Nonconvex Lagrangian-Based Optimization: Monitoring Schemes and Global Convergence

Jérôme Bolte* Shoham Sabach† Marc Teboulle‡

Accepted for publication in "Mathematics of Operations Research", August 27, 2017

Abstract

We introduce a novel approach addressing global analysis of a difficult class of nonconvex-nonsmooth optimization problems within the important framework of Lagrangian-based methods. This genuine nonlinear class captures many problems in modern disparate fields of applications. It features complex geometries, qualification conditions, and other regularity properties do not hold everywhere. To address these issues we work along several research lines to develop an original general Lagrangian methodology which can deal, all at once, with the above obstacles. A first innovative feature of our approach is to introduce the concept of Lagrangian sequences for a broad class of algorithms. Central to this methodology is the idea of turning an arbitrary descent method into a multiplier method. Secondly, we provide these methods with a transitional regime allowing us to identify in finitely many steps a zone where we can tune the step-sizes of the algorithm for the final converging regime. Then, despite the min-max nature of Lagrangian methods, using an original Lyapunov method we prove that each bounded sequence generated by the resulting monitoring schemes are globally convergent to a critical point for some fundamental Lagrangian-based methods in the broad semialgebraic setting, which to the best of our knowledge, are the first of this kind.

1 Introduction.

Consider the following nonconvex and nonlinear composite minimization problem

$$\text{(CM)} \qquad \text{minimize}\left\{f(x) \equiv f_0(x) + h(F(x)) : x \in \mathbb{R}^n\right\},$$

where

- $f_0 : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function.
- $F : \mathbb{R}^n \to \mathbb{R}^m$ $(m \leq n)$ is a continuously differentiable mapping defined by

$$F(x) := (f_1(x), f_2(x), \ldots, f_m(x)).$$

- $h : \mathbb{R}^m \to (-\infty, +\infty]$ is a proper and lower semi-continuous (lsc) function.

*TSE (Université Toulouse I), Manufacture des Tabacs, 21 allée de Brienne, 31015 Toulouse, France. E-mail: jerome.bolte@tse-fr.eu.

†Faculty of Industrial Engineering, 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.

The structure of the composite model (CM) offers extreme versatility over the traditional nonlinear programming formulation. The smooth assumptions are in the mapping F and the function f_0, while constraints, penalties and nonconvex/nonsmooth terms can be handled by the nonconvex and nonsmooth function h. The composite structure allows to beneficially model a given problem and exploit data information, and essentially captures most optimization problems. This is illustrated below in Section 1.1.

The main objective of this paper is to layout the main theoretical tools to achieve a deep understanding of augmented Lagrangian based methods and their fundamental properties in the nonconvex setting described by model (CM).

The Augmented Lagrangian (AL) methodology has a long history which can be traced back to the works of Hestenes [20], Powell [29] and Haarhoff and Buys [19] with the so-called multipliers method for problems with equality constraints. The AL algorithmic framework was a major breakthrough in nonlinear optimization providing the ground to fundamental algorithms and applications which have been extensively studied in the literature for various classes of problems. For classical results on the subject including many key results, extensions and closely related schemes such as the Proximal Methods of Multipliers (PMM) [30] and the Alternating Direction of Multipliers (ADM) [16, 17], we refer the reader to the monographs of Bertsekas [9] and Bertsekas-Tsitsiklis [8] and references therein.

Recently, there has been an intensive renewed interest in augmented Lagrangian based methods, and in particular within the ADM scheme. This surge of interest is mainly due to the emergence of new and modern applications arising in a broad diversity of applications areas such as signal processing, sparse approximation in data analysis and machine learning. These problems share particular structures which often adapt well to ADM and lead to computationally attractive schemes. A typical prototype which has been extensively studied is when all the data is *convex* with F being a *linear* mapping, and problem (CM) reduces to the *convex linear composite problem*:

$$\text{(CM-L)} \qquad \text{minimize} \left\{ f_0(x) + h(Fx) : x \in \mathbb{R}^n \right\}.$$

The recent literature on ADM for this convex problem is voluminous and clearly it is not the purpose of this paper to review it here. See, for instance, the recent work [32] for an account of old and new results on the convergence analysis of various augmented Lagrangian schemes, as well as many relevant references to earlier works and to more modern and recent contributions in the convex setting.

This work is a complete departure from the classical convex linear composite model. Indeed, in many of the modern applications alluded above the optimization model turns out to be not only nonsmooth but also includes inherent nonlinearities which the nonlinear composite model (CM) conveniently captures. Unfortunately, while as just mentioned, the analysis of Lagrangian based methods has been extensively studied in the convex case, the situation in the nonconvex setting is far from being well understood, and global analysis of Lagrangian methods for the general model (CM) remains scarce. In fact, only very recently some progress has been initiated in the nonconvex case, but *only for the linear* composite model (CM-L), see *e.g.*, [24] and references therein. Even in the simpler linear composite model, the situation is not trivial and the authors in [24] have to rely on various assumptions on the problem's data. Out of studies on the linear composite model, we are not aware of any work attempting to fully understand Lagrangian based methods for the general nonlinear composite model (CM) considered here. The objective of the present work is to address this situation, and to develop the main theoretical tools to achieve a deeper understanding

of Lagrangian based methods and their fundamental properties in the nonconvex setting described by the nonlinear composite model (CM).

Before outlining some details on our approach, main contributions and results, we first recall some of the major obstacles met in the study of Lagrangian methods by evoking three most salient theoretical issues:

1. *AL methods are non-feasible methods:* this is due to the very nature of the penalty approach used to construct an augmented Lagrangian. As a consequence feasibility issues have to be dealt with particular care as they have a direct damaging impact on qualification conditions, as explained next.

2. *Failure of qualification conditions*: A major problem with non-feasible methods is that qualification conditions must hold in a larger sense in order to allow for the good behavior of the algorithm when the current point is far from the feasible set. Yet, for very simple constraints, for instance spherical constraints (see Example 1.3 and Remark 2.3), assuming a qualification condition everywhere is not a viable option.

3. *Oscillation issues*: AL methods are particularly well designed to handle problems having complex geometry, like for instance nonlinear inequality/equality constrained problems. A typical and difficult problem in this context is to tame oscillations of minimizing sequences[1]. Moreover, AL methods are of *min-max dynamics* and thus, by nature, the values taken by the augmented Lagrangian function alternatively increase and decrease even if the sequence eventually converges. This oscillatory behavior makes the use and the design of Lyapunov functions particularly difficult.

One of the goals of this paper is to provide the reader with an original general Lagrangian methodology which can deal, all at once, with the above obstacles under general and mild assumptions on the problem's data. Let us briefly outline our exact contributions now.

The first innovative feature of our approach is to introduce and to study a broad class of algorithms through sequences that we call *Lagrangian sequences*. At the heart of this methodology is the idea of turning an arbitrary descent method into a multiplier method. The rationale is simple, once a method or mechanism is chosen, it is implemented on the primal variable(s) of the augmented Lagrangian, while the multiplier variable is updated in the classical and straightforward fashion. An illustrative but very informative instance of this approach is the famous proximal method of multipliers (PMM) alluded above which is modeled through an augmented Lagrangian with an added proximal term and consists of performing a proximal step on the primal variable while the multiplier is updated as in the classical AL method.

Based on the above methodology, we proceed and describe how we address the three points evoked above.

To circumvent the qualification failures and the lack of knowledge of fundamental constants, we introduce the notion of *information zone*. It is a subset of the space containing the feasible set and on which Lipschitz continuity and qualification conditions are known to hold and are quantifiable by simple real numbers (Lipschitz constants and regularity modulus). Then we provide our methodology with an *adaptive regime* that aims at detecting this zone and at forcing the iterates to stay within the zone. The detection of the zone is made by tuning dynamically the penalization

[1]Similar difficulties occur in other approaches, see for instance, [12] for an illustration in the context of sequentially convex programming approaches, and [1] in the context of an exact penalty approach.

parameter of the augmented Lagrangian at a sufficiently high value. This approach is shown to identify the zone in finitely many steps and deals thus with points 1 and 2.

Once the information zone is found, another crucial issue remains to address: rule out oscillations to ensure descent properties of the method, this is point 3 above. This is done by using once more the adaptive idea to detect an adequate Lyapunov function. At a technical level this function is nonincreasing but the rate of decrease is only controlled for one block of the primal sequence which is a departure from classical analysis.

The proposed novel approach and theoretical analysis developed in Sections 2 to 5 allow us to eliminate the difficulties evoked above and to derive a generic Adaptive Lagrangian Based mUltiplier Method (ALBUM) for tackling the general nonconvex and nonlinear composite model (CM) which encompasses fundamental Lagrangian methods. This paves the way to derive convergence results, and in particular, global convergence results to a critical point of problem (CM) with semi-algebraic data, by relying on the nonsmooth Kurdyka-Łojasiewicz (KL) inequality [25, 22, 11]. The potential of our results is demonstrated through the study of two major Lagrangian schemes whose convergence was never analyzed in the proposed general setting: the proximal multiplier method and the proximal alternating direction of multipliers scheme, this is done in Section 6 where we also consider some additional interesting variants. We end the introduction with some examples illustrating the versatility of model (CM).

1.1 Examples of model (CM).

Below we give some examples which exhibit the versatility of model (CM). The first example describes various well-known and classical models in the nonlinear optimization literature, while the remaining four examples describe models arising in some recent applications.

Example 1.1 (Nonlinear programming)**.** The standard nonlinear program with equality and inequality constraints:

$$\text{(NLP)} \qquad \inf_{x\in\mathbb{R}^n} \{f_0(x) :\ f_i(x) \leq 0,\ i=1,2,\ldots,p,\ \ f_i(x) = 0,\ i=p+1,p+2,\ldots,m\},$$

can be reformulated through the composite model (CM) by defining the separable model function $h(u) := \sum_{i=1}^{m} h_i(u_i)$, where

$$h_i(u_i) = i_{(-\infty,0]}(u_i),\ i=1,2,\ldots,p, \quad \text{and} \quad h_i(u_i) = i_{\{0\}}(u_i),\ i=p+1,p+2,\ldots,m.$$

Lagrangians and Smooth penalties. The standard Lagrangian associated to (NLP) as well as linear and quadratic penalty terms can easily be reformulated through model (CM) with a separable model function h and an adequate choice of $h_i,\ i=1,2,\ldots,m$. For instance with $h_i(u_i) = y_i u_i,\ i=1,2,\ldots,m$, the standard Lagrangian of problem (NLP) is recovered. Likewise the usual penalized counterpart of the problem (NLP) given by

$$\text{(P-NLP)} \qquad \inf\left\{ f_0(x) + \sum_{i=1}^{p} \mu_i \max\{0, f_i(x)\}^2 + \sum_{i=p+1}^{m} \mu_i |f_i(x)|^2 \right\},\ (\mu_i > 0),$$

is recovered through model (CM) with the obvious choices

$$h_i(u_i) = \mu_i \max\{0,u_i\}^2,\ i=1,2,\ldots,p, \quad \text{and} \quad h_i(u_i) := \mu_i |u_i|^2,\ i=p+1,p+2,\ldots,m.$$

Obviously, the classical augmented Lagrangian itself for NLP can easily be recovered from model (CM) as well, with an adequate piecewise quadratic choice of h_i, $i = 1, 2, \ldots, m$, for the inequality constraints.

Nonsmooth and nonseparable h. A classical nonsmooth model is the ℓ_1-norm penalized problem for equality constraints ($p \equiv 0$ in (NLP)) given by

$$\inf_{x \in \mathbb{R}^n} \left\{ f_0(x) + \sum_{i=1}^{m} w_i |f_i(x)| \right\},$$

which is covered by model (CM) with $h_i(u_i) := w_i |u_i|$ for some $w_i > 0$, $i = 1, 2, \ldots, m$.

Nonseparable nonsmooth: mini-max problems. Let $f_0 \equiv 0$ and $h(u) := \max\{u_1, u_2, \ldots, u_m\}$. Then, model (CM) produces the classical nonlinear mini-max problem

$$\inf_{x \in \mathbb{R}^n} \max_{1 \le i \le m} f_i(x).$$

The above example exhibit the versatility of model (CM) for traditional NLP. In all these examples h was convex. We now give three examples with *nonconvex* h which include a broad variety of fundamental problems arising in applications.

Example 1.2 (Sparsity constrained problems)**.** These problems arise in many areas of applications, for example, compressive sensing and machine learning see *e.g.*, [33]. A basic model (see [5]) reads

$$\min\{f(x) : \|x\|_0 \le s\},$$

where $\|\cdot\|_0$ stands for the usual counting function, *i.e.*, the number of nonzero coordinates of x, $s > 0$ is the desired sparsity level, and f can be any smooth fidelity criterion (*e.g.*, least squares). Let $S := \{x : \|x\|_0 \le s\}$. Then, the above problem is a special case of model (CM) with $f_0(x) \equiv f(x)$, $F(x) \equiv x$ and h is the nonconvex function described by the indicator of the closed set S, *i.e.*, $h(u) \equiv i_S(u)$.

Matrix rank minimization problems can be similarly formulated in the space of symmetric matrices using a constraint of the form $\mathrm{rank}(x) \le s$.

Moreover, nonconvex *penalized approximations* of the following form have also been considered and found useful (see, *e.g.*, [26] and references therein)

$$\min\left\{ f(x) + \rho \sum_{i=1}^{n} \varphi(|x_i|) \; x \in \mathbb{R}^n \right\}, \quad (\rho > 0 \text{ is a penalty parameter}),$$

where φ is a concave (increasing) function on $\mathbb{R}$ used to approximate the l_0-quasi norm. A typical example is the l_p-quasi norm with $\varphi(t) := t^p$, $0 < p < 1$, and model (CM) covers this case as well, with an obvious identification for the nonconvex function h.

