Optimization. Springer, Cham, 2015.

[75] Émile Picard. Memoire sur la theorie des equations aux derivees partielles et la methode des approximations successives. *Journal de Mathématiques pures et appliquées*, 6:145–210, 1890.

[76] Boris Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4:1–17, 12 1964.

[77] Nelly Pustelnik. *Méthodes proximales pour la résolution de problèmes inverses: application à la tomographie par émission de positrons*. PhD thesis, Université Paris-Est, 2010.

[78] Simeon Reich. Weak convergence theorems for nonexpansive mappings in Banach spaces. *Journal of Mathematical Analysis and Applications*, 67(2):274–276, 1979.

[79] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[80] Ernest K Ryu, Adrien B Taylor, Carolina Bergeling, and Pontus Giselsson. Operator splitting performance estimation: tight contraction factors and optimal parameter selection. *SIAM Journal on Optimization*, 30(3):2251–2271, 2020.

[81] Yekini Shehu. Convergence rate analysis of inertial Krasnoselskii–Mann type iteration with applications. *Numerical Functional Analysis and Optimization*, 39(10):1077–1091, 2018.

[82] Yekini Shehu, Aviv Gibali, and Simone Sagratella. Inertial projection-type methods for solving quasi-variational inequalities in real hilbert spaces. *Journal of Optimization Theory and Applications*, 184(3):877–894, 2020.

[83] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[84] Adrien Taylor, Bryan Van Scoy, and Laurent Lessard. Lyapunov functions for first-order methods: Tight automated convergence guarantees. In *International Conference on Machine Learning*, pages 4897–4906. PMLR, 2018.

[85] Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3):1283–1313, 2017.

[86] Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2017.

[87] Manu Upadhyaya, Sebastian Banert, Adrien B Taylor, and Pontus Giselsson. Automated tight Lyapunov analysis for first-order methods. *arXiv preprint arXiv:2302.06713*, 2023.

[88] Bang Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

[89] Hui Zhang and Lizhi Cheng. Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization. *Optimization Letters*, 9:961–979, 2015.