Lagrangian Methods for Composite Optimization

Shoham Sabach* Marc Teboulle†

April 7, 2019

Abstract

Lagrangian based methods have been on the market for over 50 years. These methods are robust and often can handle optimization problems with complex geometries through efficient computational steps. The last decade of research have generated a large volume of literature on various practical and theoretical aspects of many Lagrangian based algorithms. This chapter reviews the basic elements of Lagrangian based methods for composite minimization in the convex and nonconvex setting. In the convex case, the focus is on global rate of convergence results, which are derived here through a novel approach and very simple proof technique. In the much harder nonconvex case, we survey a very recent methodology which allows to establish global pointwise convergence results for a broad class of genuine nonlinear composite semi-algebraic problems.

2010 Mathematics Subject Classification: 90C25, 65K05.
Keywords: Lagrangian multiplier methods, proximal multiplier algorithms, convex and nonconvex composite minimization, alternating minimization, decomposition schemes, global rate of convergence analysis, Kurdyka-Łosiajewicz property, semi-algebraic optimization, global pointwise convergence.

1 Introduction

Recently, we are witnessing a resurgence of Lagrangian based methods and their related decomposition schemes. Augmented Lagrangian methods, also known as multiplier methods, are probably the first Lagrangian based methods developed for optimization. These were introduced about half a century ago by Hestenes [33] and Powell [49] for solving nonlinear optimization problems with nonlinear equality constraints. Some earlier works can be traced back to the late fifties with the volume of Arrow, Hurwicz and Uzawa [2] which have launched the fundamental idea of decomposition schemes (also known as splitting methods), which are intimately related to Lagrangian schemes. This renewed interest in Lagrangian methods has emerged from new applications arising in many fundamental scientific and engineering problems in image science, machine learning, and other fields, which are often modeled by structured and very large scale minimization problems, often nonsmooth, and even nonconvex. These methods have gained popularity due to their ability to adapt to problem structures and to exploit data information, to produce scalable and tractable updating steps. A voluminous recent literature reflects this intense renewal interest on multiplier methods and their relatives decomposition algorithms. It is not our intention, neither is the scope of this review chapter, to discuss all past and more recent literature. Instead, throughout the

*Faculty of Industrial Engineering and Management, The Technion, Haifa, 32000, Israel. E-mail: ssabach@ie.technion.ac.il.

†School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel. E-mail: teboulle@post.tau.ac.il.
This research was partially supported by the Israel Science Foundation, under ISF Grant 1844-16

chapter we provide some pointers from earlier foundational works, to some recent advances which have evolved over the last 50 years.

A classical reference providing the central developments and ideas underlying multiplier methods since their inception is the monograph of Bertsekas [8], while decomposition methods can be found in the books of Lasdon [38], and of Bertsekas and Tsitsiklis [11] for a more modern treatment. Decomposition methods can also be developed via the abstract setting of maximal monotone operators [52, 53, 48, 41]. We refer the reader to the recent monograph of Bauschke and Combettes [5], which covers fundamental results in this direction, and also includes an extensive bibliography.

In this survey, we present the essential tools needed to build and analyze Lagrangian based methods for a composite optimization model which is general enough to capture most instances of fundamental modern applications both in the convex and nonconvex settings. In the convex case, the focus is on global rate of convergence results, derived here through a novel approach leading to a simplified proof technique permitting unification. In the much harder nonconvex case, we review a very recent methodology for global pointwise convergence analysis to a broad class of genuine nonlinear composite semi-algebraic problems. This is developed in three main sections, whose contents are now briefly outlined in some more detail.

In section 2, we start with the main model of the chapter, which is a nonlinear composite minimization, and describe in informal ways the nature and basic steps in constructing Lagrangian based methods through the penalty and proximal regularization concepts. This has the advantage of explaining in simple terms, and with minimal sophistication, the basic elements underlying Lagrangian based schemes in the absence of convexity for a very general composite optimization model, and to reveal the difficulties associated with Lagrangian based methods (which will be further elaborated more precisely later on in section 4). We also introduce some basic results in the convex setting and the starring role of the proximal maps in Lagrangian based schemes. Finally, we end the section with a sample of application examples, which illustrate the flexibility of the composite model, many other applications can be found in the cited literature.

In section 3, we focus on the convex setting. When convexity is present in a given optimization model, it is well-known that an augmented Lagrangian method is a *dual* method. More precisely, it is a manifestation of the proximal minimization method applied to the dual formulation of the given optimization problem. Starting with the pioneering work of Moreau [47], followed by the proximal minimization algorithm of Martinet [44], this decisive dual approach interpretation of multiplier methods was developed by Rockafellar, see e.g., [53, 52]. Building on the more general Proximal Method of Multipliers initially developed by Rockafellar [52], we present a new and simplified approach to analyze Lagrangian based schemes for the convex composite model. This novel approach relies on a simple observation, which allows to show that all well-known Lagrangian methods and their decomposition variants can be captured and analyzed through a *single* scheme called the *perturbed* proximal method of multipliers. This allows us to derive a single unifying global rate of convergence result through an elementary proof, see Theorem 3.1, which as we show, can then be easily applied to obtain rate of convergence results for various fundamental decomposition schemes.

In section 4, we move on to the much harder and challenging nonconvex setting. In sharp contrast to the convex case, where Lagrangian based methods and their convergence properties have been extensively studied in past and current literature, in the nonconvex setting, global convergence analysis for general nonlinear composite models has remained scarce, and far from being well-understood. Only very recently some progress has been initiated in the nonconvex case, but mainly for the *linear* composite model, see e.g., [40] and references therein. The results we

reviewed in this section are in fact very recent, and were developed in the just published work of Bolte, Sabach and Teboulle [15], which to the best of our knowledge, appears to be the first work capable of handling general nonconvex and nonlinear composite minimization, and producing the basis of a theoretical framework to derive global convergence analysis of Lagrangian based methods for the broad class of semi-algebraic problems.

The definitions and notations used in this chapter, in relation to convex and variational analysis, are classical and can be found in the book of Rockafellar [51] and the monograph of Rockafellar and Wets [54].

2 The Lagrangian Framework

Consider the following general composite optimization model

$$\text{(M)} \qquad \min_{u\in\mathbb{R}^d} \left\{\Phi(u) := \varphi(u) + h(F(u))\right\}, \tag{2.1}$$

where $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$ and $h : \mathbb{R}^m \to (-\infty, +\infty]$ are proper and lower semi-continuous functions, while $F : \mathbb{R}^d \to \mathbb{R}^m$ is a C^1 mapping.

The choice of the model (M) is for the sake of convenience to clearly delineate the role played by additional structures and properties in every part of the triplet $[\varphi, h, F]$ in the design and analysis of various Lagrangian based algorithms that will be described and analyzed in the following parts of this chapter. In section 2.3, we illustrate the flexibility of model (M) which can be used to describe a wide spectrum of disparate applications. In the following two subsections we first describe in informal ways the underlying elements and basic mechanism of every Lagrangian based method for tackling the general model (M), which is then followed by the fundamental proximal framework.

2.1 Lagrangian Based Methods: Basic Elements and Mechanism

We start with an elementary and useful equivalent reformulation of the optimization model (M), which reads as

$$\text{(M)} \qquad \min_{u\in\mathbb{R}^d, v\in\mathbb{R}^m} \left\{\varphi(u) + h(v) :\ F(u) = v\right\}.$$

Attaching a multiplier $y \in \mathbb{R}^m$ to the equality constraint, we define the *Lagrangian* $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^m \to (-\infty, +\infty]$, which associated to problem (M), via

$$\mathcal{L}(u, v, y) \equiv \varphi(u) + h(v) + \langle y, F(u) - v\rangle, \tag{2.2}$$

and for any $\rho > 0$ the corresponding *augmented Lagrangian* $\mathcal{L}_\rho : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^m \to (-\infty, +\infty]$ is defined by

$$\begin{aligned}\mathcal{L}_\rho(u, v, y) &:= \mathcal{L}(u, v, y) + \frac{\rho}{2}\|F(u) - v\|^2 \\ &= \varphi(u) + h(v) + \langle y, F(u) - v\rangle + \frac{\rho}{2}\|F(u) - v\|^2,\end{aligned} \tag{2.3}$$

namely, the *penalized* Lagrangian $\mathcal{L}(\cdot) \equiv \mathcal{L}_0(\cdot)$.

Equipped with $\mathcal{L}_\rho(\cdot)$, a basic Lagrangian scheme starts with any given triplet (u, v, y), and consists of minimizing the augmented Lagrangian $\mathcal{L}_\rho(\cdot)$ with respect to (u, v) followed by an update of the corresponding multiplier y, and reads as follows

Augmented Lagrangian Scheme (ALS)

1. **Initialization:** Start with any $(u, v, y) \in \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^m$.

2. **Main step:** Update the new point (u^+, v^+, y^+) via:

$$\left(u^+, v^+\right) \in \operatorname{argmin}\left\{\mathcal{L}_\rho(u, v, y) : (u, v) \in \mathbb{R}^d \times \mathbb{R}^m\right\}, \tag{2.4}$$

$$y^+ = y + \rho\left(F\left(u^+\right) - v^+\right). \tag{2.5}$$

The rational underlying the iterative steps of ALS is quite simple and natural. Indeed, suppose we have adequate calculus rules at hand (technical details will be given later), then formally we can write the optimality conditions for ALS, which are characterized (u^+, v^+) through

$$(\mathbf{0}, \mathbf{0}) \in \partial_{(u,v)} \mathcal{L}_\rho\left(u^+, v^+, y\right),$$

where ∂ stands for some (sub)differential operation. The above is equivalent to the following system of two inclusions

$$\begin{aligned} &\mathbf{0} \in \partial\varphi\left(u^+\right) + \nabla F\left(u^+\right)^T\left(y + \rho\left(F\left(u^+\right) - v^+\right)\right), \\ &\mathbf{0} \in \partial h\left(v^+\right) - y - \rho\left(F\left(u^+\right) - v^+\right), \end{aligned}$$

where $\nabla F \in \mathbb{R}^{m \times d}$ is the Jacobian matrix associated with F. Now, thanks to (2.5), the system reduces to

$$\mathbf{0} \in \partial\varphi\left(u^+\right) + \nabla F\left(u^+\right)^T y^+ \qquad \text{and} \qquad \mathbf{0} \in \partial h\left(v^+\right) - y^+,$$

that is, we have obtained that $\mathbf{0} \in \partial\varphi\left(u^+\right) + \nabla F\left(u^+\right)^T \partial h\left(v^+\right)$. Hence, when $F\left(u^+\right) = v^+$, the latter is exactly the optimality condition for model (M), at certain triplet $(\bar{u}, \bar{v}, \bar{y})$, which reads (through the usual Lagrangian $\mathcal{L}(\cdot)$) as:

$$\begin{aligned} &\mathbf{0} \in \partial\varphi(\bar{u}) + \nabla F(\bar{u})^T \bar{y} \\ &\mathbf{0} \in \partial h(\bar{v}) - \bar{y} \qquad\qquad \Longleftrightarrow \qquad \mathbf{0} \in \partial\varphi(\bar{u}) + \nabla F(\bar{u})^T \partial h(F(\bar{u})). \\ &F(\bar{u}) = \bar{v} \end{aligned}$$

This simple description and resulting mechanism has the advantage of being essentially independent of any specific assumptions on the problem's data (except for satisfying some adequate calculus rules), e.g., convexity of the involved parts. We will now look at two interpretations of the ALS, that will throw light on the form of the scheme. In fact, if we set $y \equiv \mathbf{0}$, the resulting ALS is nothing else but a usual *penalty scheme* applied to the problem (M)

$$\min_{u \in \mathbb{R}^d, v \in \mathbb{R}^m} \{\varphi(u) + h(v) : F(u) = v\},$$

where $\rho > 0$ stands for the penalty parameter associated to the equality constraint and the term $(1/2)\|F(u) - v\|^2$ plays the role of the penalty function, which controls the "violation" of the constraint. Before we go into the second interpretation of ALS, we note that problem (M) is obviously equivalent to the following *min-max* reformulation

$$\text{(MM)} \qquad \min_{u \in \mathbb{R}^d, v \in \mathbb{R}^m} \max_{y \in \mathbb{R}^m} \{\mathcal{L}_0(u, v, y) \equiv \varphi(u) + h(v) + \langle y, F(u) - v\rangle\}.$$

A natural question, which emerges is why not simply use the Lagrangian $\mathcal{L}_0(\cdot)$ instead of the "complicated" augmented Lagrangian $\mathcal{L}_\rho(\cdot)$? In fact, the (MM) formulation reveals the problematic

issue whereby the Lagrangian $\mathcal{L}_0(\cdot)$ precisely generates the "hard penalty" reformulation of problem (M), i.e., we have

$$\text{(MM)} \qquad \min_{u\in\mathbb{R}^d, v\in\mathbb{R}^m} \{\varphi(u) + h(v) + H(u,v)\},$$

where

$$H(u,v) = \sup_{y\in\mathbb{R}^m} \langle y, F(u) - v\rangle = \begin{cases} 0, & \text{if } F(u) = v, \\ \infty, & \text{otherwise.} \end{cases}$$

In the absence of a duality theory that would allow to reverse the order of the min-max operation (i.e., under some convexity/regularity assumptions), we are thus left with an intractable nonsmooth problem. This naturally leads to the second interpretation of the augmented Lagrangian scheme that can be viewed as a *smoothing* or *proximal regularization* of the Lagrangian $\mathcal{L}(\cdot)$, i.e., of $H(u,v)$, defined above. Indeed, let

$$H_\rho(u,v) = \sup_{\eta\in\mathbb{R}^m} \left\{ \langle \eta, F(u) - v\rangle - \frac{1}{2\rho}\|\eta - y\|^2 \right\}.$$

Then, an easy computation shows that an optimal solution, η^*, is given by

$$\eta^* = y + \rho(F(u) - v).$$

Plugging this optimal solution η^* in $H_\rho(u,v)$, we then get

$$\begin{aligned} H_\rho(u,v) &= \langle y, F(u) - v\rangle + \rho\|F(u) - v\|^2 - \frac{\rho}{2}\|F(u) - v\|^2 \\ &= \langle y, F(u) - v\rangle + \frac{\rho}{2}\|F(u) - v\|^2. \end{aligned}$$

Hence, using this observation in problem (MM), naturally leads to the following equivalent problem

$$\min_{u\in\mathbb{R}^d, v\in\mathbb{R}^m} \left\{ \varphi(u) + h(v) + \langle y, F(u) - v\rangle + \frac{\rho}{2}\|F(u) - v\|^2 \equiv \mathcal{L}_\rho(u,v,y) \right\},$$

which is exactly the minimization step of the augmented Lagrangian in ALS (see step (2.4)).