Example 1.3 (Matrix minimization on Stiefel manifolds)**.** Optimization problems with matrix orthogonality constraints arise in many applications of science and engineering (*e.g.*, polynomial optimization, combinatorial optimization, eigenvalue problems, sparse PCA, matrix rank minimization, etc., [15]). A basic problem reads as:

$$\text{(O)} \quad \min\left\{ \Psi(X) : X^T X = I, \, X \in \mathbb{R}^{n \times p} \right\},$$

where $\Psi : \mathbb{R}^{n\times p} \to \mathbb{R}$ is a smooth function (often quadratic), and I stands for the $p \times p$ identity matrix. The feasible set $\mathcal{S}_{n,p} := \left\{X \in \mathbb{R}^{n\times p} : X^T X = I\right\}$ is known as the *Stiefel manifold*, which for $p = 1$ reduces to the unit-sphere manifold $\mathcal{S}_{n,1} \equiv \mathcal{S}_n = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. Clearly, with h being the nonconvex function described by the indicator of the closed set $\mathcal{S}_{n,p}$, problem (O) can easily be seen as a special case of model (CM) with the obvious identification for f_0 and F in the space of real matrices $\mathbb{R}^{n\times p}$.

Example 1.4 (Nonconvex feasibility)**.** Let $S_1, S_2, \ldots, S_p$ (for $p \geq 2$) be nonempty and closed subsets of $\mathbb{R}^n$. The nonconvex feasibility problem consists in finding a point in the intersection $\cap_{i=1}^p S_i$. These type of problems abound in many applications such as phase retrieval, network sensors localizations or protein conformation, see *e.g.*, [18] for some recent developments. One standard way to tackle the feasibility problem is simply to reformulate it as an optimization problem:

$$\min\left\{\frac{1}{2\,(p-1)}\sum_{i=2}^{p}\|x_1 - x_i\|^2 + \sum_{i=1}^{p} i_{S_i}(x_i) :\ (x_1, x_2, \ldots, x_p) \in \mathbb{R}^{n\times p}\right\},$$

Observe that $\bar{x} \in \cap_{i=1}^p S_i$ if and only if the optimal value of the above optimization problem at $(\bar{x}, \bar{x}, \ldots, \bar{x}) \in \mathbb{R}^{n\times p}$ is zero.

Choosing $\mathbb{R}^{n\times p}$ as the base space, setting $f_0(x_1, x_2, \ldots, x_p) = (2\,(p-1))^{-1}\sum_{i=2}^p \|x_1 - x_i\|^2$ (which is obviously a $C^{1,1}$ function), $F(x_1, x_2, \ldots, x_p) = (x_1, x_2, \ldots, x_p)$ and $h(x_1, x_2, \ldots, x_p) = \sum_{i=1}^p i_{S_i}(x_i)$, we see that the above optimization problem fits our general model (CM).

Notations. For any vector $w \in \mathbb{R}^d$, the standard Euclidean norm is simply denoted by $\|w\|$. Unless otherwise stated, for the subdifferential operators $\hat{\partial}$, ∂ and ∂^∞ and other objects coming from variational analysis, we adopt the notations and definitions of the monograph by Rockafellar and Wets [31].

2 The Lagrangian for nonlinear composite problems.

This section outlines the first steps toward the generic algorithm we develop and analyze in this paper. We define the augmented Lagrangian associated to problem (CM), basic qualification condition and assumptions, and in particular, we introduce the fundamental and new concept of *information zone* which play a central role in the forthcoming analysis.

2.1 Lagrangian and qualification condition.

In analogy to standard NLP, one can construct a natural Lagrangian for problem (CM) as follows. We first reformulate problem (CM) in the equivalent split form:

$$\text{(CM)} \qquad \inf\{f_0(x) + h(u) :\ u = F(x),\ (x,u) \in \mathbb{R}^n \times \mathbb{R}^m\}.$$

For this abstract equality constrained reformulation, the classical *Lagrangian* is defined by $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \to (-\infty, +\infty]$ via

$$\mathcal{L}(x, u, y) \equiv f_0(x) + h(u) + \langle y, F(x) - u\rangle.$$

An *augmented Lagrangian* is a quadratic penalized version of the Lagrangian:

$$\begin{aligned}\mathcal{L}^\sharp_\rho(x, u, y) &:= \mathcal{L}(x, u, y) + \frac{\rho}{2}\|F(x) - u\|^2 \\ &= f_0(x) + h(u) + \langle y, F(x) - u\rangle + \frac{\rho}{2}\|F(x) - u\|^2, \end{aligned} \tag{2.1}$$

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

To ensure the well-posedness of the algorithms to come, throughout this paper we assume:

$$\inf_{x,u} \mathcal{L}_\rho^\sharp(x,u,y) > -\infty \text{ for any fixed } y \in \mathbb{R}^m. \tag{2.2}$$

We assume below that model (CM) satisfies a standard qualification condition which we express in the compact form provided by variational analysis [31, Chapter 10, pp. 428–430]. We denote by $\nabla F(x) \in \mathbb{R}^{m\times n}$ the Jacobian matrix of F, whose rows are given by the gradient vectors $[\nabla f_i(x)]_{i=1}^m$.

Assumption A. The following constraint qualification holds for problem (CM),

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

For the classical NLP case, which can be obtained from model (CM) as described in Example 1.1, the condition [CQ] reduces to the classical Mangasarian-Fromovitz constraint qualification, see *e.g.*, [9].

The condition [CQ] is not only essential to provide smoothness and regularity of the constraint set, at a technical level, it is also important to provide a chain rule for the objective function of model (CM). This allows us to derive the first order necessary conditions for this model.

Definition 2.1 (First order optimality condition)**.** Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a continuously differentiable mapping, and let $h : \mathbb{R}^m \to (-\infty, +\infty]$ be a proper and lsc function. If x is a local minimizer of problem (CM) satisfying Assumption A, then there exists $y \in \mathbb{R}^m$ such that

$$\nabla f_0(x) + \nabla F(x)^T y = 0 \quad \text{with} \quad y \in \partial h(F(x)).$$

The set of critical points of a function ψ, is denoted by $\operatorname{crit}\psi$. For problem (CM) with the objective function f, we have

$$\operatorname{crit} f = \left\{ x \in \mathbb{R}^n : \, 0 \in \nabla f_0(x) + \nabla F(x)^T \partial h(F(x)) \right\}. \tag{2.3}$$

2.2 The information zone.

Lagrangian based methods require to handle simultaneously penalty parameters, constants, and qualification condition which is a delicate matter. An important aspect of this work is to address these issues.

Augmented Lagrangian methods are based on relaxing the classical Lagrangian and therefore by nature these are unfeasible methods. Measures of unfeasibility of these methods are naturally connected to the "looseness" of the relaxation. The looser is the relaxation, the more unfeasible is the method. Over relaxation could even result in absurd behaviors.

The augmented Lagrangian $\mathcal{L}_\rho^\sharp$ as given in (2.1) is

$$\mathcal{L}_\rho^\sharp(x,u,y) := f_0(x) + h(u) + \langle y, F(x) - u\rangle + \frac{\rho}{2}\|F(x) - u\|^2, \quad \text{with } \rho > 0.$$

In this context the looseness/sharpness of the relaxation is embodied within the penalty parameter ρ which is used to penalize the constraint $F(x) = u$ in the augmented Lagrangian $\mathcal{L}_\rho^\sharp$. At an analytic level this penalty reflects the fact that for a fixed triple (x,u,y) one has

$$\lim_{\rho\to+\infty} \mathcal{L}_\rho^\sharp(x,u,y) = \begin{cases} f_0(x) + h(F(x)), & \text{if } F(x) = u, \\ +\infty, & \text{otherwise,} \end{cases}$$

which amounts, in some sense, to the convergence of $\mathcal{L}_\rho^\sharp$ to $\mathcal{L}$ as $\rho \to +\infty$.

A major drawback of such unfeasible methods, easily guessed from the above, is that they generate points that might be out of control in the sense that:

– constraint qualification conditions may fail,

– assumptions on the problem's data, such as global Lipschitz constants of the various objects involved may become unknown or out of reach.

On the other hand, assuming a global control is very demanding and could be unrealistic in practice.

To remedy these obstacles all at once our approach is twofold: we first define an information zone, denoted by $\mathcal{Z}$, to be a region for which regularity is under control and constants are known. Second we provide a generic Lagrangian scheme described below with an extra-adaptive search made to reach the information zone[2]

Let $\operatorname{dom} h = \{u \in \mathbb{R}^m : h(u) < \infty\}$ which is nonempty and closed. Then the feasible set of problem (CM) is defined by

$$\mathcal{F} = \{x \in \mathbb{R}^n : F(x) \in \operatorname{dom} h\}.$$

Definition 2.2 (Information zone)**.** Given the feasible set $\mathcal{F}$ for problem (CM), an information zone is a subset $\mathcal{Z}$ of $\mathbb{R}^n$ such that there exists $\bar{d} \in (0, +\infty]$ for which

$$\mathcal{Z} \supset \left\{x \in \mathbb{R}^n : \operatorname{dist}(F(x), \operatorname{dom} h) \leq \bar{d}\right\} \supset \mathcal{F}. \tag{2.4}$$

The information zone is an enlargement of the feasible set $\mathcal{F}$. It should be noted that the information zone $\mathcal{Z}$ depends on the parameter $\bar{d}$. For simplicity of exposition, in the forthcoming section, this dependence is not explicitly mentioned. In the next definition we recall a fundamental and classical regularity assumption (see, *e.g.*, Milnor [27]).

Definition 2.3 (Uniform regularity)**.** Let Ω be an open subset of $\mathbb{R}^n$, $F : \Omega \to \mathbb{R}^m$ be a continuously differentiable mapping, and let S be a nonempty subset of Ω. We say that F is uniformly regular on S with constant $\gamma > 0$ if the following holds:

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

Remark 2.1. For a given $x \in \Omega$, asserting that

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

is nonzero is equivalent to the fact that $\nabla F(x)$ is surjective or $\nabla F(x) \nabla F(x)^T$ is positive definite. In nonlinear optimization it is also known as Mangasarian-Fromovitz condition at x. Geometrically it means that the set $\{y \in U : F(y) = F(x)\}$ is a C^1 manifold for any small open neighborhood around x.

Note also that

$$\gamma \equiv \gamma(F, x) = \sqrt{\lambda_{\min}\left(\nabla F(x) \nabla F(x)^T\right)}, \tag{2.5}$$

where $\lambda_{\min}(A)$ denotes the smallest eigenvalue of a real symmetric matrix A.

[2] As we shall see soon the adaptive regime allows also for dynamic adjustment of the step-sizes to other geometrical features.

2.3 Basic assumptions for model (CM).

We introduce the following essential assumptions.

Assumption B. Given an information zone $\mathcal{Z}$, we assume that:

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

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

(iii) ∇f_0 is $L(f_0)$ Lipschitz continuous over $\mathcal{Z}$.

Remark 2.2. (a) Naturally, the Lipschitz continuity and the uniform regularity are not required on the whole space $\mathbb{R}^n$, but only on the information zone $\mathcal{Z}$. This is a departure from the usual setting.

(b) When ∇f_0 is known to be Lipschitz continuous on the whole space $\mathbb{R}^n$, and the mapping F is assumed to be linear, *i.e.*, $F(x) = Fx$ for some matrix $F \in \mathbb{R}^{n\times m}$ with full row rank, then Assumption B holds with $\mathcal{Z} \equiv \mathbb{R}^n$ (*i.e.*, $\bar{d} = +\infty$) and $FF^T \succeq \gamma I_n$ where $\gamma = \sqrt{\lambda_{\min}(FF^T)} > 0$.

Let us illustrate the concept of the information zone on a simple but fundamental and emblematic situations (*cf.* Example 1.3).

Example 2.1 (Spherical constraints)**.** Assume that $F(x) = \|x\|^2$ and $h = i_{\{1\}}$. For simplicity we also assume that f_0 is globally Lipschitz.

One has $\nabla F(x) = 2x$ and thus for a fixed x, $\gamma(F, x) = 2\|x\|$. Take $r_1 \in (0, 1)$, and define $\mathcal{Z} = \{x \in \mathbb{R}^n : r_1 \leq \|x\|\}$. We see that F is $2r_1$ regular on $\mathcal{Z}$ and ∇F is 2-Lipschitz continuous. Hence $\mathcal{Z}$ can be chosen as an information zone as long as we show that (2.4) holds true. Take $\bar{d} = 1 - r_1^2$, it is easy to check that $\left|\|x\|^2 - 1\right| \leq \bar{d}$ implies, in particular, that $1 - \|x\|^2 \leq 1 - r_1^2$. Note that $0 \notin \mathcal{Z}$ and that $\mathbb{R}^n$ could not be an acceptable choice for an information zone because of the degeneracy of ∇F at $\{0\}$.

Remark 2.3 (Systematic failure of global CQ with compact equality constraints)**.** The preceding example reveals a simple and systematic phenomenon which motivates strongly the use of an information zone. Consider a C^1 function $F : \mathbb{R}^n \to \mathbb{R}$ such that $[F = 0]$ is a compact manifold and assume that int $[F \leq 0] = [F < 0]$. Then, necessarily there exists x^* such that $\nabla F(x^*) = 0$. Indeed, by taking x^* to be a minimizer of F over the compact set $[F \leq 0]$ and since this minimizer lies within $[F < 0]$ it follows that $\nabla F(x^*) = 0$. This shows that in general, it is not possible, to have $\mathcal{Z} = \mathbb{R}^n$.