Much like the penalized interpretation, this second interpretation for building an augmented Lagrangian method has the advantage of being of great generality, and does not need any duality and/or convexity arguments. Furthermore, it naturally explains the regularization (smoothing) effect that occurs through the so-called proximal approach that will be described in the next subsection, and proven to be at the heart of any Lagrangian based method. Recapitulating at this point, we end this sub-section by outlining the main difficulty with Lagrangian based methods described by ALS.

Main Difficulties with ALS

Quite clearly, the central issue to address in ALS, is how to solve the minimization step (2.4) efficiently? Indeed, in all Lagrangian based methods, the updated multiplier y^+ is given by the simple and explicit formula (2.5):

$$y^+ = y + \rho\left(F\left(u^+\right) - v^+\right).$$

Therefore, all the efforts are toward addressing the difficult minimization step (2.4). In fact, all the past, current and novel schemes in relation to Lagrangian based methods simply rely on choosing a minimization algorithm that can address both scalability and efficiency by judiciously exploiting

the specific structures and data information present in the model (M) and captured by the triplet $[\varphi, h, F]$.

In the following sections, we will demonstrate how many different variants of ALS can be built and analyzed through different choices of the minimization procedures. Moreover, one should also notice that the choice of the penalty/regularization parameter $\rho > 0$ also plays a fundamental role. This will be illustrated throughout the methods we elaborate below, both in the convex and nonconvex settings, under various structural scenarios for the triplet $[\varphi, h, F]$. In all these developments, as already alluded above, a key player is the so-called proximal mapping associated to a given function, which is introduced next.

2.2 Proximal Mappings and Minimization

From here, we will start with a precise description of the involved objects allowing us also to introduce a few definitions and relevant terminology from variational and convex analysis. We record some central properties of proximal mappings and their corresponding envelopes for convex functions. All the material described here is well-known and relies on the seminal work of Moreau [47]. However, we present here some of the fundamental properties of proximal mappings and envelopes of convex functions within the slightly more general setting of a weighted proximal term, and for completeness we have included their simple proofs. Classical convex analysis monographs describe in far more details further properties of proximal mappings and their envelopes, see e.g., [35, 5, 54], and the recent books of Bertsekas [10] and Beck [6], which also provide more recent results, applications and references. Note that proximal mappings and related schemes can also be extended beyond the Euclidean setting, whereby the classical quadratic norm is replaced by a non-Euclidean distance-like function, see the recent survey of Teboulle [57], which includes ample relevant references.

For convenience and flexibility towards the analysis of Lagrangian based schemes developed in the next section, we adopt the slightly more general version of proximal mappings as follows. Let $P \in \mathbb{S}^d_{++}$, where $\mathbb{S}^d_{++}$ is the set of all $d \times d$ positive definite matrices, and $\psi : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper, lower semi-continuous and convex function. Then,

$$\text{prox}_P^{\psi}(z) := \text{argmin}_{x \in \mathbb{R}^d} \left\{ \psi(x) + \frac{1}{2} \|x - z\|_P^2 \right\}, \tag{2.6}$$

where, for any $x \in \mathbb{R}^d$, we define $\|x\|_P = \langle Px, x \rangle^{1/2}$. Clearly, when $P := (1/\lambda) I$ with $\lambda > 0$, we recover the usual (unweighed) definition of a proximal mapping.

The corresponding *proximal envelope* of ψ, in that case, is defined by the function

$$\psi_P(z) := \inf_{x \in \mathbb{R}^d} \left\{ \psi(x) + \frac{1}{2} \|x - z\|_P^2 \right\}.$$

Below we collect some fundamental properties of $\text{prox}_P^{\psi}(\cdot)$ and $\psi_P(\cdot)$.

Lemma 2.1 (Basic properties of proximal maps/envelopes)**.** *Let $\psi : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper, lower semi-continuous and convex function and let $P \in \mathbb{S}^d_{++}$. The following statements hold true:*

(i) *The proximal mapping $\text{prox}_P^{\psi}(\cdot)$ is the unique solution of the problem (2.6), characterized via*

$$P\left(z - \text{prox}_P^{\psi}(z)\right) \in \partial\psi\left(\text{prox}_P^{\psi}(z)\right). \tag{2.7}$$

(ii) *The proximal envelope $\psi_P(\cdot)$ is finite everywhere, convex and differentiable. Moreover, its gradient is given by*

$$\nabla\psi_P(z) = P\left(z - \operatorname{prox}_P^{\psi}(z)\right), \tag{2.8}$$

and is Lipschitz continuous with Lipschitz constant $\lambda_{max}(P)$, the maximal eigenvalue of the matrix P, i.e., we have

$$\|\nabla\psi_P(z_1) - \nabla\psi_P(z_2)\| \leq \lambda_{max}(P)\|z_1 - z_2\|, \quad \forall\ z_1, z_2 \in \mathbb{R}^d.$$

(iii) *A vector $\bar{z}$ is a minimizer of ψ if and only if $\bar{z}$ is a minimizer of ψ_P and we have*

$$\inf_{x\in\mathbb{R}^d}\psi_P(x) = \inf_{x\in\mathbb{R}^d}\psi(x) = \psi(\bar{z}) = \psi\left(\operatorname{prox}_P^{\psi}(\bar{z})\right) = \psi_P(\bar{z}).$$

Proof. (i) Since $x \to \psi(x) + (1/2)\|x - z\|_P^2$ is strongly convex, it immediately follows that the minimum in (2.6) is uniquely attained at $p := \operatorname{prox}_P^{\psi}(z)$ such that $\mathbf{0} \in \partial\psi(p) + P(p - z)$.

(ii) For any $\bar{x} \in \operatorname{dom}\psi$ we clearly have

$$\psi_P(z) \leq \psi(\bar{x}) + \frac{1}{2}\|\bar{x} - z\|_P^2 < \infty,$$

and since $\Psi(x, z) := \psi(x) + (1/2)\|x - z\|_P^2$ is jointly convex in (x, z), it follows that $\psi_P(z) = \inf_{x\in\mathbb{R}^d}\Psi(x, z)$ is convex on $\mathbb{R}^d$.

Now, to establish that $\psi_P(\cdot)$ is differentiable, first note that

$$\psi_P(z) = \psi\left(\operatorname{prox}_P^{\psi}(z)\right) + \frac{1}{2}\left\|\operatorname{prox}_P^{\psi}(z) - z\right\|_P^2,$$

and, for all $v \in \mathbb{R}^d$, we have

$$\psi_P(z + v) \leq \psi\left(\operatorname{prox}_P^{\psi}(z)\right) + \frac{1}{2}\left\|\operatorname{prox}_P^{\psi}(z) - (z + v)\right\|_P^2.$$

Combining these two inequalities and some algebra yields that for any $v \in \mathbb{R}^d$

$$\gamma(v) := \psi_P(z + v) - \psi_P(z) - \left\langle v, P\left(z - \operatorname{prox}_P^{\psi}(z)\right)\right\rangle \leq \frac{1}{2}\|v\|_P^2.$$

Now, since $\psi_P(\cdot)$ is convex, then so is the function $v \to \gamma(v)$. Hence, since $\gamma(\mathbf{0}) = 0$ we get that $\gamma(v) + \gamma(-v) \geq 0$. Therefore, the last inequality also implies that $\gamma(v) \geq -(1/2)\|v\|_P^2$, and hence it follows that

$$\lim_{\|v\|_P\to 0}\frac{|\gamma(v)|}{\|v\|_P} = 0,$$

showing that $P\left(z - \operatorname{prox}_P^{\psi}(z)\right) = \nabla\psi_P(z)$. In order to prove the Lipschitz continuity of $\nabla\psi_P(\cdot)$, using (2.8), we have

$$\nabla\psi_P(z_1) - \nabla\psi_P(z_2) = P(z_1 - z_2) - P\left(\operatorname{prox}_P^{\psi}(z_1) - \operatorname{prox}_P^{\psi}(z_2)\right).$$

Since $P \in \mathbb{S}_{++}^d$ clearly P^{-1} exists and using the above we get

$$\begin{aligned}\left\langle P^{-1}\left(\nabla\psi_P(z_1) - \nabla\psi_P(z_2)\right), \nabla\psi_P(z_1) - \nabla\psi_P(z_2)\right\rangle &= \langle z_1 - z_2, \nabla\psi_P(z_1) - \nabla\psi_P(z_2)\rangle \\ &\quad - \left\langle \operatorname{prox}_P^{\psi}(z_1) - \operatorname{prox}_P^{\psi}(z_2), \nabla\psi_P(z_1) - \nabla\psi_P(z_2)\right\rangle \\ &\leq \langle z_1 - z_2, \nabla\psi_P(z_1) - \nabla\psi_P(z_2)\rangle,\end{aligned}$$

where the inequality uses the monotonicity of $\partial\psi$ (since ψ is convex), (2.7) and (2.8) proven above. From the latter inequality we immediately obtain that

$$\lambda_{min}\left(P^{-1}\right)\left\|\nabla\psi_P\left(z_1\right)-\nabla\psi_P\left(z_2\right)\right\|\leq\left\|z_1-z_2\right\|,$$

and this proves the desired result since $\lambda_{min}\left(P^{-1}\right)=\lambda_{max}\left(P\right)^{-1}$.

(iii) This easily follows from the very basic definition of $\psi_P\left(\cdot\right)$ and the relations established above. □

The relations established in Lemma 2.1(iii) are the basis for the so-called proximal minimization algorithm for tackling the problem $\inf_{x\in\mathbb{R}^d}\psi\left(x\right)$, which in turn is looking for a vector $z\in\mathbb{R}^d$ such that $\nabla\psi_P\left(z\right)=\mathbf{0}$ (for some $P\in\mathbb{S}_{++}^d$), i.e., , thanks to 2.8 to a solution of the fixed point equation $z=\mathrm{prox}_P^{\psi}\left(z\right)$. Thus, given any z, the proximal minimization algorithm computes the next update via

$$z^+=\mathrm{prox}_P^{\psi}\left(z\right)=\mathrm{argmin}_{x\in\mathbb{R}^d}\left\{\psi\left(x\right)+\frac{1}{2}\left\|x-z\right\|_P^2\right\}.$$

The proximal step plays a central role and leads to the fundamental proximal inequality that will be systematically used in the sequel. Before stating this result, we also recall the following fundamental *Pythagoras three points identity*. For any $P\in\mathbb{S}_{++}^d$, we have

$$2\left\langle w-v,P\left(v-u\right)\right\rangle=\left\|w-u\right\|_P^2-\left\|w-v\right\|_P^2-\left\|u-v\right\|_P^2,\quad\forall\ u,v,w\in\mathbb{R}^d. \tag{2.9}$$

The next result gives the announced fundamental proximal inequality.