3 Adaptive Lagrangian based multiplier method.

From now on Assumptions A and B form our blanket assumptions.

As explained previously, difficult obstacles are faced both in the design and the study of Lagrangian based methods: lack of descent, and above all, feasibility issues. The adaptive idea we develop here is precisely meant to put us in a position where these issues are treated in a dynamical fashion: both the information zone and the "energy functional" $\mathcal{E}_\beta$ which we introduce now come into a play.

3.1 Lagrangian and a Lyapunov function.

We shall need to work with an auxiliary function which is very similar to the augmented Lagrangian $\mathcal{L}_\rho^\sharp$ (defined in (2.1)). This is a classical approach often called the "Lyapunov" methodology. It will reveal the optimizing property of the generic Lagrangian scheme we introduce next.

Let $\beta > 0$ and $w \in \mathbb{R}^n$, here we consider the Lyapunov function which is defined by

$$\mathcal{E}_\beta(x,u,y,w) := \mathcal{L}_\rho^\sharp(x,u,y) + \beta \|x-w\|^2. \tag{3.1}$$

Below, we record the relationships between the critical point sets of the three relevant functions f, $\mathcal{L}_\rho^\sharp$ and $\mathcal{E}_\beta$. These relations already suggest the pivotal role to be played by $\mathcal{E}_\beta$. Recall that condition [CQ] is always assumed, *i.e.*, Assumption A holds.

Proposition 3.1 (Critical points relationships)**.** *Let $x \in \mathbb{R}^n$ and $u, y \in \mathbb{R}^m$. The following implications hold:*

$$(x,u,y,x) \in \operatorname{crit} \mathcal{E}_\beta \implies (x,u,y) \in \operatorname{crit} \mathcal{L}_\rho^\sharp \implies x \in \operatorname{crit} f,$$

for all $\beta, \rho > 0$.

Proof. The result follows easily from standard subdifferential calculus rules. Indeed, from the definition of $\mathcal{E}_\beta$ (see (3.1)) we have that $(x,u,y,w) \in \operatorname{crit} \mathcal{E}_\beta$ if and only if

$$(0,0,0,0) \in \left(\nabla_x \mathcal{L}_\rho^\sharp(x,u,y) + 2\beta(x-w), \partial_u \mathcal{L}_\rho^\sharp(x,u,y), \nabla_y \mathcal{L}_\rho^\sharp(x,u,y), 2\beta(w-x)\right). \tag{3.2}$$

On the other hand, using the definition of $\mathcal{L}_\rho^\sharp$ (see (2.1)) we obtain

$$\nabla_x \mathcal{L}_\rho^\sharp(x,u,y) = \nabla f_0(x) + \nabla F(x)^T (y + \rho(F(x)-u)), \tag{3.3}$$

$$\partial_u \mathcal{L}_\rho^\sharp(x,u,y) = \partial h(u) + \rho(u - F(x)) - y, \tag{3.4}$$

$$\nabla_y \mathcal{L}_\rho^\sharp(x,u,y) = F(x) - u. \tag{3.5}$$

Therefore, taking $w = x$ in (3.2), the first implication in the proposition follows. The second implication follows by noticing that with $(x,u,y) \in \operatorname{crit} \mathcal{L}_\rho^\sharp$, the three relations (3.3), (3.4) and (3.5) reduce to $0 = \nabla f_0(x) + \nabla F(x)^T y$ and $0 \in \partial h(F(x)) - y$. Hence, using Definition 2.1, we obtain that $x \in \operatorname{crit} f$. This complete the proof. □

3.2 A generic algorithm: ALBUM.

In order to describe the forthcoming generic scheme, we first need to introduce a primal black-box map which governs the mechanism of the global convergence methodology to be developed in Section 3.3.

Definition 3.1 (Lagrangian algorithmic map)**.** Consider the optimization model (CM) and its associated augmented Lagrangian $\mathcal{L}_\rho^\sharp$ which is defined in (2.1). Let $(x,u,y) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ be any given triple. A *primal black-box map* $\mathcal{A}_\rho$ generates a couple (x^+, u^+) by

$$\left(x^+, u^+\right) \in \mathcal{A}_\rho(x,u,y).$$

A primal black-box map $\mathcal{A}_\rho$ is called a *Lagrangian algorithmic map* if there are two positive constants a and b such that

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

and

$$\text{(ii)} \quad \left\|\nabla_x \mathcal{L}_\rho^\sharp\left(x^+, u^+, y\right)\right\| \le b\left\|x^+ - x\right\|.$$

Thus, once we chose the Lagrangian algorithmic map $\mathcal{A}_\rho$, this choice fully determine the constants a and b, which play an important role in the generic algorithm outlined below. Note that these constants might depend on the problem's data input (*e.g.*, Lipschitz constant, uniform regularity constant, or/and algorithmic constants, *e.g.*, proximal/penalty parameters). We deferred to Section 6 for two instances of fundamental Lagrangian algorithmic maps.

The proposed generic adaptive algorithm aims at forcing x^k to enter the information zone, which is a minimal requirement if we hope for a good behavior of our unfeasible schemes.

Adaptive Lagrangian-Based mUltiplier Method – ALBUM

1. Input: $\mathcal{A}_\rho$ a Lagrangian algorithmic map.
2. Initialization: Fix $\delta, \rho_0 > 0$ and start with any $\left(x^0, u^0, y^0\right) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$.
3. For each $k = 0, 1, \ldots$ generate a sequence $\left\{\left(x^k, u^k, y^k\right)\right\}_{k\in\mathbb{N}}$ as follows:

   3.1. Primal step
   $$\left(x^{k+1}, u^{k+1}\right) \in \mathcal{A}_{\rho_k}\left(x^k, u^k, y^k\right). \tag{3.6}$$

   3.2. Multiplier step
   $$y^{k+1} = y^k + \rho_k\left(F\left(x^{k+1}\right) - u^{k+1}\right). \tag{3.7}$$

   3.3. Adaptive step: choose $\tau \in \left(0, \frac{a}{2}\right)$ and set $\beta_k := \frac{b^2}{\rho_k \gamma}$. If $x^{k+1} \notin \mathcal{Z}$ or
   $$\tau\left\|x^{k+1} - x^k\right\|^2 > \mathcal{E}_{\beta_k}\left(x^k, u^k, y^k, x^{k-1}\right) - \mathcal{E}_{\beta_k}\left(x^{k+1}, u^{k+1}, y^{k+1}, x^k\right), \tag{3.8}$$
   set $\rho_{k+1} = \rho_k + \delta$. Otherwise, set $\rho_{k+1} = \rho_k$.

The relations between a, b, the penalty parameters sequence $\{\rho_k\}_{k\in\mathbb{N}}$ and other data input constants will be made more precise whence we develop our analytic framework in Section 4.

We record here a simple consequence which will be useful in our analysis that immediately follows from the definitions of ρ_k and β_k (see Step 3.3):

$$\rho_k \geq \rho_0 > 0 \quad \text{and} \quad \beta_k \leq \beta_0, \ \text{for all } k \in \mathbb{N}. \tag{3.9}$$

Remark 3.1. In some cases the penalty parameters ρ_k, $k \in \mathbb{N}$, can be adjusted so that Step 3.3 automatically holds with $\rho_k = \rho$ for all $k \in \mathbb{N}$. In this case the iterations boils down to Steps 3.1 and 3.2 only. This will happen for instance in the case when the information zone is the whole space, *e.g.*, when F is linear (*cf.* Remark 2.2 and Remark 4.3 below).

3.3 A methodology for Lagrangian based methods.

First note that, once the input Lagrangian algorithmic map $\mathcal{A}_\rho$ is chosen, **ALBUM** generates a sequence $\left\{z^k\right\}_{k\in\mathbb{N}} := \left\{\left(x^k, u^k, y^k\right)\right\}_{k\in\mathbb{N}}$, which thanks to Definition 3.1, must satisfy the following two conditions

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

C2 There exists a positive constant b such that

$$\left\|\nabla_x \mathcal{L}^{\sharp}_{\rho_k}\left(x^{k+1}, u^{k+1}, y^k\right)\right\| \leq b\left\|x^{k+1} - x^k\right\|, \quad \forall\ k \geq 0.$$

Independently of the algorithmic map $\mathcal{A}_\rho$ which governs the mechanism of a primal black-box, we also need two additional assumptions on the corresponding generated sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ which we record now:

C3 There exists a positive constant c such that

$$\left\|v^{k+1}\right\| \leq c\left\|x^{k+1} - x^k\right\|, \quad \forall\ k \geq 0,$$

for some $v^{k+1} \in \partial_u \mathcal{L}^{\sharp}_{\rho_k}\left(x^{k+1}, u^{k+1}, y^k\right)$.

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

Some comments are now in order. First, note that the proposed methodology, while similar in spirit, is fundamentally different from the general methodology recently proposed in [13], which is unfortunately not applicable for **ALBUM**, due to the primal-dual structure of this scheme. In particular,

- The first condition **C1** is a *partial descent property* on $\mathcal{L}^{\sharp}_{\rho}(\cdot)$. It pertains to the primal variables (x, u), since by nature the dual variable y is an "ascent variable". The dissymmetry between x and u in the descent condition could be removed by further generalizing our approach. For the sake of simplicity, we only consider the case when the quantity of decrease in x is known.
- Conditions **C2** and **C3** provide subgradient bounds for $\mathcal{L}^{\sharp}_{\rho}(\cdot)$ with respect to the primal variables.
- The sequential assumption on h, that is, condition **C4**, is a minimal and extremely weak requirement. This property holds for instance when $h : \operatorname{dom} h \to \mathbb{R}$ is continuous.

From now on, and through the rest of this paper we adopt the following terminology:

> A sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ which is generated by **ALBUM** and satisfies conditions **C1**–**C4** is called a *Lagrangian sequence*.

As we shall see soon, many fundamental Lagrangian based methods produce Lagrangian sequences. This allows us to derive convergence results in a unified way for such methods and their variants. We postpone the description of these methods to Section 6, and we announce next, our main convergence results for **ALBUM**, which will be proved in the following sections.

3.4 Main convergence results for ALBUM.

Our central theoretical contributions on the convergence of **ALBUM** to a critical point of problem (CM) are stated in the following two results.

Theorem 3.1 (Subsequence convergence)**.** *Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a bounded Lagrangian sequence and let $(\bar{x}, \bar{u}, \bar{y})$ be a limit point of $\left\{z^k\right\}_{k\in\mathbb{N}}$. Then $\bar{x}$ is a critical point of the original problem* (CM)*.*

Considering semi-algebraic or definable data, and relying on the so-called nonsmooth KL property [11], we can rule out oscillatory behaviors and establish the global convergence of the whole sequence.

Theorem 3.2 (Global convergence)**.** *Under the premises of Theorem 3.1, and assuming that f_0, F and h are semi-algebraic, the whole sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ converges to a point $(\bar{x},\bar{u},\bar{y})$ such that $\bar{x}$ is a critical point of problem* (CM)*.*

Remark 3.2. (i) Standard arguments show that convergence rates of the sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ of the type $O\left(k^{-s}\right)$ could be established with $s>0$. We refer to the technique in [2].

(ii) The essential tools for convergence are elementary stability questions and the nonsmooth Kurdyka-Łojasiewicz inequality, and thus semi-algebraicity can be replaced by definability in a o-minimal structure on $\mathbb{R},+,\times$.

The next section develops our analytically framework. We present the main ideas underlying the proposed algorithm, the main obstacles that need to be addressed, and the key tools necessary for developing the convergence analysis of **ALBUM**.

4 A key lemma: penalty parameter stabilization.

In this section, we establish a central result which is essential in our approach. It asserts that the sequence of penalty parameters $\{\rho_k\}_{k\in\mathbb{N}}$ becomes stationary and that the information zone $\mathcal{Z}$ is reached within finitely many steps. To establish this result, we provide in a preliminary subsection some simple but yet fundamental properties.

4.1 Fundamental properties of Lagrangian sequences.

The first elementary result identifies when an iterate enters the information zone $\mathcal{Z}$.

Lemma 4.1 (Information lemma)**.** *Let $\mathcal{Z}$ be a given information zone. Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a Lagrangian sequence and assume that the multiplier sequence $\left\{y^k\right\}_{k\in\mathbb{N}}$ is bounded. Then, there exists an index $k_{\text{info}}\in\mathbb{N}$, such that $x^k\in\mathcal{Z}$ for all $k\geq k_{\text{info}}$.*

Proof. We argue by contradiction and assume that $x^k\notin\mathcal{Z}$ for $k\in I$ where I is an infinite set. On one hand, by the definition of the information zone $\mathcal{Z}$, we have for all $k\in I$ that

$$\operatorname{dist}\left(F\left(x^k\right),\operatorname{dom}h\right)>\bar{d}. \tag{4.1}$$

On the other hand, for all $k\in\mathbb{N}$ we have

$$\begin{aligned}
\operatorname{dist}\left(F\left(x^k\right),\operatorname{dom}h\right) &= \inf_{u\in\operatorname{dom}h}\left\|u-F\left(x^k\right)\right\| \\
&\leq \left\|u^k-F\left(x^k\right)\right\| && \left[u^k\in\operatorname{dom}h\right] \\
&= \frac{1}{\rho_{k-1}}\left\|y^k-y^{k-1}\right\| && [(3.7)] \\
&\leq \frac{M}{\rho_{k-1}}. && \left[\left\{y^k\right\}_{k\in\mathbb{N}}\text{ is assumed bounded}\right]
\end{aligned}$$

By Step 3.3 of the algorithm and the fact that I is an infinite set, it follows that $\rho_k \to \infty$ as $k \to \infty$, thus there exists $k_{\text{info}} \in \mathbb{N}$ such that

$$\operatorname{dist}\left(F(x^k), \operatorname{dom} h\right) \leq \frac{M}{\rho_k} \leq \bar{d}, \quad \forall\ k \geq k_{\text{info}},$$

which obviously contradicts (4.1). □

The next result provides an important relation on the sequences $\left\{x^k\right\}_{k\in\mathbb{N}}$ and $\left\{y^k\right\}_{k\in\mathbb{N}}$ produced by **ALBUM** and reflects the min-max dynamics at the root of these methods.

Lemma 4.2. *Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a Lagrangian sequence. The following inequality holds true for any $k \geq 0$*

$$\mathcal{L}^{\sharp}_{\rho_k}\left(x^{k+1}, u^{k+1}, y^{k+1}\right) - \mathcal{L}^{\sharp}_{\rho_k}\left(x^k, u^k, y^k\right) \leq \frac{1}{\rho_k}\left\|y^{k+1} - y^k\right\|^2 - \frac{a}{2}\left\|x^{k+1} - x^k\right\|^2.$$

Proof. From condition **C1**,

$$\mathcal{L}^{\sharp}_{\rho_k}\left(x^{k+1}, u^{k+1}, y^k\right) - \mathcal{L}^{\sharp}_{\rho_k}\left(x^k, u^k, y^k\right) \leq -\frac{a}{2}\left\|x^{k+1} - x^k\right\|^2. \tag{4.2}$$

Using the definition of $\mathcal{L}^{\sharp}_{\rho}$ (*cf.* (2.1)) we have from (3.7) that

$$\begin{aligned}\mathcal{L}^{\sharp}_{\rho_k}\left(x^{k+1}, u^{k+1}, y^{k+1}\right) - \mathcal{L}^{\sharp}_{\rho_k}\left(x^{k+1}, u^{k+1}, y^k\right) &= \left\langle y^{k+1} - y^k, F\left(x^{k+1}\right) - u^{k+1}\right\rangle \\ &= \frac{1}{\rho_k}\left\|y^{k+1} - y^k\right\|^2.\end{aligned}$$

Adding the latter to (4.2) yields the desired result. □

The next result relates the evolution of the multiplier sequence $\left\{y^k\right\}_{k\in\mathbb{N}}$ with that of the primal sequence $\left\{x^k\right\}_{k\in\mathbb{N}}$.

Lemma 4.3. *Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a Lagrangian sequence. Assume that the multiplier sequence $\left\{y^k\right\}_{k\in\mathbb{N}}$ is bounded by some $\Lambda > 0$. Then, the following inequality holds true for any $k \geq k_{\text{info}}$,*

$$\left\|y^{k+1} - y^k\right\|^2 \leq d_1\left\|x^{k+1} - x^k\right\|^2 + d_2\left\|x^k - x^{k-1}\right\|^2, \tag{4.3}$$

where

$$d_1 = \frac{2}{\gamma^2}\left(L(f_0) + L(F)\Lambda + b\right)^2 \quad \text{and} \quad d_2 = \frac{2b^2}{\gamma^2}. \tag{4.4}$$

Proof. For convenience, we define

$$\Delta_k := \nabla F\left(x^{k+1}\right)^T y^{k+1} - \nabla F\left(x^k\right)^T y^k.$$

Then, by Lemma 4.1 and Assumption B(i) and (ii) which warrants that F is uniform regular on $\mathcal{Z}$ with constant γ and ∇F is Lipschitz continuous on $\mathcal{Z}$, respectively, it follows for all $k \geq k_{\text{info}}$ that

$$\begin{aligned}\|\Delta_k\| &= \left\|\nabla F\left(x^{k+1}\right)^T\left(y^{k+1} - y^k\right) + \left(\nabla F\left(x^{k+1}\right) - \nabla F\left(x^k\right)\right)^T y^k\right\| \\ &\geq \gamma\left\|y^{k+1} - y^k\right\| - L(F)\Lambda\left\|x^{k+1} - x^k\right\|.\end{aligned} \tag{4.5}$$

On the other hand, from the definition of $\mathcal{L}_{\rho}^{\sharp}$ (see (2.1)), we have that

$$\begin{aligned}\nabla_x \mathcal{L}_{\rho_k}^{\sharp}\left(x^{k+1}, u^{k+1}, y^k\right) &= \nabla f_0\left(x^{k+1}\right) + \nabla F\left(x^{k+1}\right)^T \left(y^k + \rho_k\left(F\left(x^{k+1}\right) - u^{k+1}\right)\right) \\ &= \nabla f_0\left(x^{k+1}\right) + \nabla F\left(x^{k+1}\right)^T y^{k+1},\end{aligned}$$

where the second equality uses the the multiplier update given in (3.7). Thus, using the latter, thanks to condition **C2** we obtain for all $k \geq 0$ that there exists $b > 0$ such that

$$\left\|\nabla f_0\left(x^{k+1}\right) + \nabla F\left(x^{k+1}\right)^T y^{k+1}\right\| \leq b\left\|x^{k+1} - x^k\right\|. \tag{4.6}$$

Therefore, we obtain for all $k \geq k_{\text{info}}$,

$$\begin{aligned}\|\Delta_k\| &= \left\|\nabla F\left(x^{k+1}\right)^T y^{k+1} - \nabla F\left(x^k\right)^T y^k\right\| \\ &= \left\|\nabla F\left(x^{k+1}\right)^T y^{k+1} + \nabla f_0\left(x^{k+1}\right) - \nabla F\left(x^k\right)^T y^k - \nabla f_0\left(x^k\right) + \nabla f_0\left(x^k\right) - \nabla f_0\left(x^{k+1}\right)\right\| \\ &\leq \left\|\nabla F\left(x^{k+1}\right)^T y^{k+1} + \nabla f_0\left(x^{k+1}\right)\right\| + \left\|\nabla F\left(x^k\right)^T y^k + \nabla f_0\left(x^k\right)\right\| \\ &+ \left\|\nabla f_0\left(x^{k+1}\right) - \nabla f_0\left(x^k\right)\right\| \\ &\leq \left(L(f_0) + b\right)\left\|x^{k+1} - x^k\right\| + b\left\|x^k - x^{k-1}\right\|,\end{aligned} \tag{4.7}$$

where the last inequality uses (4.6), and the Lipschitz continuity of ∇f_0 over $\mathcal{Z}$ (see Assumption B(iii)). Combining (4.5) and (4.7), we thus obtain for any $k \geq k_{\text{info}}$

$$\gamma\left\|y^{k+1} - y^k\right\| \leq \left(L(f_0) + L(F)\Lambda + b\right)\left\|x^{k+1} - x^k\right\| + b\left\|x^k - x^{k-1}\right\|. \tag{4.8}$$

Therefore, squaring the last inequality and using the fact that $(r+s)^2 \leq 2r^2 + 2s^2$ for all $r, s \in \mathbb{R}$, the claimed assertion follows. □

4.2 Finite stabilization of the penalty sequence $\{\rho_k\}_{k\in\mathbb{N}}$.

We are now ready to establish the promised key lemma which asserts that the sequence of penalizing parameters $\{\rho_k\}_{k\in\mathbb{N}}$ becomes stationary from a certain iteration-index k_{statio}. A "Lyapunov zone" for $\mathcal{E}_\beta$ is thus reached within finitely many steps.

Lemma 4.4 (Finite stabilization of the sequence $\{\rho_k\}_{k\in\mathbb{N}}$)**.** *Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a Lagrangian sequence. Assume that the multiplier sequence $\left\{y^k\right\}_{k\in\mathbb{N}}$ is bounded. Then, there exists an index $k_{\text{statio}} \in \mathbb{N}$ such that*

$$\rho_k = \rho_{k_{\text{statio}}}, \quad \forall\ k \geq k_{\text{statio}}.$$

Moreover, for all $k \geq k_{\text{statio}}$ we have $x^k \in \mathcal{Z}$, and there exists $\tau > 0$ such that

$$\tau\left\|x^{k+1} - x^k\right\|^2 \leq \mathcal{E}_{\beta_{k_{\text{statio}}}}\left(x^k, u^k, y^k, x^{k-1}\right) - \mathcal{E}_{\beta_{k_{\text{statio}}}}\left(x^{k+1}, u^{k+1}, y^{k+1}, x^k\right). \tag{4.9}$$

Proof. Lemma 4.1 warrants that $x^k \in \mathcal{Z}$ for all $k \geq k_{\text{info}}$ and by applying Lemma 4.2, we obtain for all $k \geq 0$ that

$$\mathcal{L}^{\sharp}_{\rho_k}\left(z^k\right) - \mathcal{L}^{\sharp}_{\rho_k}\left(z^{k+1}\right) \geq \frac{a}{2}\left\|x^{k+1} - x^k\right\|^2 - \frac{1}{\rho_k}\left\|y^{k+1} - y^k\right\|^2. \tag{4.10}$$

Using Lemma 4.3, we get for all $k \geq k_{\text{info}}$,

$$\left\|y^{k+1} - y^k\right\|^2 \leq d_1\left\|x^{k+1} - x^k\right\|^2 + d_2\left\|x^k - x^{k-1}\right\|^2, \tag{4.11}$$

where d_1 and d_2 are given in (4.4). Hence, by combining (4.10) and (4.11), it follows for all $k \geq k_{\text{info}}$, that

$$\mathcal{L}^{\sharp}_{\rho_k}\left(z^k\right) - \mathcal{L}^{\sharp}_{\rho_k}\left(z^{k+1}\right) \geq \left(\frac{a}{2} - \frac{d_1}{\rho_k}\right)\left\|x^{k+1} - x^k\right\|^2 - \frac{d_2}{\rho_k}\left\|x^k - x^{k-1}\right\|^2. \tag{4.12}$$

Using the definition of $\mathcal{E}_\beta$ (see (3.1)) and setting $\beta := \beta_k$ for all $k \geq 0$, we get

$$\begin{aligned} V_k &:= \mathcal{E}_{\beta_k}\left(x^k, u^k, y^k, x^{k-1}\right) - \mathcal{E}_{\beta_k}\left(x^{k+1}, u^{k+1}, y^{k+1}, x^k\right) \\ &= \mathcal{L}^{\sharp}_{\rho_k}\left(z^k\right) - \mathcal{L}^{\sharp}_{\rho_k}\left(z^{k+1}\right) + \beta_k\left\|x^k - x^{k-1}\right\|^2 - \beta_k\left\|x^{k+1} - x^k\right\|^2. \end{aligned} \tag{4.13}$$

Therefore, with (4.12), we deduce that for all $k \geq k_{\text{info}}$

$$\begin{aligned} V_k &\geq \left(\frac{a}{2} - \frac{d_1}{\rho_k} - \beta_k\right)\left\|x^{k+1} - x^k\right\|^2 - \left(\frac{d_2}{\rho_k} - \beta_k\right)\left\|x^{k-1} - x^k\right\|^2 \\ &= \left(\frac{a}{2} - \frac{d_1}{\rho_k} - \beta_k\right)\left\|x^{k+1} - x^k\right\|^2, \end{aligned} \tag{4.14}$$

where the equality follows from the definition of β_k given in Step 3.3 of **ALBUM**. Hence, using (4.3), we get that

$$\beta_k = \frac{d_2}{\rho_k} = \frac{2b^2}{\rho_k \gamma^2}.$$

In addition, one has for all $k \geq k_{\text{info}}$ that

$$\frac{a}{2} - \frac{d_1}{\rho_k} - \beta_k = \frac{a}{2} - \frac{d_1 + d_2}{\rho_k}. \tag{4.15}$$

Thus (4.14) rewrites

$$V_k \geq \left(\frac{a}{2} - \frac{d_1 + d_2}{\rho_k}\right)\left\|x^{k+1} - x^k\right\|^2. \tag{4.16}$$

The sequence $\{\rho_k\}_{k\in\mathbb{N}}$ cannot increase indefinitely else we would get from (4.16) that

$$\mathcal{E}_{\beta_k}\left(x^k, u^k, y^k, x^{k-1}\right) - \mathcal{E}_{\beta_k}\left(x^{k+1}, u^{k+1}, y^{k+1}, x^k\right) \geq \tau\left\|x^{k+1} - x^k\right\|^2,$$

for all k sufficiently large, where $\tau > 0$ is the parameter given in the **ALBUM** scheme. Thus we obtain the existence of an iteration-index $k_{\text{statio}} \geq k_{\text{info}}$ such that $\rho_k = \rho_{k_{\text{statio}}}$ for all $k \geq k_{\text{statio}}$, and the desired result follows. □

Remark 4.1 (Adaptive process and the dynamics of $\{\rho_k\}_{k\in\mathbb{N}}$)**.** Lemma 4.4 establishes that **ALBUM**, within Step 3.3, relies on two fundamental tests:

– a weak[3] feasibility test, *i.e.*, $x^k \in \mathcal{Z}$,

– a surrogate[4] descent test for $\mathcal{E}_\beta$ which implicitly tunes the algorithm to match the natural step-sizes attached to f_0 and F.

Lemma 4.4 tells us that ρ_k can be automatically tuned to an acceptable value $\rho_{k_{\text{statio}}}$ in finitely many steps. As a consequence, and it is a fundamental fact, we have the descent property:

$$\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(x^k, u^k, y^k, x^{k-1}\right) - \mathcal{E}_{\beta_{k_{\text{statio}}}}\left(x^{k+1}, u^{k+1}, y^{k+1}, x^k\right) \geq \tau \left\|x^{k+1} - x^k\right\|^2, \quad \forall\ k \geq k_{\text{statio}}.$$

In short and to conclude, one could say that the adaptive protocol leads to the finite identification of the information zone and to a sufficient descent property.