Lemma 2.2 (Proximal Inequality)**.** *Let $\psi:\mathbb{R}^d\to(-\infty,+\infty]$ be a proper, lower semi-continuous and convex function and let $P\in\mathbb{S}_{++}^d$. Then, for $z^+=\mathrm{prox}_P^{\psi}\left(z\right)$ and for any $x\in\mathbb{R}^d$, we have*

$$\psi\left(z^+\right)-\psi\left(x\right)\leq\left\langle x-z^+,P\left(z^+-z\right)\right\rangle=\frac{1}{2}\left(\left\|x-z\right\|_P^2-\left\|x-z^+\right\|_P^2-\left\|z^+-z\right\|_P^2\right). \tag{2.10}$$

Proof. Writing the optimality condition, which characterize $z^+=\mathrm{prox}_P^{\psi}\left(z\right)$, yields

$$\mathbf{0}\in\partial\psi\left(z^+\right)+P\left(z^+-z\right).$$

Using the subgradient inequality for the convex function ψ, we have for all $x\in\mathbb{R}^d$

$$\psi\left(z^+\right)-\psi\left(x\right)\leq\left\langle x-z^+,-\gamma\right\rangle,\quad\gamma\in\partial\psi\left(z^+\right),$$

and hence, from the first inclusion, it follows that

$$\psi\left(z^+\right)-\psi\left(x\right)\leq\left\langle x-z^+,P\left(z^+-z\right)\right\rangle,$$

which proves the first inequality in (2.10). Invoking the identity (2.9) immediately yields the equality. □

In relation to Lagrangian based methods, we will also need the following useful variant of the proximal inequality.

Lemma 2.3 (Composite proximal inequality)**.** *Let $\psi:\mathbb{R}^d\to(-\infty,+\infty]$ be a proper, lower-semi continuous and convex function and let $S:\mathbb{R}^d\to\mathbb{R}$ be a differentiable function. Given a matrix $P\in\mathbb{S}_{++}^d$, we define:*

$$z^+:=\mathrm{argmin}_{x\in\mathbb{R}^d}\left\{\psi\left(x\right)+S\left(x\right)+\frac{1}{2}\left\|x-z\right\|_P^2\right\}. \tag{2.11}$$

Then, for any $x\in\mathbb{R}^d$, we have

$$\psi\left(z^+\right)-\psi\left(x\right)+\left\langle\nabla S\left(z^+\right),z^+-x\right\rangle\leq\frac{1}{2}\left(\left\|x-z\right\|_P^2-\left\|x-z^+\right\|_P^2-\left\|z^+-z\right\|_P^2\right).$$

Proof. The proof follows through identical arguments of the proof given in Lemma 2.2. □

To illustrate the viability of the general optimization model (M), we end this section with some application examples.

2.3 Application Examples

We briefly present few examples of challenging applications that can be cast as particular instances of the general model (M). We discuss below both convex and nonconvex models.

Example 1. *Convex Nonlinear Programming* [51, 9] is a constrained minimization of a convex function φ under a set of equality and inequality constraints, which is explicitly given by

$$\min_{u\in\mathbb{R}^d}\left\{\varphi(u):\ f_j(u)\leq b_j,\ (j=1,2,\ldots,p),\ f_j(u)=b_j,\ (j=p+1,p+2,\ldots,m)\right\},$$

where $f_j:\mathbb{R}^d\to(-\infty,+\infty]$, $j=1,2,\ldots,p$, are proper, lower semi-continuous and convex functions, while $f_j:\mathbb{R}^d\to(-\infty,+\infty]$, $j=p+1,p+2,\ldots,m$, are affine functions. Therefore, this fits into the general model (M), where

$$F(u):=\left(f_1(u),f_2(u),\cdots,f_m(u)\right)^T,$$

and $h(v):=\sum_{j=1}^m h_j(v_j)$ with $h_j=\iota_{(-\infty,b_j]}(v_j)$ for $j=1,2,\ldots,p$ and $h_j=\iota_{\{b_j\}}(v_j)$ for $j=p+1,p+2,\ldots,m$ (the indicator function of a set C is defined as $\iota_C(x)=0$ for $x\in C$ and $\iota_C(x)=+\infty$ for $x\notin C$). In this case, the function h is convex and nonsmooth.

Example 2. *Regularized Image Deblurring* [26, 55, 18, 61, 20] consists of finding an "image", which solves the following optimization problem

$$\min_{U\in\mathbb{R}^{p\times q}}\left\{\frac{1}{2}\left\|\mathcal{A}U-B\right\|^2+\lambda R(U)\right\},$$

where $\mathcal{A}$ is a linear mapping (usually called deblurring transformation) and $B\in\mathbb{R}^{p\times q}$. Here, we assume that both $\mathcal{A}$ and B are known in advance. The role of the regularizer $R(\cdot)$ and the regularizing parameter $\lambda>0$, is to control the obtained solution. For example, if a sparse solution is desired, then the convex regularizer $R(U)=\|U\|_1$ is very useful. Another example, which is very classic, is when one is interested in controlling the total variation of the obtained image. In this case, appropriate choice for a regularizer would be the ℓ_1-based anisotropic total variation, i.e., $R(U)=TV_1(U)$, where TV_1 stands for

$$TV_1(U)=\sum_{i=1}^{p-1}\sum_{j=1}^{q-1}\left(|u_{i,j}-u_{i+1,j}|+|u_{i,j}-u_{i,j+1}|\right)+\sum_{i=1}^{p-1}|u_{i,q}-u_{i+1,q}|+\sum_{j=1}^{q-1}|u_{p,j}-u_{p,j+1}|.$$

It is well-known that $TV_1(U)=\left\|L^T(U)\right\|_1$ where $L:\mathbb{R}^{(p-1)\times q}\times\mathbb{R}^{p\times(q-1)}\to\mathbb{R}^{p\times q}$ is defined by the formula

$$L(x,y)_{i,j}=x_{i,j}+y_{i,j}-x_{i-1,j}-y_{i,j-1},i=1,2,\ldots,p\ \text{ and }\ j=1,2,\ldots,q,$$

where we assume that $x_{0,j}=x_{p,j}=y_{i,0}=y_{i,q}=0$ for every $i=1,2,\ldots,p$ and $j=1,2,\ldots,q$. This means that in this case $h(\cdot)=\|\cdot\|_1$ and $F=L^T$ where the operator $L^T:\mathbb{R}^{p\times q}\to\mathbb{R}^{(p-1)\times q}\times\mathbb{R}^{p\times(q-1)}$ is the adjoint to L, which is given by

$$L^T(U)=(x,y),$$

where $x \in \mathbb{R}^{(p-1)\times q}$ and $y \in \mathbb{R}^{p\times(q-1)}$ are the matrices defined by

$$x = u_{i,j} - u_{i+1,j}, i = 1, 2, \ldots, p-1 \text{ and } j = 1, 2, \ldots, q,$$
$$y = u_{i,j} - u_{i,j+1}, i = 1, 2, \ldots, p \text{ and } j = 1, 2, \ldots, q-1.$$

Example 3. *Fused Lasso* [58, 59, 34] aims at solving a linear system of equations, by minimizing the corresponding least squares error measure, with the additional requirements that the solution is both sparse and "smooth". More precisely, let $\{(a_i, b_i)\}_{i=1}^N$ be a set of samples, where $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$, the problem can be then formulated as follows:

$$\min_{u\in\mathbb{R}^d} \left\{ \frac{1}{2} \|Au - b\|^2 + \lambda_1 \|u\|_1 + \lambda_2 \|Du\|_1 \right\},$$

where $A = \left[a_1^T, a_2^T, \ldots, a_N^T\right]^T \in \mathbb{R}^{N\times d}$, $b \in \mathbb{R}^N$, $D \in \mathbb{R}^{d\times(d-1)}$ defined by $(Du)_i = u_{i+1} - u_i$ and $\lambda_1, \lambda_2 > 0$ are regularizing parameters. This problem also fits well in our general model (M) by taking $h(v) = \|v\|_1$ and $F = [I, D]^T$.

Another closely related problem, is when the least squares error measure is replaced by the logistic regression measure. In this case, we obtain the following optimization problem

$$\min_{u\in\mathbb{R}^d} \left\{ \frac{1}{N} \sum_{i=1}^N \log\left(1 + \exp\left(b_i u^T a_i\right)\right) + \lambda_1 \|u\|_1 + \lambda_2 \|Du\|_1 \right\},$$

where in this case $b_i \in \{-1, 1\}$, $i = 1, 2, \ldots, N$.

In the above examples, the function h was convex. We turn to some example with a nonconvex h.

Example 4. *Orthogonality Constrained Minimization* [25, 1, 43] also known as minimization on the Stiefel manifold, can be generally formulated as follows:

$$\min_{U\in\mathbb{R}^{p\times q}} \left\{ \varphi(U) : U^T U = I \right\},$$

where the nonconvex constraint surface is called *Stiefel manifold*. In this case by choosing the model function h to be the indicator of the the nonconvex set described by the Stiefel manifold, we immediately see that this challenging class of problems fits into the general model (M).

Example 5. *Blind Image Deconvolution* [36, 39, 17] is one of the most challenging problems in imaging sciences. Similarly to the problem described above in Example 2, we are given a blurred image $B \in \mathbb{R}^{p\times q}$, and the goal is to recover a sharp image U and an unknown blurring transformation $\mathcal{A}$, which generated the blurred image B through the following linear deconvolution process $B = \mathcal{A} * U + E$, where $*$ denotes the convolution operation (of the corresponding dimensions), $\mathcal{A} \in \mathbb{R}^{r\times s}$ and $U \in \mathbb{R}^{p\times q}$, while $E \in \mathbb{R}^{p\times q}$ is an additive noise. Therefore, by adopting again the least squares error measure we are focusing on the following optimization problem:

$$\min_{A\in\mathbb{R}^{r\times s}, U\in\mathbb{R}^{p\times q}} \left\{ \frac{1}{2} \|\mathcal{A} * U - B\|^2 + h(U, \mathcal{A}) \right\},$$

where $h : \mathbb{R}^{p\times q} \times \mathbb{R}^{r\times s} \to (-\infty, +\infty]$ is a regularizer (nonconvex, and possibly nonsmooth) that controls both the recovered image U and the blurring transformation $\mathcal{A}$.

Example 6. *Regularized Structured Total Least Squares* [50, 45, 7] is another interesting application in imaging sciences that seeks to recover a sharp image out of an observed blurred image. Here, the blurring transformation is also unknown and should be recovered, but assumed to have some

linear structure. Given matrices $A, A_1, A_2, \ldots, A_p$ in $\mathbb{R}^{n\times d}$ and a vector of measurements $b \in \mathbb{R}^n$, the resulting model consists of the following nonconvex minimization problem:

$$\min_{u\in\mathbb{R}^d, w\in\mathbb{R}^p} \left\{ \frac{1}{2} \left\| \left(A + \sum_{i=1}^{p} w_i A_i \right) u - b \right\|^2 + \frac{1}{2} \|w\|^2 + h(u, w) \right\},$$

where $w_1, w_2, \ldots, w_p \in \mathbb{R}$ are unknown structure components, and $h : \mathbb{R}^d \times \mathbb{R}^p \to (-\infty, +\infty]$ is a regularizer that controls on one hand the recovered image u and on the other hand the structure components stored in w.

3 The Convex Setting

We focus on the convex setting, and consider the following particular instance of model (M), which is the *linear composite* convex optimization problem defined by

$$\text{(CM)} \qquad \min_{u\in\mathbb{R}^d} \{\varphi(u) + h(Fu)\},$$

where here $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$ and $h : \mathbb{R}^m \to (-\infty, +\infty]$ are proper, lower semi-continuous and convex functions, while $F : \mathbb{R}^d \to \mathbb{R}^m$ is a linear mapping.

The main goal of this section is to present a new and simple approach to analyze Lagrangian based schemes for tackling model (CM), and derive their convergence rates. This relies on a simple observation, which allows to show that all well-known (as well as other variants, cf. Remark 3.2) Lagrangian based schemes can be captured and analyzed through *one perturbed* proximal method of multipliers that we introduce below. We start by recalling some basic and standard assumptions with respect to the convex problem (CM).

3.1 Preliminaries on the Convex Model (CM)

We first reformulate problem (CM) in the equivalent form:

$$\text{(CM)} \qquad \min_{u\in\mathbb{R}^d, v\in\mathbb{R}^m} \{\varphi(u) + h(v) : Fu = v\},$$

for which according to section 2, the corresponding Lagrangian and augmented Lagrangian, are respectively defined by

$$\mathcal{L}(u, v, y) = \varphi(u) + h(v) + \langle y, Fu - v\rangle,$$

and, for any $\rho > 0$, by

$$\mathcal{L}_\rho(u, v, y) = \mathcal{L}(u, v, y) + \frac{\rho}{2} \|Fu - v\|^2.$$

Throughout this section the following assumption will be needed.

Assumption A. The Lagrangian $\mathcal{L}$ has a saddle point, that is, there exists (u^*, v^*, y^*) such that

$$\mathcal{L}(u^*, v^*, y) \le \mathcal{L}(u^*, v^*, y^*) \le \mathcal{L}(u, v, y^*), \quad \forall\, u \in \mathbb{R}^d, \quad \forall\, v, y \in \mathbb{R}^m.$$

Under Assumption A, and the convex data assumption for the triplet $[\varphi, h, F]$ in problem (CM), it follows that $p^* := \mathcal{L}(u^*, v^*, y^*) \in \mathbb{R}$. Furthermore, a saddle point (u^*, v^*, y^*) exists if and only if the following two conditions hold:

(i) (u^*, v^*) is a solution of problem (CM), i.e., $\varphi(u^*)$ and $h(v^*)$ are finite valued and $Fu^* = v^*$.