Remark 4.2. One observes from the proof, that the descent property on $\mathcal{E}_\beta$ is ensured once we know that

$$\frac{a}{2} - \frac{d_1 + d_2}{\rho_k} > \tau, \quad \forall\ k \geq k_{\text{info}}. \tag{4.17}$$

In order to shunt the surrogate descent test, it is thus tempting to fix a value ρ_0 a priori (before running the method), so that the above holds directly. Yet it is important to understand that this cannot be done in general, since d_1 (*cf.* (4.4)) is a constant that *depends* on a bound Λ of the sequence $\left\{y^k\right\}_{k\in\mathbb{N}}$ which by itself depends on $\{\rho_k\}_{k\in\mathbb{N}}$!

Remark 4.3 (Special case with F assumed to be linear)**.** (i) In that case the dependence of d_1 with Λ given in Lemma 4.3 *disappears*. This allows for a more direct and simplified approach. Indeed, exploiting the linearity of F, the inequality (4.5) reduces to $\|\Delta_k\| \geq \gamma \left\|y^{k+1} - y^k\right\|$ for all $k \geq 0$, where here $\gamma \equiv \sqrt{\lambda_{\min}(FF^T)} > 0$, *cf.* Remark 2.2. Therefore, the boundedness of $\left\{y^k\right\}_{k\in\mathbb{N}}$ is not needed, and it immediately follows that the proof of inequality (4.3) holds true in Lemma 4.3 for all $k \geq 0$, with

$$d_1 = \frac{2}{\lambda_{\min}(FF^T)}\left(L(f_0) + b\right)^2 \quad \text{and} \quad d_2 = \frac{2b^2}{\lambda_{\min}(FF^T)}. \tag{4.18}$$

Secondly, as mentioned before (*cf.* Remark 2.2) the information zone can be taken as the whole space *i.e.*, $\mathcal{Z} \equiv \mathbb{R}^n$, and in that case the adaptive regime is not anymore necessary. Thus we set $\rho_k \equiv \rho > 0$ for all $k \in \mathbb{N}$, and Step 3.3 of **ALBUM** is simply removed (see also Remark 3.1). Therefore, in order to guarantee sufficient descent of the Lyapunov $\mathcal{E}_\beta$, all we need is that (4.17) holds true, that is (with $\tau = 0$), it reduces to

$$\rho > \bar{\rho} := \frac{2\left(d_1 + d_2\right)}{a}, \tag{4.19}$$

where d_1 and d_2 are given in (4.18). Therefore, in the special linear case, this allows for determining explicitly the threshold value $\bar{\rho}$, for a chosen Lagrangian algorithmic map $\mathcal{A}_\rho$ which provides the constants a and b and to obtain the corresponding convergence results via a straightforward application of Theorems 3.1 and 3.2.

(ii) Interestingly, this also provides a positive answer to a question posed in [24, Remark 4(3) p. 2451], where the authors pointed out that it would be interesting to see if global convergence of a proximal ADM could be derived; see also Section 6 for more results.

[3] Weak because we do not ask for actual feasibility.

[4] Surrogate because we do not ask for the augmented Lagrangian function $\mathcal{L}_\rho^\sharp$ to be Lyapunov, but rather that the auxiliary function $\mathcal{E}_\beta$ is Lyapunov.

5 Proof of the main convergence results.

Equipped with the results we have established, we can now apply our methodology to prove the main convergence results of **ALBUM** announced in Section 3.4.

5.1 Subgradient bound for the Lyapunov function $\mathcal{E}_\beta$.

As mentioned previously, we work with the function $\mathcal{E}_\beta$ to overcome the descent obstacle and to detect hidden descent mechanisms. Now the third condition **C3** of our methodology comes into a play. We derive below an upper bound on a subgradient of the Lyapunov function $\mathcal{E}_\beta$.

Lemma 5.1. *Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a bounded Lagrangian sequence. Then, for each $k\in\mathbb{N}$, there exist positive constants σ_1 and σ_2 together with $q^{k+1}\in\partial\mathcal{E}_{\beta_k}\left(x^{k+1},u^{k+1},y^{k+1},x^k\right)$, such that for all $k\geq k_{\text{info}}$*

$$\left\|q^{k+1}\right\|\leq\sigma_1\left\|x^{k+1}-x^k\right\|+\sigma_2\left\|x^k-x^{k-1}\right\|. \tag{5.1}$$

Proof. Consider the quadruplet $q^{k+1}=\left(q_1^{k+1},q_2^{k+1},q_3^{k+1},q_4^{k+1}\right)\in\partial\mathcal{E}_{\beta_k}\left(x^{k+1},u^{k+1},y^{k+1},x^k\right)$. Using the definition of $\mathcal{E}_\beta$ (see (3.1)), subdifferential calculus rules, and recalling the multiplier update rule (3.7), a direct computation shows that:

$$\begin{aligned}
q_1^{k+1}&=\nabla_x\mathcal{L}_{\rho_k}^\sharp\left(x^{k+1},u^{k+1},y^{k+1}\right)+2\beta_k\left(x^{k+1}-x^k\right)\\
&=\nabla_x\mathcal{L}_{\rho_k}^\sharp\left(x^{k+1},u^{k+1},y^k\right)+\nabla F\left(x^{k+1}\right)^T\left(y^{k+1}-y^k\right)+2\beta_k\left(x^{k+1}-x^k\right), &(5.2)\\
q_2^{k+1}&\in\partial_u\mathcal{L}_{\rho_k}^\sharp\left(x^{k+1},u^{k+1},y^{k+1}\right)=\partial_u\mathcal{L}_{\rho_k}^\sharp\left(x^{k+1},u^{k+1},y^k\right)-\left(y^{k+1}-y^k\right), &(5.3)\\
q_3^{k+1}&=\nabla_y\mathcal{L}_{\rho_k}^\sharp\left(x^{k+1},u^{k+1},y^{k+1}\right)=F\left(x^{k+1}\right)-u^{k+1}=\rho_k^{-1}\left(y^{k+1}-y^k\right), &(5.4)\\
q_4^{k+1}&=2\beta_k\left(x^k-x^{k+1}\right). &(5.5)
\end{aligned}$$

Since $\left\{x^k\right\}_{k\in\mathbb{N}}$ is assumed bounded and ∇F is continuous (see Assumption B(ii)) it follows that there exists $B>0$ such that

$$\sup_{k\geq k_{\text{info}}}\left\|\nabla F\left(x^k\right)\right\|\leq B. \tag{5.6}$$

Moreover, recall that from (3.9), we have $\rho_k\geq\rho_0$ and $\beta_k\leq\beta_0$ for all $k\in\mathbb{N}$. Therefore, using condition **C2** and the expressions for q_j^{k+1}, $j=1,2,3,4$ derived above, we get the following estimates:

$$\begin{aligned}
\left\|q_1^{k+1}\right\|&\leq\left\|\nabla_x\mathcal{L}_{\rho_k}^\sharp\left(x^{k+1},u^{k+1},y^k\right)\right\|+B\left\|y^{k+1}-y^k\right\|+2\beta_0\left\|x^{k+1}-x^k\right\|\\
&\leq b\left\|x^{k+1}-x^k\right\|+B\left\|y^{k+1}-y^k\right\|+2\beta_0\left\|x^{k+1}-x^k\right\|\\
&=B\left\|y^{k+1}-y^k\right\|+(b+2\beta_0)\left\|x^{k+1}-x^k\right\|.
\end{aligned}$$

Likewise, thanks to condition **C3** we have with $v^{k+1}\in\partial_u\mathcal{L}_{\rho_k}^\sharp\left(x^{k+1},u^{k+1},y^k\right)$ that $\left\|v^{k+1}\right\|\leq d\left\|x^{k+1}-x^k\right\|$, and hence by defining $q_2^{k+1}=v^{k+1}-\left(y^{k+1}-y^k\right)$, it immediately follows that $q_2^{k+1}\in\partial_u\mathcal{L}_{\rho_k}^\sharp\left(x^{k+1},u^{k+1},y^{k+1}\right)$, and from (5.3)

$$\left\|q_2^{k+1}\right\|\leq\left\|v^{k+1}\right\|+\left\|y^{k+1}-y^k\right\|\leq d\left\|x^{k+1}-x^k\right\|+\left\|y^{k+1}-y^k\right\|.$$

Finally, from (5.4) and (5.5) we immediately obtain (recall (3.9))

$$\left\|q_3^{k+1}\right\| \leq \frac{1}{\rho_0}\left\|y^{k+1}-y^k\right\| \quad \text{and} \quad \left\|q_4^{k+1}\right\| \leq 2\beta_0\left\|x^{k+1}-x^k\right\|.$$

Therefore, summing these inequalities, we obtain for all $k \geq k_{\text{statio}}$

$$\left\|q^{k+1}\right\| \leq \sum_{j=1}^{4}\left\|q_j^{k+1}\right\| \leq \left(B+1+\frac{1}{\rho_0}\right)\left\|y^{k+1}-y^k\right\| + (4\beta_0+b+d)\left\|x^{k+1}-x^k\right\|.$$

Using the proof of Lemma 4.3, for all $k \geq k_{\text{statio}}$, we know from (4.8) that

$$\gamma\left\|y^{k+1}-y^k\right\| \leq (L(f_0)+L(F)\Lambda+b)\left\|x^{k+1}-x^k\right\| + b\left\|x^k-x^{k-1}\right\|. \tag{5.7}$$

Combining this with the above inequality yields the desired estimation (5.1) by choosing

$$\sigma_1 = \frac{1}{\gamma}\left(B+1+\frac{1}{\rho_0}\right)(L(f_0)+L(F)\Lambda+b)+4\beta_0+b+d \quad \text{and} \quad \sigma_2 = \frac{b}{\gamma}\left(B+1+\frac{1}{\rho_0}\right).$$

This completes the proof. □

Equipped with Lemma 4.4 we immediately obtain the following result.

Proposition 5.1. *Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a Lagrangian sequence. Assume that the multiplier sequence $\left\{y^k\right\}_{k\in\mathbb{N}}$ is bounded. Then*

$$\sum_{k=1}^{\infty}\left\|x^{k+1}-x^k\right\|^2 < \infty \quad \textit{and} \quad \sum_{k=1}^{\infty}\left\|y^{k+1}-y^k\right\|^2 < \infty.$$

Proof. Invoking Lemma 4.4 which holds true under the stated assumptions, we have that

$$\tau\left\|x^{k+1}-x^k\right\|^2 \leq \mathcal{E}_{\beta_{k_{\text{statio}}}}\left(x^k,u^k,y^k,x^{k-1}\right) - \mathcal{E}_{\beta_{k_{\text{statio}}}}\left(x^{k+1},u^{k+1},y^{k+1},x^k\right), \tag{5.8}$$

for all $k \geq k_{\text{statio}}$. Summing (5.8) over $k = k_{\text{statio}}, k_{\text{statio}}+1, \ldots, k_{\text{statio}}+p$ we obtain

$$\begin{aligned}\tau\sum_{k=k_{\text{statio}}}^{k_{\text{statio}}+p}\left\|x^{k+1}-x^k\right\|^2 &\leq \mathcal{E}_{\beta_{k_{\text{statio}}}}\left(x^1,u^1,y^1,x^0\right) - \mathcal{E}_{\beta_{k_{\text{statio}}}}\left(x^{p+1},u^{p+1},y^{p+1},x^p\right)\\ &\leq \mathcal{E}_{\beta_{k_{\text{statio}}}}\left(x^1,u^1,y^1,x^0\right),\end{aligned}$$

where the last inequality follows from the fact that $\inf_{(x,u)}\mathcal{E}_{\beta_{k_{\text{statio}}}} > -\infty$ (thanks to (2.2) since $\mathcal{E}_{\beta_{k_{\text{statio}}}}(\cdot) \geq \mathcal{L}^{\sharp}_{\rho_{k_{\text{statio}}}}(\cdot)$). Letting $p \to \infty$ yields

$$\sum_{k=1}^{\infty}\left\|x^{k+1}-x^k\right\|^2 < \infty.$$

Therefore, from Lemma 4.3, it also follows that $\sum_{k=1}^{\infty}\left\|y^{k+1}-y^k\right\|^2 < \infty$, as required. □

We are now ready to prove our first convergence result for the generic scheme **ALBUM**.