(ii) The multiplier y^* is an optimal solution of the dual problem associated to problem (CM), which is given by

$$\text{(DM)} \qquad d^* = \sup_{y \in \mathbb{R}^m} \left\{ d(y) := \varphi^* \left(-F^T y \right) + h^*(y) \right\},$$

where φ^* and h^* stand for the Fenchel conjugate functions of φ and h, respectively.

Thanks to the convexity of problem (CM), recall that the existence of an optimal dual Lagrange multiplier y^* attached to the constraint $Fu = v$ is ensured under the standard constraint qualification for problem (CM), that can be formulated as follows[1]

$$\text{there exists a } \bar{u} \in \operatorname{ri}(\operatorname{dom} \varphi) \text{ and a } \bar{v} \in \operatorname{ri}(\operatorname{dom} \mathrm{h}) \text{ satisfying } F\bar{u} = \bar{v},$$

and that strong duality holds, i.e., $p^* = d^* \in \mathbb{R}$. Moreover, the set of optimal dual solutions is nonempty, convex and compact.

We end these preliminaries part on the convex model (CM) with the following elementary result, which will be useful in the forthcoming analysis.

Lemma 3.1 (Objective and Feasibility Approximation)**.** *Let (u^*, v^*, y^*) be a saddle point for $\mathcal{L}$. Suppose, for every $\delta \geq 0$ and $\alpha > 0$, that*

$$\varphi(u) + h(v) - (\varphi(u^*) + h(v^*)) + \alpha \|Fu - v\| \leq \delta, \quad \forall\, u \in \mathbb{R}^d, \quad \forall\, v \in \mathbb{R}^m. \tag{3.1}$$

Then, for every couple $(u, v) \in \mathbb{R}^d \times \mathbb{R}^m$ which satisfies (3.1), the following assertions hold:

(i) $\varphi(u) + h(v) - (\varphi(u^*) + h(v^*)) \leq \delta$.

(ii) $\|Fu - v\| \leq 2\alpha^{-1}\delta$ *whenever* $2\|y^*\| \leq \alpha$.

Proof. Since $\alpha \|Fu - v\| \geq 0$, the first assertion follows at once from the premise (3.1). Now, since (u^*, v^*, y^*) is a saddle point of $\mathcal{L}$, we have that $\mathcal{L}(u^*, v^*, y^*) \leq \mathcal{L}(u, v, y^*)$, i.e.,

$$\varphi(u^*) + h(v^*) \leq \varphi(u) + h(v) + \langle y^*, Fu - v \rangle, \quad \forall\, u \in \mathbb{R}^d, \quad \forall\, v \in \mathbb{R}^m,$$

which combined with (3.1) yields

$$\alpha \|Fu - v\| \leq \delta + \langle y^*, Fu - v \rangle \leq \delta + \|y^*\| \cdot \|Fu - v\|,$$

and hence, with $\alpha > 0$ such that $2\|y^*\| \leq \alpha$, the desired claim (ii) immediately follows. □

3.2 Proximal Method of Multipliers and Fundamental Lagrangian Based Schemes

The Proximal Method of Multipliers (PMM) invented by Rockafellar [52], emerges from the Augmented Lagrangian scheme (discussed above in section 2) by adding two (weighted) quadratic proximal terms in the minimization step with respect to the two primal variables u and v. The PMM takes the following form:

[1] For any convex function $F : \mathbb{R}^d \to (-\infty, +\infty]$ we denote by $\operatorname{dom} F := \{u : F(u) < \infty\}$ its *domain*, and for any convex set S, $\operatorname{ri} S$ stands for the *relative interior* of S.

PMM – Proximal Method of Multipliers

1. **Input:** $M_1 \in \mathbb{S}^d_+$, $M_2 \in \mathbb{S}^m_+$ and $\mu \in (0,2)$.
2. **Initialization:** Start with any $\left(u^0, v^0, y^0\right) \in \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^m$.
3. **Main step:** For $k = 0, 1, \ldots$ generate the sequence $\left\{\left(u^k, v^k, y^k\right)\right\}_{k\in\mathbb{N}}$ as follows

$$\left(u^{k+1}, v^{k+1}\right) = \operatorname{argmin}_{u,v} \left\{ \mathcal{L}_\rho\left(u, v, y^k\right) + \frac{1}{2}\|u - u^k\|^2_{M_1} + \frac{1}{2}\|v - v^k\|^2_{M_2} \right\}, \tag{3.2}$$

$$y^{k+1} = y^k + \mu\rho\left(Fu^{k+1} - v^{k+1}\right). \tag{3.3}$$

The PMM of Rockafellar [52] (with $M_1 = M_2 = cI$ for some $c > 0$, and $\mu = 1$) emerges from a primal-dual application of the proximal point algorithm (as opposed to a dual application, which leads to the usual ALS described above). One advantage of the PMM is that as long as $M_1, M_2 \succ \mathbf{0}$, it manufactures strongly convex minimization problems. As a consequence, as shown in [52], it can be proven that the generated sequence converges to a saddle point of the associated Lagrangian to problem (CM), under mild assumptions on the problem's data. Note that here we introduce a positive paramater $\mu \in (0,2)$ which allows to scale the step-size in the multiplier update for a larger step-size which could enhance the overall performance of the scheme. This idea was suggested in [27] with μ being the golden ratio in the context of the so-called ADMM (which we discuss below), and as shown in [27], under some additional assumptions preserves the convergence of this scheme.

Clearly, the basic ALS is recovered from the PMM when $M_1 = M_2 = \mathbf{0}$. In that case, we need to assume that the minimization step is well-defined, i.e., admits a nonempty solution set. This can be warranted, for instance, under usual coercivity, or more generally through additional conditions on the problem's data that can be obtained through standard asymptotic (recession) analysis tools, see e.g., [4]. In what follows, we always assume that the minimization steps are well-defined.

Remark 3.1. It should be noted that actually, the augmented Lagrangian $\mathcal{L}_\rho$ already includes a quadratic proximal term with respect to the variable v, and therefore we could avoid giving an additional proximal term for this variable. However, sometimes the model (CM) can be described through a more general linear constraint of the form $Fu + Gv = b$, for which it can then be useful to add a proximal term for the v variable. For simplicity of exposition, we have adopted here the simpler formulation of the form $Fu = v$, i.e., $G = -I$ and $b = \mathbf{0}$, which in general is sufficient to describe most applications of interest. All our results can be adapted for the more general linear constraint. Thus, for the sake of completeness, and potential use within the more general linear constraint, here we will keep the proximal terms for both variables u and v.

Much like the classical ALS, a main drawback of the PMM is that for convex problems like model (CM), which exhibit nice separability in the variables u and v, the presence of the *coupling quadratic term* $Q(u,v) := \|Fu - v\|^2$ destroys it. Thus, on the practical side, this renders the joint minimization a very difficult or impossible computational task. Nevertheless, as we shall see below, the PMM remains of central importance from the theoretical perspective.

To overcome the just alluded difficulty in the PMM, and beneficially exploit the structure of model (CM), we now describe three classical approaches to eliminate the quadratic coupling term which lead to the following three fundamental Lagrangian based schemes. Recall that since the multiplier update remains the same for any Lagrangian based scheme, we only describe the minimization steps of these methods with respect to the primal variables u and v.

- **Linearized PMM (L-PMM).** This idea was developed by Chen and Teboulle [21] in the context of the PMM. By linearizing the coupling quadratic term Q around the current iteration $\left(u^k, v^k\right)$, the joint minimization step of PMM splits into the favorable easier steps (presuming as usual access to the prox maps of φ and h) that can be performed in *parallel*:

$$u^{k+1} = \operatorname{argmin}_u \left\{ \varphi(u) + \left\langle F^T\left(y^k + \rho\left(Fu^k - v^k\right)\right), u - u^k \right\rangle + \frac{1}{2}\|u - u^k\|^2_{M_1} \right\}, \tag{3.4}$$

$$v^{k+1} = \operatorname{argmin}_v \left\{ h(v) + \left\langle y^k + \rho\left(v^k - Fu^k\right), v - v^k \right\rangle + \frac{1}{2}\|v - v^k\|^2_{M_2} \right\}. \tag{3.5}$$

  This class of algorithms is particularly useful for developing parallel methods for solving nonlinear programs with block structures.

- **Proximal Alternating Direction Method of Multipliers (PADMM).** The Proximal ADMM (with $\mu = 1$) was introduced by Eckstein [23] (with $M_1 = c_1 I$ and $M_2 = c_2 I$ for some $c_1, c_2 > 0$). It consists of applying the alternating minimization approach to the joint minimization of PMM in order to eliminate the quadratic coupling:

$$u^{k+1} = \operatorname{argmin}_u \left\{ \mathcal{L}_\rho\left(u, v^k, y^k\right) + \frac{1}{2}\|u - u^k\|^2_{M_1} \right\}, \tag{3.6}$$

$$v^{k+1} = \operatorname{argmin}_v \left\{ \mathcal{L}_\rho\left(u^{k+1}, v, y^k\right) + \frac{1}{2}\|v - v^k\|^2_{M_2} \right\}. \tag{3.7}$$

  The use of positive definite matrices allows for more flexibility to consider other practical variants of PADMM (see examples below). With $M_1 = M_2 = 0$, we recover the so-called well-known Alternating Direction of Multipliers (ADMM) introduced by Glowinski and Marroco [30] and Gabay and Mercier [29]. This method has received much attention in the past and recent literature, see e.g., [24, 11, 27, 28, 31, 60, 62, 16], and as shown in [24] ADMM is in fact a special case of the Douglas-Rachford splitting method [22]. Actually, splitting/decomposition methods take their origin through the work of Douglas-Rachford [22] and were developed for the more general problem of finding the zero of the sum of two monotone operators via the abstract setting of maximal monotone operators [52, 53, 48, 41]. We refer to the monograph of Bauschke and Combettes [5] for an extensive and comprehensive treatment within the general operator setting, and references therein.

- **Linearized PADMM (L-PDAMM).** Here, we combine both strategies above, namely linearization and alternating minimization. For example, with a fixed v^k, by linearizing the quadratic coupling term $u \to 2^{-1}\rho\|Fu - v^k\|^2$ around the current iteration u^k, we obtain a Linearized PADMM

$$u^{k+1} = \operatorname{argmin}_u \left\{ \varphi(u) + \rho\left\langle F^T\left(Fu^k - v^k\right), u - u^k \right\rangle + \frac{1}{2}\|u - u^k\|^2_{M_1} \right\}, \tag{3.8}$$

$$v^{k+1} = \operatorname{argmin}_v \left\{ \mathcal{L}_\rho\left(u^{k+1}, v, y^k\right) + \frac{1}{2}\|v - v^k\|^2_{M_2} \right\}. \tag{3.9}$$

  Here, the v-step is simple as long as the function h is proximal "friendly". However, in the presence of a more general constraint (see Remark 3.1), we could also use a linearizing technique in this step to derive a simple step.

He and Yuan [32] was the first work proving ergodic global rate of convergence for PADMM and L-PADMM through a variational inequality framework. Shefi and Teboulle [56] provide a unified

analysis to establish ergodic global rate of convergence for L-PMM and PADMM and L-PADMM, and also show that the latter includes the algorithm of Chambolle and Pock [19].

We will now show that all the above methods (and other variants) can be seen through the lens of one simple scheme, which in turn leads to derive a simplified proof technique and unified analysis to establish their convergence properties.

3.3 One Scheme for All: A Perturbed PMM and its Global Rate Analysis

Main observation: The just described PMM, L-PMM, PADMM and L-PADMM schemes (and other possible variants) can simply be recovered and analyzed *at once*, through what we call a *Perturbed* PMM that we introduce next.

Before proceeding, for simplicity of exposition it will be convenient to use the following compact notations.

- Let $\xi := (u, v)$, and $L := (F, -I_m)$. Then, problem (CM) reads

$$\min\left\{\Psi(\xi) :\ L\xi = 0, \xi \in \mathbb{R}^d \times \mathbb{R}^m\right\} \equiv \min\left\{\Psi(\xi) :\ \xi \in \mathcal{F}\right\},$$

  with objective $\Psi(\xi) := \varphi(u) + h(v)$, and feasible set $\mathcal{F} = \left\{(u, v) \in \mathbb{R}^d \times \mathbb{R}^m :\ Fu = v\right\} \equiv \left\{\xi \in \mathbb{R}^d \times \mathbb{R}^m :\ L\xi = \mathbf{0}\right\}$.

- $M := \mathrm{Diag}(M_1, M_2)$, is the diagonal block matrix with diagonal blocks M_1 and M_2.

- For any sequence $\left\{\left(u^k, v^k, y^k\right)\right\}_{k\in\mathbb{N}}$ and any integer $N \geq 1$, we define the corresponding ergodic sequence $\left\{\left(\mathbf{u}^N, \mathbf{v}^N, \mathbf{y}^N\right)\right\}_{N\in\mathbb{N}}$ by

$$\mathbf{u}^N = \frac{1}{N}\sum_{k=0}^{N-1} u^{k+1}, \quad \mathbf{v}^N = \frac{1}{N}\sum_{k=0}^{N-1} v^{k+1}, \quad \mathbf{y}^N = \frac{1}{N}\sum_{k=0}^{N-1} y^{k+1}, \text{ and } \xi^N = (\mathbf{u}^N, \mathbf{v}^N). \tag{3.10}$$

- For any matrix $P \in \mathbb{S}^q_+$, any $k \geq 0$ and any three vectors $a, a^k, a^{k+1} \in \mathbb{R}^q$, we define

$$\Delta_k(a, P) := \frac{1}{2}\left\|a - a^k\right\|_P^2 - \frac{1}{2}\left\|a - a^{k+1}\right\|_P^2. \tag{3.11}$$

  When $P \equiv I_q$, the identity matrix, we simply write $\Delta_k(a, I_q) \equiv \Delta_k(a)$.

With these notations at hand, we introduce the following main scheme, which stands at the heart of our ability to unify the analysis of all the four mentioned methods.

Perturbed PMM

1. **Input:** $M = \mathrm{Diag}(M_1, M_2)$, with $M_1 \in \mathbb{S}^d_+$, $M_2 \in \mathbb{S}^m_+$ and $\mu \in (0, 2)$.
2. **Initialization:** Start with any $(\xi^0, y^0) \equiv \left(u^0, v^0, y^0\right) \in \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^m$.
3. **Main step:** For $k = 0, 1, \ldots$ generate the sequence $\left\{\left(u^k, v^k, y^k\right)\right\}_{k\in\mathbb{N}}$ as follows: choose $d^k \in \mathbb{R}^{d\times m}$ and compute

$$\xi^{k+1} = \mathrm{argmin}_\xi\left\{\mathcal{L}_\rho\left(\xi, y^k\right) + \left\langle d^k, \xi\right\rangle + \frac{1}{2}\|\xi - \xi^k\|_M^2\right\}, \tag{3.12}$$

$$y^{k+1} = y^k + \mu\rho L\xi^{k+1}. \tag{3.13}$$

The perturbation vector d^k, $k \in \mathbb{N}$, plays a central role that will be made clearer later. In fact below we will show that besides the PMM, which is obviously obtained with $d^k \equiv \mathbf{0}$, $k \in \mathbb{N}$, all fundamental Lagrangian based schemes described above can be simply recovered and analyzed at once, through a proper choice of the vector d^k, of the matrix M and of the parameter μ.

The forthcoming simple result on the Perturbed PMM is essentially the key to establish the main convergence properties of all the Lagrangian based algorithms just discussed above.

Lemma 3.2 (Fundamental Estimate for the Perturbed PMM)**.** *Let $\{(\xi^k, y^k)\}_{k\in\mathbb{N}}$ be a sequence generated by the Perturbed PMM. For any $\xi \in \mathcal{F}$ and $y \in \mathbb{R}^m$, we have*

$$\mathcal{L}\left(\xi^{k+1}, y\right) - \mathcal{L}\left(\xi, y\right) \leq \Delta_k\left(\xi, M\right) + \frac{1}{\mu\rho}\Delta_k\left(y\right) + R_k\left(\xi, M\right), \tag{3.14}$$

where $R_k\left(\xi, M\right) := -\frac{1}{2}\|\xi^{k+1} - \xi^k\|_M^2 - \left(\frac{2-\mu}{2\mu^2\rho}\right)\|y^{k+1} - y^k\|^2 - \left\langle d^k, \xi^{k+1} - \xi\right\rangle$.

Proof. Let $S\left(\xi\right) = \left\langle y^k, L\xi\right\rangle + \frac{\rho}{2}\|L\xi\|^2 + \left\langle d^k, \xi\right\rangle$. Then, within our compact notations, the primal iteration (3.12) reads:

$$\xi^{k+1} = \mathrm{argmin}_\xi \left\{\Psi\left(\xi\right) + S\left(\xi\right) + \frac{1}{2}\|\xi - \xi^k\|_M^2\right\}.$$

Invoking Lemma 2.3 with the above functions Ψ and S, we obtain for any $\xi \in \mathbb{R}^d \times \mathbb{R}^m$

$$\Psi\left(\xi^{k+1}\right) - \Psi\left(\xi\right) + \left\langle y^k + \rho L\xi^{k+1}, L\xi^{k+1} - L\xi\right\rangle + \left\langle d^k, \xi^{k+1} - \xi\right\rangle \leq \Delta_k\left(\xi, M\right) - \frac{1}{2}\|\xi^{k+1} - \xi^k\|_M^2.$$

Since $\mathcal{L}\left(\xi, y^k\right) = \Psi\left(\xi\right) + \left\langle y^k, L\xi\right\rangle$, the above can be written as,

$$\mathcal{L}\left(\xi^{k+1}, y^k\right) - \mathcal{L}\left(\xi, y^k\right) - \rho\left\langle L\xi^{k+1}, L\xi\right\rangle + \left\langle d^k, \xi^{k+1} - \xi\right\rangle \leq \Delta_k\left(\xi, M\right) - \frac{1}{2}\|\xi^{k+1} - \xi^k\|_M^2 - \rho\|L\xi^{k+1}\|^2.$$

Therefore, for any $\xi \in \mathcal{F}$ (i.e., with $L\xi = 0$), using the multiplier update (see (3.13)), the above reduces to

$$\mathcal{L}\left(\xi^{k+1}, y^k\right) - \mathcal{L}\left(\xi, y^k\right) + \left\langle d^k, \xi^{k+1} - \xi\right\rangle \leq \Delta_k\left(\xi, M\right) - \frac{1}{2}\|\xi^{k+1} - \xi^k\|_M^2 - \frac{1}{\mu^2\rho}\|y^{k+1} - y^k\|^2.$$

On the other hand, using again (3.13), followed by the three points identity, we have

$$\mathcal{L}\left(\xi^{k+1}, y\right) - \mathcal{L}\left(\xi^{k+1}, y^k\right) = \langle y - y^k, L\xi^{k+1}\rangle = \frac{1}{\mu\rho}\langle y - y^k, y^{k+1} - y^k\rangle = \frac{1}{\mu\rho}\Delta_k\left(y\right) + \frac{1}{2\mu\rho}\|y^{k+1} - y^k\|^2.$$

Adding the two relations above and collecting terms, we obtain for any $\xi \in \mathcal{F}$ and any $y \in \mathbb{R}^m$,

$$\mathcal{L}\left(\xi^{k+1}, y\right) - \mathcal{L}\left(\xi, y^k\right) \leq \Delta_k\left(\xi, M\right) + \frac{1}{\mu\rho}\Delta_k\left(y\right) + R_k\left(\xi, M\right),$$

where

$$R_k\left(\xi, M\right) = -\frac{1}{2}\|\xi^{k+1} - \xi^k\|_M^2 - \left(\frac{2-\mu}{2\mu^2\rho}\right)\|y^{k+1} - y^k\|^2 - \left\langle d^k, \xi^{k+1} - \xi\right\rangle.$$

Noting that $\mathcal{L}\left(\xi, y^k\right) \equiv \mathcal{L}\left(\xi, y\right)$ for any $\xi \in \mathcal{F}$, the proof is completed. □

The Residual Bound. As we shall see, the quantity

$$R_k(\xi, M) := -\frac{1}{2}\|\xi^{k+1} - \xi^k\|_M^2 - \left(\frac{2-\mu}{2\mu^2\rho}\right)\|y^{k+1} - y^k\|^2 - \left\langle d^k, \xi^{k+1} - \xi \right\rangle, \tag{3.15}$$

as derived in Lemma 3.2, *governs* the rate of convergence analysis of the perturbed PMM, and hence of all the four methods described above. For that purpose, we will just need to verify the following assumption, which will be shown to hold true for any of the well-known Lagrangian based schemes, by simply choosing an adequate matrix M.

Assumption R. For any $\xi \in \mathcal{F}$, and any integer $N \geq 1$, there exists a nonnegative constant $C_\rho(\xi)$ such that:

$$\sum_{k=0}^{N-1} R_k(\xi, M) \leq \frac{C_\rho(\xi)}{2}.$$

Note that the constant $C_\rho(\cdot)$ is, of course, independent of N. It will determine the value of the constant in the bound of the efficiency estimate (rate of convergence) of each method. As an appetizer, for the PMM, we clearly have $d^k = \mathbf{0}$, $k \in \mathbb{N}$, and hence since $\mu \in (0, 2)$, from (3.15) we obviously have that $R_k(\xi, M) \leq 0$, and hence Assumption **R** trivially holds with $C_\rho(\xi) \equiv 0$.

We are now ready to state and prove the main rate of convergence result for the perturbed PMM.

Theorem 3.1 (An $O(1/N)$ Rate of Convergence for the Perturbed PMM)**.** *Let $\{(\xi^k, y^k)\}_{k\in\mathbb{N}}$ be a sequence generated by the Perturbed PMM, and let (ξ^*, y^*) be a saddle point of problem (CM). Assume that Assumption **R** holds. Then, for any integer $N \geq 1$, and any positive $\alpha > 0$ such that $\alpha \geq 2\|y^*\|$, we have*

$$\Psi\left(\xi^N\right) - \Psi(\xi^*) \leq \frac{B_{\rho,\alpha}(\xi^*, M)}{2N}, \tag{3.16}$$

$$\|L\xi^N\| \leq \frac{B_{\rho,\alpha}(\xi^*, M)}{\alpha N}, \tag{3.17}$$

where $B_{\rho,\alpha}(\xi^*, M) := \|\xi^* - \xi^0\|_M^2 + C_\rho(\xi^*) + \frac{1}{\mu\rho}\left(\|y^0\| + \alpha\right)^2$.

Proof. Applying Lemma 3.2, and summing (3.14) over $k = 0, 1, \ldots, N-1$, it follows that for any $\xi \in \mathcal{F}$ and $y \in \mathbb{R}^m$,

$$\sum_{k=0}^{N-1} \mathcal{L}\left(\xi^{k+1}, y\right) - N\mathcal{L}(\xi, y) \leq \frac{1}{2}\|\xi - \xi^0\|_M^2 + \frac{1}{2\mu\rho}\|y - y^0\|^2 + \sum_{k=0}^{N-1} R_k(\xi, M).$$

Since $\xi \to \mathcal{L}(\xi, y)$ is convex, thanks to Jensen's inequality, and Assumption **R**, it follows that

$$\mathcal{L}\left(\xi^N, y\right) - \mathcal{L}(\xi, y) \leq \frac{1}{2N}\left(\|\xi - \xi^0\|_M^2 + \frac{1}{\mu\rho}\|y - y^0\|^2 + C_\rho(\xi)\right), \quad \forall\, \xi \in \mathcal{F}, \quad \forall\, y \in \mathbb{R}^m.$$

Recalling our compact notations, and the definition of $\mathcal{L}$, applying the latter at $\xi = \xi^* \in \mathcal{F}$ we get:

$$\Psi(\xi^N) - \Psi(\xi^*) + \langle y, L\xi^N\rangle \leq \frac{1}{2N}\left(\|\xi^* - \xi^0\|_M^2 + C_\rho(\xi^*) + \frac{1}{\mu\rho}\|y - y^0\|^2\right), \quad \forall\, y \in \mathbb{R}^m.$$

Maximizing both sides of this inequality over $\|y\| \leq \alpha$, we thus obtain,

$$\Psi\left(\xi^N\right) - \Psi(\xi^*) + \alpha\|L\xi^N\| \leq \frac{1}{2N}B_{\rho,\alpha}(\xi^*, M),$$

with $B_{\rho,\alpha}(\xi^*, M) = \|\xi^* - \xi^0\|_M^2 + C_\rho(\xi^*) + \frac{1}{\mu\rho}\left(\|y^0\| + \alpha\right)^2$, and the first assertion (3.16) follows. Since, $\alpha \geq 2\|y^*\|$, applying Lemma 3.1 with $\delta := (2N)^{-1}B_{\rho,\alpha}(\xi^*, M)$, proves the second assertion (3.17). □

This generic rate of convergence result is all we need to derive rate of convergence for the four fundamental Lagrangian based schemes mentioned above, as well as other variants , cf. Remark 3.2.

3.4 Special Cases of the Perturbed PMM: Fundamental Schemes

In order to apply the result obtained above on various Lagrangian based schemes, we need to identify the perturbation vector d^k (used in step (3.12)) in each method, namely to cast each method as a particular case of the perturbed PMM, and then verify for each case that Assumption **R** holds. Let us illustrate this on the PADMM.