5.2 Proof of Theorem 3.1 – subsequence convergence.

The sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ is bounded and therefore there exists a subsequence $\{z^{m_k}\}_{k\in\mathbb{N}}$ which converges to $\bar{z}=(\bar{x},\bar{u},\bar{y})$. We first prove that $(\bar{x},\bar{u},\bar{y},\bar{x})$ is a critical point of $\mathcal{E}_{\beta_{k_{\text{statio}}}}$, that is,

$$(0,0,0,0)\in\partial\mathcal{E}_{\beta_{k_{\text{statio}}}}(\bar{x},\bar{u},\bar{y},\bar{x}).$$

Since h is lower semi-continuous we have that

$$\liminf_{k\to\infty} h\left(u^{m_k}\right)\geq h\left(\bar{u}\right),$$

which combined with condition **C4** yields that $h\left(u^{m_k}\right)$ converges to $h\left(\bar{u}\right)$ as $k\to\infty$. Therefore, from Proposition 5.1 and the continuity of f_0 and F (see Assumption B(ii) and (iii)), we obtain that

$$\begin{aligned}\lim_{k\to\infty}\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^{m_k},x^{m_k-1}\right)&=\lim_{k\to\infty}\left[\mathcal{L}^{\sharp}_{\rho_{k_{\text{statio}}}}\left(x^{m_k},u^{m_k},y^{m_k}\right)+\beta_{k_{\text{statio}}}\left\|x^{m_k}-x^{m_k-1}\right\|^2\right]\\&=\mathcal{L}^{\sharp}_{\rho_{k_{\text{statio}}}}\left(\bar{x},\bar{u},\bar{y}\right)\\&=\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(\bar{z},\bar{x}\right).\end{aligned}$$

We know from Lemma 5.1 that there exist $\sigma_1,\sigma_2>0$ and $q^{k+1}\in\partial\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^{k+1},x^k\right)$ for which

$$\left\|q^{k+1}\right\|\leq\sigma_1\left\|x^{k+1}-x^k\right\|+\sigma_2\left\|x^k-x^{k-1}\right\|.$$

On the other hand, from Proposition 5.1 it follows that

$$\lim_{k\to\infty}\left\|x^{k+1}-x^k\right\|=0.$$

Thus $q^{k+1}\to 0$ as $k\to\infty$. Using the closedness property of the graph of the subdifferential $\partial\mathcal{E}_\beta$, we obtain that $(0,0,0,0)\in\partial\mathcal{E}_{\beta_{k_{\text{statio}}}}(\bar{x},\bar{u},\bar{y},\bar{x})$. This shows that $(\bar{x},\bar{u},\bar{y},\bar{x})$ is a critical point of $\mathcal{E}_{\beta_{k_{\text{statio}}}}$. Proposition 3.1 now implies that $\bar{x}$ is a critical point of the objective function f of model (CM), and the proof is completed.

Next, in order to prove the second main global convergence result of our algorithm **ALBUM**, we need to introduce adequate and necessary material on the nonsmooth KL property [11].

Let $\eta\in(0,+\infty]$. We denote by Φ_η the class of all concave and continuous functions $\varphi:[0,\eta)\to\mathbb{R}_+$ which satisfy the following conditions

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

(ii) φ is C^1 on $(0,\eta)$ and continuous at 0;

(iii) for all $s\in(0,\eta)$: $\varphi'(s)>0$.

The next result plays a crucial role, see [13, Lemma 6].

Lemma 5.2 (Uniformized KL property)**.** *Let Ω be a compact set and let $\sigma:\mathbb{R}^d\to(-\infty,\infty]$ be a proper and lower semicontinuous function. Assume that σ is constant on Ω and satisfies the KL property at each point of Ω. Then, there exist $\varepsilon>0$, $\eta>0$ and $\varphi\in\Phi_\eta$ such that for all $\overline{u}$ in Ω and all u in the following intersection*

$$\left\{u\in\mathbb{R}^d:\operatorname{dist}(u,\Omega)<\varepsilon\right\}\cap\left[\sigma(\overline{u})<\sigma(u)<\sigma(\overline{u})+\eta\right],\tag{5.9}$$

one has,

$$\varphi'\left(\sigma(u)-\sigma(\overline{u})\right)\operatorname{dist}\left(0,\partial\sigma(u)\right)\geq 1.\tag{5.10}$$

Equipped with these results we proceed with the proof of the second main theorem, *i.e.*, convergence of the whole sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ to a critical point of problem (CM) with semi-algebraic data f_0, h and F. Note that the technique used below is patterned after the recent work [13]. However, as explained previously, we cannot apply directly these results to **ALBUM**, since the descent requirements stated there clearly do not hold in our framework.

5.3 Proof of Theorem 3.2 – global convergence.

Since $\left\{z^k\right\}_{k\in\mathbb{N}}$ is bounded there exists a subsequence $\{z^{m_k}\}_{k\in\mathbb{N}}$ such that $z^{m_k}\to\bar{z}$ as $k\to\infty$. In a similar way as in Theorem 3.1 we get that

$$\lim_{k\to\infty}\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k,x^{k-1}\right)=\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(\bar{z},\bar{x}\right). \tag{5.11}$$

If there exists an integer $\bar{k}\geq k_{\text{statio}}$ for which $\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^{\bar{k}},x^{\bar{k}-1}\right)=\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(\bar{z},\bar{x}\right)$ then the decreasing property obtained in Lemma 4.4 would imply that $z^{\bar{k}+1}=z^{\bar{k}}$. A trivial induction show then that the sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ is stationary and the announced result is obvious.

Since $\left\{\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k,x^{k-1}\right)\right\}_{k\in\mathbb{N}}$ is a nonincreasing sequence, it is clear from (5.11) that

$$\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(\bar{z},\bar{x}\right)<\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k,x^{k-1}\right)\text{ for all }k\geq k_{\text{statio}}.$$

Again from (5.11), for any $\eta>0$ there exists $k_0\geq k_{\text{statio}}$ such that

$$\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k,x^{k-1}\right)<\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(\bar{z},\bar{x}\right)+\eta,\ \forall\, k>k_0.$$

From Theorem 3.1 we know that $\lim_{k\to\infty}\operatorname{dist}\left(z^k,\omega\left(z^0\right)\right)=0$. This means that for any $\varepsilon>0$ there exists a positive integer $k_1\geq k_{\text{statio}}$ such that $\operatorname{dist}\left(z^k,\omega\left(z^0\right)\right)<\varepsilon$ for all $k>k_1$. Summing up all these facts, we get that z^k belongs to the intersection in (5.9) for all $k>l:=\max\left\{k_0,k_1\right\}\geq k_{\text{statio}}$.

We denote by $\omega\left(z^0\right)$ the set of all limit points. By Theorem 3.1, $\omega\left(z^0\right)$ is nonempty and compact (since by definition, it can viewed as an intersection of compact sets). Now, we show that $\mathcal{E}_{\beta_{k_{\text{statio}}}}$ is finite and constant on $\omega\left(z^0\right)$. Indeed, by our standing assumption (see (2.2)) we know that $\mathcal{L}_\rho^\sharp\left(z^k\right)>-\infty$ for all $k\in\mathbb{N}$, therefore from the definitions of $\mathcal{L}_\rho^\sharp$ and $\mathcal{E}_\beta$ (see (2.1) and (3.1), respectively) we have that $\left\{\mathcal{E}_{\beta_k}\left(z^k,x^{k-1}\right)\right\}_{k\in\mathbb{N}}$ is bounded from below. Lemma 4.4 now guarantees that $\left\{\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k,x^{k-1}\right)\right\}_{k\in\mathbb{N}}$ converges to a finite limit, say l. From (5.11) it follows that $l=\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(\bar{z},\bar{x}\right)$, which proves that $\mathcal{E}_{\beta_{k_{\text{statio}}}}$ is finite and constant on $\omega\left(z^0\right)$.

Thus, since $\mathcal{E}_{\beta_{k_{\text{statio}}}}$ is a KL function, we can apply the Uniformization Lemma 5.2 with $\Omega=\omega\left(z^0\right)$. Therefore, for any $k>l$, we have

$$\varphi'\left(\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k,x^{k-1}\right)-\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(\bar{z},\bar{x}\right)\right)\cdot\operatorname{dist}\left(0,\partial\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k,x^{k-1}\right)\right)\geq 1. \tag{5.12}$$

This makes sense since we know that $\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k,x^{k-1}\right)>\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(\bar{z},\bar{x}\right)$ for any $k>l\geq k_{\text{statio}}$. Using Lemma 5.1 (recalling that $k_{\text{statio}}\geq k_{\text{info}}$), we get that

$$\begin{aligned}\varphi'\left(\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k,x^{k-1}\right)-\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(\bar{z},\bar{x}\right)\right)&\geq\frac{1}{\operatorname{dist}\left(0,\partial\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k,x^{k-1}\right)\right)}\\&\geq\left(\sigma\left\|x^k-x^{k-1}\right\|+\sigma\left\|x^{k-1}-x^{k-2}\right\|\right)^{-1},\end{aligned} \tag{5.13}$$

where $\sigma = \max\{\sigma_1, \sigma_2\}$ while σ_1 and σ_2 given in Lemma 5.1. On the other hand, from the concavity of φ we get that

$$\varphi\left(\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k, x^{k-1}\right) - \mathcal{E}_{\beta_{k_{\text{statio}}}}(\bar{z}, \bar{x})\right) - \varphi\left(\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^{k+1}, x^k\right) - \mathcal{E}_{\beta_{k_{\text{statio}}}}(\bar{z}, \bar{x})\right) \geq$$
$$\varphi'\left(\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k, x^{k-1}\right) - \mathcal{E}_{\beta_{k_{\text{statio}}}}(\bar{z}, \bar{x})\right)\left(\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^k, x^{k-1}\right) - \mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^{k+1}, x^k\right)\right). \quad (5.14)$$

For convenience, we define for all $p, q \in \mathbb{N}$ and $\bar{z}$ the following quantities

$$\Delta_{p,q} := \varphi\left(\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^p, x^{p-1}\right) - \mathcal{E}_{\beta_{k_{\text{statio}}}}(\bar{z}, \bar{x})\right) - \varphi\left(\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^q, x^{q-1}\right) - \mathcal{E}_{\beta_{k_{\text{statio}}}}(\bar{z}, \bar{x})\right).$$

Combining (5.13) and (5.14) and using Lemma 4.4 yields for any $k > l$ that

$$\Delta_{k,k+1} \geq \frac{\tau \left\|x^{k+1} - x^k\right\|^2}{\psi\left(\|x^k - x^{k-1}\| + \|x^{k-1} - x^{k-2}\|\right)}, \quad (5.15)$$

and hence

$$\left\|x^{k+1} - x^k\right\|^2 \leq \rho\Delta_{k,k+1}\left(\left\|x^k - x^{k-1}\right\| + \left\|x^{k-1} - x^{k-2}\right\|\right),$$

where $\rho = \sigma/\tau$. Using the fact that $2\sqrt{\alpha\beta} \leq \alpha + \beta$ for all $\alpha, \beta \geq 0$, we infer

$$4\left\|x^{k+1} - x^k\right\| \leq \left\|x^k - x^{k-1}\right\| + \left\|x^{k-1} - x^{k-2}\right\| + 4\rho\Delta_{k,k+1}. \quad (5.16)$$

Let us now prove that for any $k > l$ the following inequality holds

$$2\sum_{i=l+1}^{k}\left\|x^{i+1} - x^i\right\| \leq 2\left\|x^{l+1} - x^l\right\| + \left\|x^l - x^{l-1}\right\| + \rho\Delta_{l+1,k+1}.$$

Summing up (5.16) for $i = l+1, \ldots, k$ yields

$$\begin{aligned}
4\sum_{i=l+1}^{k}\left\|x^{i+1} - x^i\right\| &\leq \sum_{i=l+1}^{k}\left\|x^i - x^{i-1}\right\| + \sum_{i=l+1}^{k}\left\|x^{i-1} - x^{i-2}\right\| + 4\rho\sum_{i=l+1}^{k}\Delta_{i,i+1} \\
&\leq \sum_{i=l+1}^{k}\left\|x^{i+1} - x^i\right\| + \left\|x^{l+1} - x^l\right\| + 4\rho\sum_{i=l+1}^{k}\Delta_{i,i+1} \\
&+ \sum_{i=l+1}^{k}\left\|x^{i+1} - x^i\right\| + \left\|x^{l+1} - x^l\right\| + \left\|x^l - x^{l-1}\right\| \\
&= 2\sum_{i=l+1}^{k}\left\|x^{i+1} - x^i\right\| + 2\left\|x^{l+1} - x^l\right\| + \left\|x^l - x^{l-1}\right\| + 4\rho\Delta_{l+1,k+1},
\end{aligned}$$

where the last inequality follows from the fact that $\Delta_{p,q} + \Delta_{q,r} = \Delta_{p,r}$ for all $p, q, r \in \mathbb{N}$. Since $\varphi \geq 0$, we thus have for any $k > l$ that

$$2\sum_{i=l+1}^{k}\left\|x^{i+1} - x^i\right\| \leq 2\left\|x^{l+1} - x^l\right\| + \left\|x^l - x^{l-1}\right\| + \gamma\varphi\left(\mathcal{E}_{\beta_{k_{\text{statio}}}}\left(z^l, x^{l-1}\right) - \mathcal{E}_{\beta_{k_{\text{statio}}}}(\bar{z}, \bar{x})\right).$$

Since the right hand-side of the inequality above does not depend on k at all, it is easily shows that the sequence $\left\{x^k\right\}_{k\in\mathbb{N}}$ has finite length, that is,

$$\sum_{k=1}^{\infty}\left\|x^{k+1}-x^k\right\|<\infty. \tag{5.17}$$

This means that it is a Cauchy sequence and hence a convergent sequence. In addition, from (4.8) we also have

$$\sum_{k=1}^{\infty}\left\|y^{k+1}-y^k\right\|<\infty,$$

and thus $\left\{y^k\right\}_{k\in\mathbb{N}}$ has also finite length and therefore a convergent sequence. Now, the multiplier Step 3.2 yields, for any $k\geq k_{\text{statio}}$, that

$$u^{k+1}=F\left(x^{k+1}\right)+\frac{1}{\rho_{k_{\text{statio}}}}\left(y^k-y^{k+1}\right).$$

Since F is continuous, $\left\{x^k\right\}_{k\in\mathbb{N}}$ a convergent sequence and thanks to Proposition 5.1 it follows that $\left\{u^k\right\}_{k\in\mathbb{N}}$ is also a convergent sequence. From Theorem 3.1 it is clear that $\left\{\left(z^k,x^{k-1}\right)\right\}_{k\in\mathbb{N}}$ converges to a critical point $(\bar{x},\bar{u},\bar{y},\bar{x})$ of $\mathcal{E}_{\beta_{k_{\text{statio}}}}$. We finally conclude from Proposition 3.1 that $\bar{x}$ is a critical point of f.