Writing the optimality conditions of the two primal steps of the PADMM (see (3.6) and (3.7)) we get

$$\begin{cases} \mathbf{0} & \in \partial\varphi\left(u^{k+1}\right) + F^T\left(y^k + \rho\left(Fu^{k+1} - v^k\right)\right) + M_1\left(u^{k+1} - u^k\right), \\ \mathbf{0} & \in \partial h\left(v^{k+1}\right) - y^k - \rho\left(Fu^{k+1} - v^{k+1}\right) + M_2\left(v^{k+1} - v^k\right), \end{cases}$$

which is equivalent to

$$\begin{cases} \mathbf{0} & \in \partial_u \mathcal{L}_\rho\left(u^{k+1}, v^{k+1}, y^k\right) + \rho F^T\left(v^{k+1} - v^k\right) + M_1\left(u^{k+1} - u^k\right), \\ \mathbf{0} & \in \partial_v \mathcal{L}_\rho\left(u^{k+1}, v^{k+1}, y^k\right) + M_2\left(v^{k+1} - v^k\right). \end{cases}$$

The last two inclusions simply read

$$\xi^{k+1} = \operatorname{argmin}_\xi \left\{ \mathcal{L}_\rho\left(\xi, y^k\right) + \left\langle d^k, \xi \right\rangle + \frac{1}{2}\|\xi - \xi^k\|_M^2 \right\}, \text{ with } d^k \equiv \left(\rho F^T\left(v^{k+1} - v^k\right), \mathbf{0}\right),$$

thus showing that the PADMM is a particular instance of the perturbed PMM. The next result shows that Assumption **R** holds.

Lemma 3.3 (Residual Bound for PADMM)**.** *Let $\left\{\left(u^k, v^k, y^k\right)\right\}_{k\in\mathbb{N}}$ be a sequence generated by PADMM. Then, for any $M_1, M_2 \succeq 0$, and any $0 < \mu < 2 - \rho\left(\rho + \lambda_{\min}(M_2)\right)^{-1}$, we have $C_\rho(\xi) = \rho\|v^* - v^0\|^2$ for every $\xi \in \mathcal{F}$.*

Proof. As just shown above, here we have that $d^k \equiv \left(\rho F^T\left(v^{k+1} - v^k\right), \mathbf{0}\right)$, and hence in this case for all $\xi \in \mathcal{F}$

$$R_k(\xi, M) = -\frac{1}{2}\|\xi^{k+1} - \xi^k\|_M^2 - \left(\frac{2-\mu}{2\mu^2\rho}\right)\|y^{k+1} - y^k\|^2 - \rho\left\langle v^{k+1} - v^k, Fu^{k+1} - v\right\rangle, \tag{3.18}$$

where we have used the fact that $\xi \in \mathcal{F}$, i.e., $Fu = v$. Using the multiplier update, we have

$$\begin{aligned} \rho\left\langle v^k - v^{k+1}, Fu^{k+1} - v\right\rangle &= \rho\left\langle v^k - v^{k+1}, v^{k+1} - v\right\rangle + \rho\left\langle v^k - v^{k+1}, Fu^{k+1} - v^{k+1}\right\rangle \\ &= \rho\Delta_k(v) - \frac{\rho}{2}\|v^{k+1} - v^k\|^2 + \frac{1}{\mu}\left\langle v^k - v^{k+1}, y^{k+1} - y^k\right\rangle \\ &\leq \rho\Delta_k(v) - \frac{1}{2}\left(\rho - \frac{1}{s}\right)\|v^{k+1} - v^k\|^2 + \frac{s}{2\mu^2}\|y^{k+1} - y^k\|^2, \end{aligned}$$

where the last inequality follows from using the fact: $\langle a, b\rangle \leq (s/2)\|a\|^2 + (1/2s)\|b\|^2$ for any $s > 0$. Plugging this into (3.18) yields (recalling the definition of M and ξ)

$$R_k(\xi, M) \leq \rho\Delta_k(v) - \frac{1}{2}\|u^{k+1} - u^k\|_{M_1}^2 - \frac{1}{2\mu^2}\left(\frac{2-\mu}{\rho} - s\right)\|y^{k+1} - y^k\|^2 - \frac{1}{2}\|v^{k+1} - v^k\|_{M_2 + (\rho - 1/s)I}^2.$$

Since $\mu \in (0,2)$, by taking $s := (2-\mu)/\rho > 0$, we get

$$M_2 + \left(\rho - \frac{1}{s}\right) I = M_2 + \left(\rho - \frac{\rho}{2-\mu}\right) I \succeq 0 \quad \text{whenever} \quad 0 < \mu < 2 - \frac{\rho}{\rho + \lambda_{\min}(M_2)}.$$

Thus, it follows from the above inequality that $R_k(\xi, M) \leq \rho \Delta_k(v)$ and hence

$$\sum_{k=0}^{N-1} R_k(\xi, M) \leq \frac{\rho}{2} \sum_{k=0}^{N-1} \left(\|v - v^k\|^2 - \|v - v^{k+1}\|^2 \right) = \frac{\rho}{2} \left(\|v - v^0\|^2 - \|v - v^N\|^2 \right) \leq \frac{\rho}{2} \|v - v^0\|^2,$$

proving that $C_\rho(\xi) = \rho \|v - v^0\|^2$. □

From here, applying Theorem 3.1, we immediately get the rate of convergence of PADMM with its bound constant by plugging the obtained value for $C_\rho(\xi)$. Similarly, the same can be easily done for all the other decomposition schemes, which can be shown to be particular instances of the perturbed PMM. We omit the details, and just summarize below the results that can be easily verified for these methods.

- **PMM**: $d^k = (\mathbf{0}, \mathbf{0})$ with any $M_1, M_2 \succeq 0$ and $\mu \in (0,2)$. Then, we have $C_\rho(\xi^*) = 0$.
- **Linearized PMM**: $d^k = \rho \left(F^T \left(v^{k+1} - v^k \right), F \left(u^{k+1} - u^k \right) \right)$ with any

  $$M_1 = P - \rho F^T F \quad \text{and} \quad M_2 = Q - \rho I,$$

  where $\lambda_{\min}(P) \geq \lambda_{\max}(F^T F)$ and $\lambda_{\min}(Q) \geq \rho$. Then, for any $0 < \mu < 2 - t$ where

  $$t = 2\rho \cdot \max \left\{ \frac{1}{\rho + \lambda_{\min}(M_2)}, \frac{\lambda_{max}(F^T F)}{\rho \lambda_{max}(F^T F) + \lambda_{\min}(M_1)} \right\},$$

  we have $C_\rho(\xi) = \rho \|u^* - u^0\|^2_{F^T F} + \rho \|v^* - v^0\|^2$
- **Proximal ADMM**: $d^k = \rho \left(F^T \left(v^{k+1} - v^k \right), \mathbf{0} \right)$ with any $M_1, M_2 \succeq 0$. Then, for any $0 < \mu < 2 - \rho \left(\rho + \lambda_{\min}(M_2) \right)^{-1}$, we have $C_\rho(\xi) = \rho \|v^* - v^0\|^2$.
- **Linearized PADMM**: $d^k = \rho \left(F^T \left(v^{k+1} - v^k \right), \mathbf{0} \right)$ with any $M_1 = P - \rho F^T F$ (where $\lambda_{\min}(P) \geq \lambda_{\max}(F^T F)$) and $M_2 \succeq 0$. Then, for any $0 < \mu < 2 - \rho \left(\rho + \lambda_{\min}(M_2) \right)^{-1}$, we have $C_\rho(\xi) = \rho \|v^* - v^0\|^2$.

Remark 3.2. Other combinations of the above scenarios can be similarly derived. For instance, it is also easy to consider a variant of the perturbed PMM for the case where the function φ is smooth with a Lipschitz continuous gradient. In that case we can linearize the function φ, and then we can use the well-known descent lemma, to adapt the bound in the key estimate of Lemma 3.2, and from there derive the corresponding $O(1/N)$ rate from Theorem 3.1. Finally, note that the global pointwise convergence of the described algorithms can also be derived as a by-product of the above analysis using classical arguments. For sake of brevity we omit the details.

4 The Nonconvex Setting

We are back to our general nonconvex and nonlinear composite model

$$\text{(M)} \qquad \min_{u \in \mathbb{R}^d} \left\{ \Phi(u) \equiv \varphi(u) + h(F(u)) \right\},$$

with the following problem data:

- $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function.
- $h : \mathbb{R}^m \to (-\infty, +\infty]$ is a proper and lower semi-continuous function.
- $F : \mathbb{R}^d \to \mathbb{R}^m$ $(m \leq n)$ is a continuously differentiable map.

As exemplified in subsection 2.3, this structure is very flexible and captures many fundamental optimization problems arising in disparate applications. As opposed to the development and analysis of Lagrangian based schemes in the convex setting, the nonconvex model is much harder and quite challenging, and the literature remains scarce. Not only the lack of convexity already renders the problem difficult, but also the nonlinear nature of the map $F(\cdot)$, which in turn adds severe difficulties preventing us to apply the tools and approaches used in the convex setting. In this section, we briefly describe some very recent results as developed in Bolte et al. [15], highlighting the main difficulties, and outlining a general abstract approach, which paves the way to analyze Lagrangian based methods for the general nonconvex model (M).

4.1 The Nonconvex Nonlinear Composite Optimization - Preliminaries

As explained in section 2, we first reformulate problem (M) in the equivalent split form:

$$\text{(M)} \qquad \min_{u \in \mathbb{R}^d, v \in \mathbb{R}^m} \{\varphi(u) + h(v) : F(u) = v\}.$$

The classical *Lagrangian* $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^m \to (-\infty, +\infty]$ is defined, as before, by

$$\mathcal{L}(u, v, y) \equiv \varphi(u) + h(v) + \langle y, F(u) - v \rangle,$$

and the associated *augmented Lagrangian* is defined as follows

$$\mathcal{L}_\rho(u, v, y) := \varphi(u) + h(v) + \langle y, F(u) - v \rangle + \frac{\rho}{2} \|F(u) - v\|^2,$$

where $\rho > 0$ is a penalty parameter.

To ensure the well-posedness of the algorithms to come, we assume for any fixed $y \in \mathbb{R}^m$, that

$$\inf_{u,v} \mathcal{L}_\rho(u, v, y) > -\infty.$$

To work with nonconvex functions we need some appropriate subdifferentials that we now record. All these results can be found in greater details in the book of Rockafellar and Wets [54].

Definition 4.1 (Subdifferentials)**.** Let $\psi : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semi-continuous function with $\operatorname{dom} \psi := \{z \in \mathbb{R}^d : \psi(z) < \infty\}$.

(i) For a given $z \in \operatorname{dom} \psi$, the *Frechet subdifferential* of ψ at z, denoted by $\hat{\partial}\psi(z)$ is the set of all vectors $v \in \mathbb{R}^d$ satisfying

$$\liminf_{\substack{x \to z \\ x \neq z}} \frac{\psi(x) - \psi(z) - \langle v, x - z \rangle}{\|x - z\|} \geq 0.$$

(ii) The *limiting subdifferential* of ψ at z, denoted by $\partial\psi(z)$ is the set of all vectors $v \in \mathbb{R}^d$ for which there exist two sequences $\{z^k\}_{k \in \mathbb{N}}$ and $\{v^k\}_{k \in \mathbb{N}}$ such that $v^k \in \partial\psi\left(z^k\right)$ and

$$\lim_{k \to \infty} \left(z^k, \psi\left(z^k\right), v^k\right) = (z, \psi(z), v).$$

(iii) The *horizon subdifferential* of ψ at z, denoted by $\partial^\infty \psi(z)$ is the set of all vectors $v \in \mathbb{R}^d$ for which there exist a sequence of real numbers $\{t_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} t_k = 0$ and two sequences $\{z^k\}_{k\in\mathbb{N}}$ and $\{v^k\}_{k\in\mathbb{N}}$ such that $v^k \in \partial\psi\left(z^k\right)$ and

$$\lim_{k\to\infty}\left(z^k, \psi\left(z^k\right), t_k v^k\right) = \left(z, \psi(z), v\right).$$

To ensure the validity of calculus rules for problem (M), we need to assume a constraint qualification. We record below a central one that is relevant to our model (M).

Assumption B. The following constraint qualification condition holds for problem (M),

$$[\text{CQ}] \qquad \nabla F(u)^T z = \mathbf{0}, \quad z \in \partial^\infty h(F(u)) \implies z = \mathbf{0}.$$

In the classical nonlinear programming case, which is a special case of model (M), (cf. Example 1 in section 2.3), the general condition [CQ] simply reduces to the classical *Mangasarian-Fromovitz* constraint qualification condition: there is no vector $z \neq \mathbf{0}$, which satisfies $\nabla F(u)^T z = \mathbf{0}$.

The condition [CQ] is also crucial to guarantee the subdifferential calculus rules and the chain rule, which are needed to write the first order optimality condition for model (M). More precisely, if u is a local minimum of problem (M) satisfying condition [CQ], then there exists $z \in \mathbb{R}^m$ such that $\nabla\varphi(u) + \nabla F(u)^T z = \mathbf{0}$ with $z \in \partial h(F(u))$. Thus, for model (M), under [CQ], the set of critical points of problem (M) reads as:

$$\operatorname{crit}\Phi = \left\{u \in \mathbb{R}^d : \mathbf{0} \in \nabla\varphi(u) + \nabla F(u)^T \partial h(F(u))\right\}.$$

Next, we introduce a fundamental and classical concept of regularity needed to handle the nonlinearity of the map $F(\cdot)$, see Milnor [46].

Definition 4.2 (Uniform Regularity)**.** Let $\Omega \subset \mathbb{R}^d$ be an open set, $F : \Omega \to \mathbb{R}^m$ be a C^1 map, and $\emptyset \neq S \subset \Omega$. The mapping F is uniformly regular on S with constant $\gamma > 0$ if the following holds:

$$\left\|\nabla F(u)^T v\right\| \geq \gamma \|v\|, \ \forall\, u \in S, \ v \in \mathbb{R}^m.$$

For a given $u \in \Omega$, we can equivalently define

$$\gamma(F, u) := \min\left\{\left\|\nabla F(u)^T v\right\| : \|v\| = 1\right\},$$

and that $\gamma(F, u)$ is nonzero means, that $\nabla F(u)$ is surjective, or that $\nabla F(u)\nabla F(u)^T$ is positive definite.

A main issue in Lagrangian based methods is to handle *simultaneously* the penalty parameter, the data constants and the constraint qualification condition. However, as we know, augmented Lagrangian methods are based on *relaxing* the classical Lagrangian, and hence are obviously *unfeasible*. The looser is the relaxation, the more unfeasible is the method. The looseness of the relaxation is measured through the penalty parameter ρ, which is used to penalize the constraint $F(u) = v$ in $\mathcal{L}_\rho$, namely:

$$\lim_{\rho\to+\infty} \mathcal{L}_\rho(u, v, y) = \begin{cases} \varphi(u) + h(F(u)), & \text{if } F(u) = v, \\ +\infty, & \text{otherwise.} \end{cases}$$

While in the convex setting, and with a linear map F, we can globally control the behavior of the points generated by a Lagrangian based scheme, in the nonconvex and nonlinear setting, the lack of feasibility creates new and severe difficulties in the sense that an unfeasible method generates points that might be out of control, inducing three major obstacles:

- Constraint qualification condition may fail.
- Problem's data input, such as global Lipschitz constants, may become unknown, or out of reach.
- The inherent min-max dynamics of Lagrangian based methods clearly imply that the augmented Lagrangian function *alternatively* increases and decreases. Therefore, we can not warrant/measure the crucial *descent* property needed in the convergence analysis through it.

In the forthcoming subsection, we describe the key ideas and building blocks of the theory, which allow to address these three main obstacles.

4.2 ALBUM - Adaptive Lagrangian Based MUltiplier Method

In the remaining parts of this section, we have briefly outlined the fundamental ideas and theoretical tools that were developed in [15], and we refer the reader to that work for the specific details, proofs, more results and references.

The main underlying ideas allowing us to address the alluded obstacles rely on the following ingredients:

- Define an *information zone* to be a region for which *regularity is under control* and data constants are known.
- Build a *generic Lagrangian based scheme with an adaptive regime* aiming at detecting the information zone, and forcing the iterates to enter and stay within this zone.
- Once the the zone is found, using the adaptive regime, *detect an adequate Lyapunov function* ensuring the necessary descent properties.

We start by introducing the notion of information zone, which is in fact an enlargement of the feasible set.

Definition 4.3 (Information Zone – Enlargement of the Feasible Set)**.** An information zone is a subset $\mathcal{Z}$ of $\mathbb{R}^d$ such that there exists $\bar{d} \in (0, +\infty]$ for which

$$\mathcal{Z} \supset \left\{ u \in \mathbb{R}^d : \operatorname{dist}\left(F\left(u\right), \operatorname{dom} h\right) \leq \bar{d} \right\} \supset \mathcal{F},$$

where $\mathcal{F} = \left\{ u \in \mathbb{R}^d : F\left(u\right) \in \operatorname{dom} h \right\}$ is the feasible set of problem (M).

Equipped with an information zone $\mathcal{Z}$, our second assumption for problem (M) is *local regularity* and *local smoothness* (with respect to the information zone) of the problem's data.

Assumption C. (i) F is uniformly regular over $\mathcal{Z}$ with positive constant γ,

(ii) ∇F is $L(F)$ Lipschitz continuous over $\mathcal{Z}$,

(iii) $\nabla \varphi$ is $L(\varphi)$ Lipschitz continuous over $\mathcal{Z}$.

Throughout the rest of this section, we assume that Assumptions B and C hold.

Our next ingredient is to introduce an adequate Lyapunov function that will allow us to measure the descent property of the proposed scheme, which due to the min-max dynamics of Lagrangian based methods, cannot be detected through the usual augmented Lagrangian function.

Definition 4.4 (A Lyapunov Function)**.** Fix $\beta > 0$ and $w \in \mathbb{R}^d$. The function $\mathcal{E}_\beta(\cdot)$, which is defined by

$$\mathcal{E}_\beta(u, v, y, w) := \mathcal{L}_\rho(u, v, y) + \beta \|u - w\|^2,$$

is called a *Lyapunov function* associated to problem (M).

The next result records the key fact showing that $\mathcal{E}_\beta$ can be beneficially used to identify critical points of the original problem (M).

Proposition 4.1 (Critical Points Relationships)**.** *[15, Proposition 1]. The following implications hold for all $\beta, \rho > 0$*

$$(u, v, y, u) \in \operatorname{crit} \mathcal{E}_\beta \implies (u, v, y) \in \operatorname{crit} \mathcal{L}_\rho \implies u \in \operatorname{crit} \Phi.$$

Both the information zone and the Lyapunov function $\mathcal{E}_\beta$ just defined above, will come into play in a dynamical (adaptive) fashion in the forthcoming algorithm. As already seen, a typical Lagrangian based method is given by

$$\begin{aligned} \left(u^+, v^+\right) &\in \mathcal{A}_\rho(u, v, y), \\ y^+ &= y + \rho\left(F\left(u^+\right) - v^+\right), \end{aligned}$$

where $\mathcal{A}_\rho$ stands for any minimization black-box map, while the multiplier update y^+ remains the same for any Lagrangian based scheme. We now formalize the definition of the map $\mathcal{A}_\rho$, which captures the mechanism that will essentially govern the convergence of a generic Lagrangian based scheme, and play an essential role in controlling the relevant parameters in our generic Adaptive Lagrangian Based mUltiplier Method (**ALBUM**).

Definition 4.5 (Lagrangian Algorithmic Map)**.** A map $\mathcal{A}_\rho : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^d \times \mathbb{R}^m$ is called a *Lagrangian algorithmic map* if there exist two constants $a, b > 0$ such that

$$\text{(i)} \quad \frac{a}{2}\left\|u^+ - u\right\|^2 + \mathcal{L}_\rho\left(u^+, v^+, y\right) \leq \mathcal{L}_\rho(u, v, y),$$

and

$$\text{(ii)} \quad \left\|\nabla_u \mathcal{L}_\rho\left(u^+, v^+, y\right)\right\| \leq b\left\|u^+ - u\right\|.$$

Note that once we choose the map $\mathcal{A}_\rho$, it fully determines the constants a and b. These constants might depend on the problem's data input or/and other algorithmic constants. This will be illustrated later on. With the above framework, we are now ready to state our generic adaptive Lagrangian based multiplier method.

Adaptive Lagrangian Based mUltiplier Method – ALBUM

1. **Input:** $\mathcal{A}_\rho$ an algorithmic map, with the corresponding parameters $a, b > 0$.
2. **Initialization:** Fix $\delta, \rho_0 > 0$ and start with any $\left(u^0, v^0, y^0\right)$ where $u^{-1} \equiv u^0$.
3. **Main step:** For $k = 0, 1, \ldots$ generate the sequence $\left\{\left(u^k, v^k, y^k\right)\right\}_{k\in\mathbb{N}}$ as follows

   3.1. **Primal step**
   $$\left(u^{k+1}, v^{k+1}\right) \in \mathcal{A}_{\rho_k}\left(u^k, v^k, y^k\right).$$

   3.2. **Multiplier step**
   $$y^{k+1} = y^k + \rho_k\left(F\left(u^{k+1}\right) - v^{k+1}\right).$$

   3.3. **Adaptive step:** choose $\tau \in \left(0, \frac{a}{2}\right)$, set $\beta_k := \frac{b^2}{\rho_k \gamma}$, and compute
   $$\Delta_{\beta_k} := \mathcal{E}_{\beta_k}\left(u^k, v^k, y^k, u^{k-1}\right) - \mathcal{E}_{\beta_k}\left(u^{k+1}, v^{k+1}, y^{k+1}, u^k\right).$$

   If $u^{k+1} \notin \mathcal{Z}$ or $\tau\left\|u^{k+1} - u^k\right\|^2 > \Delta_{\beta_k}$, set $\rho_{k+1} = \rho_k + \delta$. Otherwise, set $\rho_{k+1} = \rho_k$.

4.3 A Methodology for Global Analysis of Lagrangian Based Methods

Basically, **ALBUM** aims at forcing u^k to enter the information zone $\mathcal{Z}$ in finitely many steps and stay there. The adaptive protocol essentially turns **ALBUM** into a descent scheme for the Lyapunov function $\mathcal{E}_\beta$: it leads to a finite identification of the zone $\mathcal{Z}$, and warrants sufficient descent of $\mathcal{E}_\beta$ through the adaptive step of **ALBUM**. This crucial adaptive step relies on detecting weak feasibility, in the sense that we ask for $u^k \in \mathcal{Z}$, and not for actual feasibility, which would be obviously impossible given the inherent nonfeasibility of a Lagrangian based scheme! Moreover, it detects a descent property for $\mathcal{E}_\beta$ (and again clearly not possible for $\mathcal{L}_\rho$!), which implicitly tunes the algorithm to match the natural step-sizes attached to φ and F.

A main remaining ingredient to derive convergence results for **ALBUM** is then a fundamental definition characterizing the required properties of a sequence generated by **ALBUM**, which is called a Lagrangian sequence.

Definition 4.6 (Lagrangian Sequence)**.** A sequence $\{w^k\}_{k\in\mathbb{N}} := \left\{\left(u^k, v^k, y^k\right)\right\}_{k\in\mathbb{N}}$ generated by **ALBUM** is called a *Lagrangian sequence*, if it satisfies the following conditions:

C1 There exists a positive constant a such that
$$\frac{a}{2}\left\|u^{k+1} - u^k\right\|^2 + \mathcal{L}_{\rho_k}\left(u^{k+1}, v^{k+1}, y^k\right) \leq \mathcal{L}_{\rho_k}\left(u^k, v^k, y^k\right), \quad \forall\ k \geq 0.$$

C2 There exists a positive constant b such that
$$\left\|\nabla_u \mathcal{L}_{\rho_k}\left(u^{k+1}, v^{k+1}, y^k\right)\right\| \leq b\left\|u^{k+1} - u^k\right\|, \quad \forall\ k \geq 0.$$

C3 There exists a positive constant c and $q^{k+1} \in \partial_v \mathcal{L}_{\rho_k}\left(u^{k+1}, v^{k+1}, y^k\right)$ such that
$$\left\|q^{k+1}\right\| \leq c\left\|u^{k+1} - u^k\right\|, \quad \forall\ k \geq 0.$$

C4 Let $\bar{v}$ be a limit point of a subsequence $\left\{v^k\right\}_{k\in\mathcal{K}}$ of $\{v^k\}_{k\in\mathbb{N}}$, then $\limsup_{k\in\mathcal{K}\subset\mathbb{N}} h\left(v^k\right) \leq h\left(\bar{v}\right)$.

Before proceeding, we make the following remarks regarding conditions **C1** – **C4**.

- Thanks to the definition of the Lagrangian algorithmic map $\mathcal{A}_\rho$, once the map is chosen, conditions **C1** and **C2** immediately hold for the sequence $\{w^k\}_{k\in\mathbb{N}}$.
- Condition **C1** is actually a *partial descent property* of the augmented Lagrangian $\mathcal{L}_\rho(\cdot)$ only with respect to the primal variable u:

$$\frac{a}{2}\left\|u^{k+1}-u^k\right\|^2+\mathcal{L}_{\rho_k}\left(u^{k+1},v^{k+1},y^k\right)\leq\mathcal{L}_{\rho_k}\left(u^k,v^k,y^k\right),\quad\forall\ k\geq 0.$$

  It pertains to the primal variables (u,v), since by nature the multiplier y is an "ascent variable".
- Conditions **C2** and **C3** provide (sub)gradient bounds of the augmented Lagrangian $\mathcal{L}_\rho(\cdot)$ with respect to both primal variables (u,v).
- Condition **C4** is a simple sequential assumption on h, it holds, e.g., when $h:\operatorname{dom}h\to\mathbb{R}$ is continuous.