6 Applications: specific schemes from ALBUM.

The generic scheme **ALBUM** encompasses interesting Lagrangian based methods. First recall that in any Lagrangian based method, the multiplier update is always given by an *explicit* formula (see (3.7)):

$$y^{k+1}=y^k+\rho_k\left(F\left(x^{k+1}\right)-u^{k+1}\right).$$

Thus, the main computational and algorithmic issues which emerge from **ALBUM** depend on the way we define the Lagrangian algorithmic map $\mathcal{A}_\rho$ to compute the primal step. In general, any minimization algorithm can be used at this stage. We focus on the description of two fundamental types of maps $\mathcal{A}_\rho$, yet we note that other variants can also be conceived depending on the problem's data information and the structure at hand. This point will be further developed below in Section 6.3.

6.1 Two fundamental instances of $\mathcal{A}_\rho$ and the corresponding ALBUM.

Given a triple $\left(x^k,u^k,y^k\right)$ we compute the next primal variables x^{k+1} and u^{k+1} in **ALBUM** via the algorithmic map $\mathcal{A}_\rho$ given by either one of the following minimization schemes:

- **ALBUM 1** – *Joint Minimization ≡ Proximal Multipliers Method [30]*

$$\left(x^{k+1},u^{k+1}\right)\in\operatorname{argmin}_{(x,u)}\left\{\mathcal{L}^{\sharp}_{\rho_k}\left(x,u,y^k\right)+\frac{\mu}{2}\left\|x-x^k\right\|^2\right\},\quad(\mu>0). \tag{6.1}$$

  This simple idea consists in minimizing a proximal counterpart of the augmented Lagrangian $\mathcal{L}^{\sharp}_{\rho}$, jointly with respect to both primal variables x and u, is nothing else but the classical dynamic of Proximal Method of Multipliers (PMM) of Rockafellar [30].

- **ALBUM 2** – *Alternating Minimization (aka Gauss-Seidel) ≡ Proximal ADM [17]*
  Update the variables x and u in an alternating fashion as follows:

$$u^{k+1} \in \operatorname{argmin}_u \mathcal{L}^{\sharp}_{\rho_k}\left(x^k, u, y^k\right), \tag{6.2}$$

$$x^{k+1} \in \operatorname{argmin}_x \left\{ \mathcal{L}^{\sharp}_{\rho_k}\left(x, u^{k+1}, y^k\right) + \frac{\mu}{2}\left\|x - x^k\right\|^2 \right\}, \quad (\mu > 0). \tag{6.3}$$

Remark 6.1. (i) Note that in the above two schemes the proximal regularization term was added only for the primal variable x of the augmented Lagrangian, since by the construction of $\mathcal{L}^{\sharp}_{\rho}$ (see (2.1)), we note that the primal variable u already admit a built-in proximal term.

(ii) Also, note that the flexibility of **ALBUM** provides potential for further studies within other strategies or variants that could be conceived and further developed in future work, *e.g.*, adding a proximal regularization term for u around u^k and performing a subgradient step for determining the next point u^{k+1}; or dropping one of the proximal regularization term in exchange of other assumptions on the problem's data, see section 6.3 for the latter situation.

Remark 6.2 (Tractability of the subproblems)**.** Although the practical aspects involving implementation are beyond the scope of this work, it is important to discuss some of these issues. In this regard we comment the general practicability of the steps of **ALBUM 2** whose alternating structure is often more favorable toward implementation. Recall that **ALBUM 2** features a simple dual step and two primal steps à la Gauss-Seidel, one with respect to u and one with respect to x, we discuss them below:

(i) As already mentioned the u-step, defined through (6.2), reduces to the computation of the proximal mapping of the function h. Thus, this step can be efficiently computed when the proximal map of h is accessible, *i.e.*, via an and explicit formula or via simple computations, see for instance, [26, 13, 6] for interesting examples.

(ii) The second subproblem, namely the x-step, is more involved. Let us discuss two protocols for solving this step approximately. For simplicity, suppose that $f_0 \equiv 0$. Then, the step (6.3) reduces to solve an *unconstrained Nonlinear Least Squares problem*, NLS for short. Therefore, the proposed Lagrangian methodology which allows to reduce the very general constrained nonlinear optimization model (CM) to solving sequentially unconstrained NLS subproblems, provides interesting future research avenues, whereby fundamental methods of NLS could be considered and exploited to analyze inexact variants. Indeed, NLS problems are central in scientific computation, and even though these are nonconvex problems, there exist two well-known fundamental methods: Gauss-Newton and Levenberg-Marquardt, including many of their variants, which address this key computational problem within a very large body of literature, see *e.g.*, [10, 14]; see also the interesting work [21], where SDP relaxations are shown to find global solutions of some unconstrained NLS of polynomial type. Another approach to tackle the x-step is to approximate it through convex subproblems, which can then be efficiently solved. For this, we refer the reader to Section 6.3 where we give further insights into this question, and we also introduce a new and easily implementable version of **ALBUM 2** for (CM-L) problems.

6.2 Convergence results for ALBUM 1 and ALBUM 2.

To apply our main results (*cf.* Section 3), as previously explained, we first need to verify that *joint minimization* and *alternating minimization* satisfy the two conditions of Definition 3.1, *i.e.*, they are Lagrangian algorithmic maps. Recall that following our notations, for a given point $\xi := \xi^k$ at iteration k, the next point ξ^+ stands for ξ^{k+1}.

- **ALBUM 1** – *Joint Minimization*

  From the choice of $\mathcal{A}_\rho$ (see (6.1)) we immediately get

  $$\mathcal{L}_\rho^\sharp\left(x^+, u^+, y\right) + \frac{\mu}{2}\left\|x^+ - x\right\|^2 \leq \mathcal{L}_\rho^\sharp\left(x, u, y\right),$$

  showing that Definition 3.1(i) holds true with $a = \mu$. Moreover, we also obtain

  $$(0,0) \in \left(\nabla_x \mathcal{L}_\rho^\sharp\left(x^+, u^+, y\right) + \mu\left(x^+ - x\right), \partial_u \mathcal{L}_\rho^\sharp\left(x^+, u^+, y\right)\right), \tag{6.4}$$

  hence it follows that Definition 3.1(ii) immediately holds true with $b = \mu$.

- **ALBUM 2** – *Alternating Minimization*

  Thanks to the choice of $\mathcal{A}_\rho$, we get from (6.2) that $\mathcal{L}_\rho^\sharp\left(x, u^+, y\right) \leq \mathcal{L}_\rho^\sharp\left(x, u, y\right)$ and from (6.3) we get that $\mathcal{L}_\rho^\sharp\left(x^+, u^+, y\right) + \frac{\mu}{2}\left\|x^+ - x\right\|^2 \leq \mathcal{L}_\rho^\sharp\left(x, u^+, y\right)$. Combining both inequalities shows that Definition 3.1(i) holds true with $a = \mu$. Moreover, as before it also follows immediately that Definition 3.1(ii) holds true with $b = \mu$.

We will now show that both **ALBUM 1** and **ALBUM 2** generate Lagrangian sequences $\left\{z^k\right\}_{k\in\mathbb{N}}$. To this end we have to verify that conditions **C3** and **C4** hold true for both schemes.

First, for **ALBUM 1** we obtain from (6.1) (*cf.* (6.4)) that $0 =: v^{k+1} \in \partial_u \mathcal{L}_{\rho_k}^\sharp\left(x^{k+1}, u^{k+1}, y^k\right)$, and hence condition **C3** holds true with any $c > 0$. The next result shows that condition **C3** also holds true for **ALBUM 2**.

Proposition 6.1. *Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a sequence generated by* **ALBUM 2** *which is assumed to be bounded. Then, for each $k \in \mathbb{N}$, there exist a positive constant c and $v^{k+1} \in \partial_u \mathcal{L}_{\rho_k}^\sharp\left(x^{k+1}, u^{k+1}, y^k\right)$, such that for all $k \geq k_{\text{statio}}$ we have*

$$\left\|v^{k+1}\right\| \leq c\left\|x^{k+1} - x^k\right\|.$$

Proof. Since $\left\{x^k\right\}_{k\in\mathbb{N}}$ is bounded, and for each $k \geq k_{\text{info}}$, we have that ∇F is Lipschitz continuous on $\mathcal{Z}$ (by Assumption B(ii)), it follows that there exists $B > 0$ such that

$$\sup_{k\geq k_{\text{info}}} \left\|\nabla F\left(x^k\right)\right\| \leq B.$$

From (6.2) we get that

$$0 \in \partial_u \mathcal{L}_{\rho_k}^\sharp\left(x^k, u^{k+1}, y^k\right).$$

Using the definition of $\mathcal{L}_\rho^\sharp$ (see (2.1)) we obtain that

$$\partial_u \mathcal{L}_{\rho_k}^\sharp\left(x^{k+1}, u^{k+1}, y^k\right) = \partial_u \mathcal{L}_{\rho_k}^\sharp\left(x^k, u^{k+1}, y^k\right) + \rho_k\left(F\left(x^k\right) - F\left(x^{k+1}\right)\right).$$

Therefore, using the inclusion just above, we obtain for all $k \in \mathbb{N}$ that

$$v^{k+1} \equiv \rho_k \left(F\left(x^k\right) - F\left(x^{k+1}\right)\right) \in \partial_u \mathcal{L}^\sharp_{\rho_k}\left(x^{k+1}, u^{k+1}, y^k\right),$$

and

$$\left\|v^{k+1}\right\| = \rho_k \left\|F\left(x^{k+1}\right) - F\left(x^k\right)\right\| \leq \rho_{k_{\text{statio}}} B \left\|x^{k+1} - x^k\right\|,$$

where the last inequality follows from the Mean Value Theorem[5] and the fact that $\rho_k \leq \rho_{k_{\text{statio}}}$ for all $k \geq k_{\text{statio}}$ (see Lemma 4.4). This proves that condition **C3** holds true with $c = \rho_{k_{\text{statio}}} B$. □

Having established that the three conditions **C1**, **C2** and **C3** of the basic methodology hold, to apply our main convergence results to **ALBUM 1** and **ALBUM 2**, it remains to verify the validity of the condition **C4** for h. This is done next.

Proposition 6.2. *Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a sequence generated by either* **ALBUM 1** *or* **ALBUM 2**, *which is assumed to be bounded. Let $\bar{z}$ be a limit point of a subsequence $\left\{z^k\right\}_{k\in\mathcal{K}}$ of $\left\{z^k\right\}_{k\in\mathbb{N}}$, then we have that $\limsup_{k\in\mathcal{K}\subset\mathbb{N}} h\left(u^k\right) \leq h\left(\bar{u}\right)$.*

Proof. The sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ is bounded and therefore there exists a subsequence $\{z^{m_k}\}_{k\in\mathbb{N}}$ which converges to $\bar{z} = (\bar{x}, \bar{u}, \bar{y})$.

For **ALBUM 1**: from the x-step we have for all $k \geq k_{\text{statio}}$ that

$$\mathcal{L}^\sharp_{\rho_{k_{\text{statio}}}}\left(x^{k+1}, u^{k+1}, y^k\right) + \frac{\mu}{2}\left\|x^{k+1} - x^k\right\|^2 \leq \mathcal{L}^\sharp_{\rho_{k_{\text{statio}}}}\left(\bar{x}, \bar{u}, y^k\right) + \frac{\mu}{2}\left\|\bar{x} - x^k\right\|^2.$$

We now substitute k by $m_k - 1$ and obtain from the definition of $\mathcal{L}^\sharp_\rho$ (see (2.1)) that

$$\begin{aligned} f_0\left(x^{m_k}\right) + h\left(u^{m_k}\right) + \left\langle y^{m_k-1}, F\left(x^{m_k}\right) - F\left(\bar{x}\right)\right\rangle + \left\langle y^{m_k-1}, \bar{u} - u^{m_k}\right\rangle \\ + \frac{\rho_{k_{\text{statio}}}}{2}\left\|F\left(x^{m_k}\right) - u^{m_k}\right\|^2 \leq f_0\left(\bar{x}\right) + h\left(\bar{u}\right) + \frac{\rho_{k_{\text{statio}}}}{2}\left\|F\left(\bar{x}\right) - \bar{u}\right\|^2 \\ + \frac{\mu}{2}\left\|\bar{x} - x^{m_k-1}\right\|^2. \end{aligned} \tag{6.5}$$

Likewise, for **ALBUM 2**, from the u-step (see (6.2)), we have for all $k \geq k_{\text{statio}}$ that

$$\mathcal{L}^\sharp_{\rho_{k_{\text{statio}}}}\left(x^k, u^{k+1}, y^k\right) \leq \mathcal{L}^\sharp_{\rho_{k_{\text{statio}}}}\left(x^k, \bar{u}, y^k\right).$$

We now substitute k by $m_k - 1$ and obtain from the definition of $\mathcal{L}^\sharp_\rho$ (see (2.1)) that

$$h\left(u^{m_k}\right) + \left\langle y^{m_k-1}, u^{m_k} - \bar{u}\right\rangle + \frac{\rho_{k_{\text{statio}}}}{2}\left\|F\left(x^{m_k-1}\right) - u^{m_k}\right\|^2 \leq h\left(\bar{u}\right) + \frac{\rho_{k_{\text{statio}}}}{2}\left\|F\left(x^{m_k-1}\right) - \bar{u}\right\|^2. \tag{6.6}$$

For each of the just derived inequalities (6.5) and (6.6), letting k goes to ∞ and using the continuity of f_0 and F (see Assumption B(ii) and (iii)), together with Proposition 5.1 (for the case of (6.5)) yields in both cases that

$$\limsup_{k\to\infty} h\left(u^{m_k}\right) \leq h\left(\bar{u}\right),$$

and the proof is completed. □

To summarize at this point, we have therefore shown that the two main schemes **ALBUM 1** and **ALBUM 2** produce *Lagrangian sequences* and hence our convergence results Theorems 3.1 and 3.2 are applicable. Observe that we do not only prove that these well-known methods converge in the absence of convexity for the general nonlinear composite model (CM), we also show how to apply them under weak assumptions through the use of a new adaptive regime.