We are now ready to state the main convergence results for **ALBUM** proven in [15, Theorems 1 and 2] which we record in the following theorem.

Theorem 4.1 (Sequential and Global Convergence)**.** *Let $\{w^k\}_{k\in\mathbb{N}}$ be generated by* **ALBUM** *that satisfies conditions* **C1** – **C4***, i.e., a Lagrangian sequence. Assume that the sequence $\{w^k\}_{k\in\mathbb{N}}$ is bounded. Then the following two assertions hold:*

(i) *(Subsequence convergence). Let $(\bar{u},\bar{v},\bar{y})$ be a limit point of $\{w^k\}_{k\in\mathbb{N}}$. Then, $\bar{u}$ is a critical point of problem (M).*

(ii) *(Global convergence). Assume that φ, F and h are semi-algebraic. Then, the whole sequence $\{w^k\}_{k\in\mathbb{N}}$ converges to $(\bar{u},\bar{v},\bar{y})$ where $\bar{u}$ is a critical point of problem (M).*

Note that the second part of the theorem above establishes *global convergence*, i.e., the *whole* sequence generated by **ALBUM** converges to a critical point of problem (M) with *semi-algebraic data*.[2] This relies on the fundamental nonsmooth version of the so-called *Kurdyka-Łojasiewicz property*. The KL property was introduced in the seminal works [42, 37] for the smooth case, and [12, 13] for the nonsmooth case, the later includes an in depth study of KL functions and many relevant references. The KL property is a crucial ingredient toward the global convergence analysis of descent schemes [3, 14]. A general methodology to prove global convergence of descent methods is derived in the work of Bolte, Sabach and Teboulle [14] which relies on a key uniformization of the KL property [14, Lemma 6, p. 478]. While verifying the KL property of a given function is often a very difficult task, it holds for the class of semi-algebraic functions, which are ubiquitous in many applications (see [14] for examples), thus making the KL property a powerful tool for proving convergence of descent schemes for semi-algebraic problems. The central result establishing that KL holds for the broad class of semi-algebraic functions was derived in the seminal work of Bolte, Daniilidis and Lewis [12].

[2]Recall that a subset S of $\mathbb{R}^d$ is real semi-algebraic set if constructible by finite systems of polynomial inequalities, and that a function is semi-algebraic if its graph is semi-algebraic.

4.4 ALBUM in Action: Global Convergence of Lagrangian Based Schemes

Now we illustrate the application of **ALBUM** on the two classical schemes PMM and Proximal ADMM, previously studied in the convex case (see section 3). To this end, in order to put these two methods in the general framework of **ALBUM**, we need first to identify the corresponding algorithmic maps $\mathcal{A}_\rho$ that update the primal variables u and v (recall that the multiplier update remains identical for any Lagrangian based scheme).

A Proximal Multipliers Method (PMM)

$$\left(u^{k+1}, v^{k+1}\right) \in \operatorname{argmin}_{(u,v)} \left\{ \mathcal{L}_\rho\left(u, v, y^k\right) + \frac{1}{2}\left\|u - u^k\right\|_{M_1}^2 \right\}, \quad (M_1 \succ 0).$$

It is clear that in this case, the map $\mathcal{A}_\rho$ stands for the joint minimization of the (partial) proximal counterpart of the augmented Lagrangian $\mathcal{L}_\rho$[3].

A Proximal Alternating Direction Method of Multipliers (PADMM)

$$\begin{aligned} v^{k+1} &\in \operatorname{argmin}_v \mathcal{L}_\rho\left(u^k, v, y^k\right), \\ u^{k+1} &\in \operatorname{argmin}_u \left\{ \mathcal{L}_\rho\left(u, v^{k+1}, y^k\right) + \frac{1}{2}\left\|u - u^k\right\|_{M_1}^2 \right\}, \quad (M_1 \succ 0). \end{aligned}$$

In this case we obviously have that the map $\mathcal{A}_\rho$ consists of alternating minimization applied on the (partial) proximal counterpart of the augmented Lagrangian $\mathcal{L}_\rho$.

It can be shown [15] that both algorithmic maps corresponding to the PMM and PADMM schemes are Lagrangian algorithmic maps satisfying the premises of Definition 4.5 with the same parameters:

$$a = \lambda_{\min}(M_1) > 0 \quad \text{and} \quad b = \lambda_{\max}(M_1) > 0, \quad (\text{since } M_1 \succ 0).$$

The *adaptive version* of both methods is obtained from **ALBUM** through their corresponding maps $\mathcal{A}_\rho$ just described above. In order to derive the global convergence of any bounded sequence generated by these two adaptive versions to critical points of problem (M), we need to ensure that both methods are producing *Lagrangian sequences*, i.e., that conditions **C1** – **C4** hold (cf. Definition 4.6). This was proven in [15, Section 6], which we record in the following result.

Proposition 4.2. *Let $\{w^k\}_{k\in\mathbb{N}}$ be a sequence generated by either the Adaptive PMM or the Adaptive PADMM. Then, conditions* **C1** – **C3** *hold. If, in addition, $\{w^k\}_{k\in\mathbb{N}}$ is bounded, then condition* **C4** *holds true, and hence $\{w^k\}_{k\in\mathbb{N}}$ is a Lagrangian sequence.*

Equipped with this result for these adaptive versions of PMM and PADMM, we can then apply Theorem 4.1 to obtain global convergence to critical points of problem (M) in the nonlinear and nonconvex setting for both methods.

We end by briefly indicating other decomposition variants (as discussed previously in the convex case) that fit into the **ALBUM** framework and include

- The classical ADMM. This is obtained by taking $M_1 = 0$ in the PADMM. Under the additional assumption that $u \to \mathcal{L}_\rho(u, v, y)$ is σ-strongly convex for any fixed $v, y \in \mathbb{R}^m$, the global convergence of ADMM to critical points of the nonlinear composite model (M) can be derived.

[3] Recall that the use of the partial version on the variable u is enough, since the variable v has a built-in proximal term through the definition of the augmented Lagrangian.

- The Proximal Linearized ADMM. Here the model (M) is nonconvex but, F is assumed to be a linear map, and the condition number of FF^T is strictly smaller than 2. The scheme consists of *one gradient iteration* of the u-step and it reads as:

$$v^{k+1} \in \operatorname{argmin}_v \mathcal{L}_\rho\left(u^k, v, y^k\right),$$
$$u^{k+1} \in \operatorname{argmin}_u \left\{\left\langle u-u^k, \nabla_u \mathcal{L}_\rho\left(u^k, v^{k+1}, y^k\right)\right\rangle + \frac{1}{2}\left\|u-u^k\right\|_{M_1}^2\right\}, \quad (M_1 \succ 0).$$

  Clearly, in that case, the u-step reduces to an easy explicit formula. The general convergence results of **ALBUM** can then be applied to this variant, including the determination of the penalty parameter in terms of the problem's data, see [15] for details.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[2] K. J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Non-linear Programming*. Stanford Mathematical Studies in the Social Sciences, vol. II. Stanford University Press, Stanford, Calif., 1958.

[3] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.*, 137(1-2, Ser. A):91–129, 2013.

[4] A. Auslender and M. Teboulle. *Asymptotic Cones and Functions in Optimization and Variational Inequalities*. Springer Monographs in Mathematics. Springer-Verlag, New York, 2003.

[5] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, second edition, 2017.

[6] A. Beck. *First-Order Methods in Optimization*, volume 25 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2017.

[7] A. Beck, A. Ben-Tal, and C. Kanzow. A fast method for finding the global solution of the regularized structured total least squares problem for image deblurring. *SIAM J. Matrix Anal. Appl.*, 30(1):419–443, 2008.

[8] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Computer Science and Applied Mathematics. Academic Press Inc., New York-London, 1982.

[9] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, second edition, 1999.

[10] D. P. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, Belmont, MA, 2015.

[11] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation*. Prentice-Hall International Editions, Englewood Cliffs, NJ, 1989.

[12] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.*, 17:1205–1223, 2006.

[13] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.*, 362(6):3319–3363, 2010.

[14] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2, Ser. A):459–494, 2014.

[15] J. Bolte, S. Sabach, and M. Teboulle. Nonconvex Lagrangian-based optimization: monitoring schemes and global convergence. *Math. Oper. Res.*, 43(3):1210–1232, 2018.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* 3(1):1–122, 2011.

[17] P. Campisi and K. Egiazarian, editors. *Blind Image Deconvolution: Theory and Applications*. CRC Press, Taylor & Francis Group, 2007.

[18] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vision*, 20(1-2):89–97, 2004.

[19] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.

[20] R. H. Chan, M. Tao and X. Yuan. Constrained total variation deblurring models and fast algorithms based on alternating direction method of multipliers. *SIAM J. Imaging Sci.*, 6(1):680–697, 2013.

[21] G. Chen and M. Teboulle. A proximal-based decomposition method for convex minimization problems. *Math. Programming*, 64(1, Ser. A):81–101, 1994.

[22] J. Douglas and H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Trans. Amer. Math. Soc.*, 82(2):421–439, 1956.

[23] J. Eckstein. Some saddle-function splitting methods for convex programming. *Optim. Methods Softw.*, 4(1):75–83, 1994.

[24] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Programming*, 55(3, Ser. A):293–318, 1992.

[25] A. Edelman, T. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1999.

[26] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[27] M. Fortin and R. Glowinski. *Augmented Lagrangian Methods*, volume 15 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1983.

[28] D. Gabay. Applications of the method of multipliers to variational inequalities. In M. Fortin and R. Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems*, pages 299–340. North-Holland, Amsterdam, 1983.

[29] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.*, 2(1):17–40, 1976.

[30] R. Glowinski and A. Marrocco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér.*, 9, 1975.

[31] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*. Studies in Applied and Numerical Mathematics. Society for Industrial Mathematics, 1983.

[32] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[33] M. Hestenes. Multiplier and gradient methods. *J. Optim. Theory Appl.*, 4(5):303–320, 1969.

[34] J. M. Hilbe. *Logistic Regression Models*. Chapman & Hall/CRC Texts in Statistical Science Series. CRC Press, Boca Raton, FL, 2009.

[35] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer-Verlag, Berlin, 2001.

[36] D. Kundur and D. Hatzinakos. Blind image deconvolution. *IEEE Signal Process. Mag.*, 13(3):43–64, 1996.

[37] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier (Grenoble)*, 48(3):769–783, 1998.

[38] L. S. Lasdon. *Optimization Theory For Large Systems*. Dover Publications, Inc., Mineola, NY, 2002. Reprint of the 1970 original.

[39] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1964–1971. IEEE, Miami, FL, 2009.

[40] G. Li and T. K. Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM J. Optim.*, 25(4):2434–2460, 2015.

[41] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964–979, 1979.

[42] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles (Paris, 1962)*, pages 87–89. Éditions du Centre National de la Recherche Scientifique, Paris, 1963.

[43] J. H. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Trans. Signal Process.*, 50(3):635–650, 2002.

[44] B. Martinet. Régularisation d'inéquations variationnelles par approximation ssuccessives. *Rev. Française Informat. Recherche Opérationnelle*, 4:154–158, 1970.

[45] N. Mastronardi, P. Lemmerling, A. Kalsi, D. P. O'Leary, and S. van Huffel. Implementation of the regularized structured total least squares algorithms for blind image deblurring. *Linear Algebra Appl.*, 391:203–221, 2004.

[46] J. Milnor. *Topology from the Differentiable Viewpoint*. Princeton University Press, 1931.

[47] J. J. Moreau. Proximitéet dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.

[48] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *J. Math. Anal. Appl.*, 72(2):383–390, 1979.

[49] M. J. D. Powell. A method for nonlinear constraints in minimization problems. In R. Fletcher, editor, *Optimization*, pages 283–298. Academic Press, New York, 1969.

[50] A. Pruessner and D. P. O'Leary. Blind deconvolution using a regularized structured total least norm algorithm. *SIAM J. Matrix Anal. Appl.*, 24(4):1018–1037, 2003.

[51] R. T. Rockafellar. *Convex Analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[52] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.*, 1(2):97–116, 1976.

[53] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optimization*, 14(5):877–898, 1976.

[54] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Fundamental Principles of Mathematical Sciences*. Springer-Verlag, Berlin, 1998.

[55] L. I. Rudin, S. J. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[56] R. Shefi and M. Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM J. Optim.*, 24(1):269–297, 2014.

[57] M. Teboulle. A simplified view of first order methods for optimization. *Math. Program.*, 170(1, Ser. B):67–96, 2018.

[58] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[59] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005.

[60] P. Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.*, 29(1):119–138, 1991.

[61] C. Wu and X. C. Tai. Augmented Lagrangian method, dual methods, and split Bregman iteration for ROF, vectorial TV, and high order models. *SIAM J. Imaging Sci.*, 3(3):300–339, 2010.

[62] J. Yang and Y. Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM J. Sci. Comput.*, 33(1):250–278, 2011.