[5]Recall that $\|F(u) - F(v)\| \leq \sup_{\theta\in[0,1]} \|\nabla F(v + \theta(u - v))\| \|u - v\|$, [28, p. 69]

6.3 Towards implementable variants of ALBUM.

To further illustrate the potential benefits and generality of our approach we now consider further specific instances and variants of **ALBUM** under other relevant assumptions on data information which occur in many interesting applications. This allows us to extend some recent results in the literature and even to propose a new scheme.

The classical method of alternating direction of multipliers (ADM). Consider the limiting case of **ALBUM 2** obtained with $\mu \equiv 0$. We recover the classical Alternating Direction of Multipliers (ADM) [17]. Under the additional assumption that the augmented Lagrangian $x \to \mathcal{L}_\rho^\sharp(x,u,y)$ is σ-strongly convex, for any fixed $u, y \in \mathbb{R}^m$, we can obtain global convergence of the ADM to critical points of the *nonlinear* nonconvex composite model (CM). Indeed, in this case **ALBUM 2** yields (recall (6.2) and (6.3)) that

$$\mathcal{L}_\rho^\sharp\left(x,u^+,y\right) \leq \mathcal{L}_\rho^\sharp(x,u,y), \tag{6.7}$$

and

$$\nabla_x \mathcal{L}_\rho^\sharp\left(x^+,u^+,y\right) = 0. \tag{6.8}$$

Now, by the σ-strong convexity of $x \to \mathcal{L}_\rho^\sharp(x,u^+,y)$ together with (6.8) we have that

$$\mathcal{L}_\rho^\sharp\left(x^+,u^+,y\right) + \frac{\sigma}{2}\left\|x^+ - x\right\|^2 \leq \mathcal{L}_\rho^\sharp\left(x,u^+,y\right),$$

and hence from (6.7) it follows that Definition 3.1(i) holds true with $a = \sigma$. Moreover, we also get that $\left\|\nabla_x \mathcal{L}_\rho^\sharp(x^+,u^+,y)\right\| = 0 \leq b\left\|x^+ - x\right\|$, showing that Definition 3.1(ii) immediately holds true with any $b > 0$. Now it is trivial to see that the proofs of conditions **C3** and **C4** as done for the case $\mu > 0$ for **ALBUM 2** remain valid for the case $\mu = 0$. Thus our convergence results apply, and extend the recent result [24, Theorem 4], which uses the same assumption on the Lagrangian, but was valid only for the linear case (*i.e.*, $F(x) \equiv Fx$). Furthermore, for the linear case with a matrix F full row rank, we have $\gamma = \sqrt{\lambda_{\min}(FF^T)} > 0$, and since $a = \sigma$ and b can be any positive number, (*e.g.*, we can set $b = 1$), we immediately obtain the threshold value for ρ (see (4.19) in Remark 4.3) that warrant our convergence results:

$$\rho > \frac{4((L(f_0)+1)^2+1)}{\sigma\lambda_{\min}(FF^T)}.$$

Tractable convex subproblems for ALBUM 2. In relation to Remark 6.2, we focus on the tractability of the x-step (as already mentioned, the u-step is easier for any proximable h). We illustrate here a specific but fundamental aspect of our family of methods through the important case of **ALBUM 2**. In addition to the standing assumptions, we assume that f_0 is C^2 with Lipschitz continuous gradient (for simplicity) and F is linear (so that the information zone is the whole space, *cf.* Remark 2.2). The constant $\rho > 0$ can thus be determined. We observe that for fixed couple (u,y), the function $\mathcal{L}_\rho^\sharp(\cdot,u,y)$ is C^2 whenever u is in $\operatorname{dom} h$ and that its Hessian matrix is given by $x \to \nabla^2 f_0(x) + \rho F^T F$. As a consequence of the Lipschitz continuity assumption of f_0 we have that:

$$\sup_{(x,u,y)\in\mathbb{R}^n\times\mathbb{R}^m\times\operatorname{dom} h}\left\|\nabla_x^2 \mathcal{L}_\rho^\sharp(x,u,y)\right\| \leq L(f_0) + \rho\lambda_{\max}(FF^T). \tag{6.9}$$

Thus, with $\mu = L(f_0) + \rho\lambda_{\max}(FF^T)$, the x-step in **ALBUM 2** consists in minimizing a *convex function* $x \to \mathcal{L}_\rho^\sharp\left(x,u^{k+1},y^k\right) + (\mu/2)\left\|x - x^k\right\|^2$ with *known Lipschitz continuous gradient.*

Solving general semi-algebraic feasibility problems with ALBUM 2. The specialization of **ALBUM 2** to the general feasibility problem described in Example 1.4 provides a new parallel projection method; the details of the easy derivation of the corresponding steps in this case are left to the reader. In view of our general results, the penalty parameter $\rho > 0$ can be determined and no other assumption than semi-algebraicity of the subsets S_i, $i = 1, 2, \ldots, p$, is necessary to obtain global convergence of the methods (under our classical boundedness assumptions)

A simple explicit algorithm: Proximal Linearized Alternating Minimization. We consider here a proximal linearized instance of **ALBUM 2** with proven global convergence results which seems to be new in the literature for the nonconvex composite model. Our setting here is confined to the particular, yet interesting and important case, where in the model (CM):

- The function f_0 has an $L(f_0)$-Lipschitz continuous gradient on $\mathbb{R}^n$.
- The mapping F is linear, namely $F(x) \equiv Fx$ for all $x \in \mathbb{R}^n$, for some matrix $F \in \mathbb{R}^{n\times m}$ with full row rank.

Furthermore, we additionally assume that $\kappa(FF^T) < 2$, where $\kappa(A)$ denotes the condition number of a square matrix A, namely the ratio $\lambda_{\max}(A)/\lambda_{\min}(A)$.

Note that this assumption always holds true whenever FF^T or F^TF is the identity matrix, which often occurs in applications, *e.g.*, in some problems in signal recovery [7].

Recall (cf. Remark 2.2) that under the above hypothesis on the problem's data, Assumption B holds with $\mathcal{Z} \equiv \mathbb{R}^n$, and we also have that $\gamma = \sqrt{\lambda_{\min}(FF^T)} > 0$. The augmented Lagrangian in this case reads (*cf.* (2.1)), for $\rho > 0$, as follows

$$\mathcal{L}^\sharp_\rho(x, u, y) := f_0(x) + h(u) + \langle y, Fx - u\rangle + \frac{\rho}{2}\|Fx - u\|^2.$$

We then consider approximating the x-step in **ALBUM 2** (leaving the u-step untouched) through the following scheme:

- **ALBUM 3** – *Proximal Linearized Alternating Minimization*

$$u^{k+1} \in \operatorname{argmin}_u \mathcal{L}^\sharp_{\rho_k}\left(x^k, u, y^k\right), \tag{6.10}$$

$$x^{k+1} \in \operatorname{argmin}_x \left\{\left\langle x - x^k, \nabla_x \mathcal{L}^\sharp_{\rho_k}\left(x^k, u^{k+1}, y^k\right)\right\rangle + \frac{\mu}{2}\left\|x - x^k\right\|^2\right\}, \quad (\mu > 0). \tag{6.11}$$

Thus, the x-step consists of first linearizing the augmented Lagrangian around a given point and adding a proximal term, which is a common strategy to generate a simpler approximate step (see *e.g.*, [13]), and hence (6.11) is nothing else but *one shot* of an explicit gradient step for minimizing $\mathcal{L}^\sharp_{\rho_k}\left(x, u^{k+1}, y^k\right)$, with an easy explicit formula.

To apply the convergence results of Section 3, we first need to verify that the corresponding algorithmic map $\mathcal{A}_\rho$ of **ALBUM 3** satisfies the two conditions of Definition 3.1 ,*i.e.*, is a Lagrangian algorithmic map. For that purpose, first note that given couple (u, y), the gradient of $\mathcal{L}^\sharp_\rho(x, u, y)$ with respect to x, is the mapping $x \to \nabla f_0(x) + F^Ty + \rho F^T(Fx - u)$, which is a L-Lipschitz continuous mapping, with $L := L(f_0) + \rho\|F\|^2$. Invoking the well known Descent Lemma, it

follows that condition **C1** holds with $a = \mu - L/2$. However, observe that contrary to ALBUM 1 and 2, the constant a depends on ρ through L, and $a > 0$ will be warranted thanks to Lemma 6.2 given below.

Next, using the steps of the corresponding algorithmic map $\mathcal{A}_\rho$, together with the fact that f_0 admits an $L(f_0)$-Lipschitz continuous gradient, one easily verifies that for any $k \geq 0$,

$$\begin{aligned}\left\|\nabla_x \mathcal{L}_\rho^\sharp\left(x^+, u^+, y\right)\right\| &\leq \left\|\nabla_x \mathcal{L}_\rho^\sharp\left(x^+, u^+, y\right) - \nabla_x \mathcal{L}_\rho^\sharp\left(x, u^+, y\right)\right\| + \left\|\nabla_x \mathcal{L}_\rho^\sharp\left(x, u^+, y\right)\right\| \\ &\leq \left(L(f_0) + \rho\left\|F\right\|^2 + \mu\right)\left\|x^+ - x\right\|. \end{aligned} \tag{6.12}$$

This shows that condition **C2** holds true with $b = L(f_0) + \rho\left\|F\right\|^2 + \mu$. In addition, condition **C3** is immediate, since here the optimality condition of the u-step (see (6.10)) implies that

$$\partial_u \mathcal{L}_{\rho_k}^\sharp\left(x^{k+1}, u^{k+1}, y^k\right) \ni v^{k+1} = \rho F\left(x^{k+1} - x^k\right) \;\Rightarrow\; \left\|v^{k+1}\right\| \leq \rho\left\|F\right\|\left\|x^{k+1} - x^k\right\|,$$

showing that condition **C3** holds with $c = \rho\left\|F\right\|$.

Finally, since the u-step in **ALBUM 3** is identical to the one in **ALBUM 2**, the statement and the proof of Proposition 6.2 holds in this case with the same proof (see only the part that related to **ALBUM 2**), and hence condition **C4** holds true in this case too.

Despite the fact that conditions **C1**–**C4** are satisfied it is important to realize that our general theorem does not apply at this stage because both a and b depend on ρ and may become negative if ρ is too large. In order to circumvent this difficulty and obtain the general convergence of the scheme (as in Theorems 3.1 and 3.2), it suffices to guarantee a sufficient descent of the Lyapunov function $\mathcal{E}_\beta$. For this we need that (4.17) holds true (see Remark 4.1(b)) for a couple of well chosen μ and ρ, that is,

$$\frac{a}{2} - \frac{d_1 + d_2}{\rho} > 0. \tag{6.13}$$

For that purpose let us first observe that a stronger version of Lemma 4.3 can be derived. Just follow the same proof by exploiting the linearity of F, and note that the boundedness assumption on the sequence of multipliers $\left\{y^k\right\}_{k\in\mathbb{N}}$ in not anymore needed in that case. We leave the details to the reader, and record this result below.

Lemma 6.1. *Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a Lagrangian sequence. Then, the following inequality holds true for any $k \geq 0$,*

$$\left\|y^{k+1} - y^k\right\|^2 \leq d_1\left\|x^{k+1} - x^k\right\|^2 + d_2\left\|x^k - x^{k-1}\right\|^2, \tag{6.14}$$

where

$$d_1 = \frac{2\left\|\mathcal{M}\right\|^2}{\lambda_{\min}(FF^T)}, \quad d_2 = \frac{2\left(L(f_0) + \left\|\mathcal{M}\right\|\right)^2}{\lambda_{\min}(FF^T)}, \tag{6.15}$$

and $\mathcal{M} := \mu I_n - \rho F^T F$.

Equipped with this result, we now show that we can find positive constants ρ and μ in terms of the problem's data so that (6.13) holds, and hence our convergence results for **ALBUM 3**: Theorems 3.1 and 3.2 with semi-algebraic data, apply.

Lemma 6.2 (Determining threshold value for ρ)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a linear mapping for which $\kappa(FF^T) < 2$. Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a sequence generated by* **ALBUM 3**. *Then, there exists a constant $\bar{\rho}$ such that* (6.13) *holds for any $\rho > \bar{\rho}$, and with $\mu \in (\mu_1, \mu_2)$ for some $\mu_1, \mu_2 > 0$, where both $\bar{\rho}$, μ_1 and μ_2 are given in terms of the problem's data $L(f_0)$ and γ.*

Proof. For convenience we denote $\ell := L(f_0)$. Using Lemma 6.1 and the fact that $a = \mu - \left(\ell + \rho \|F\|^2\right)/2$, in order to satisfy (6.13), we need to find $\rho > 0$ and $\mu > 0$ such that

$$\frac{\mu - \frac{\ell + \rho\|F\|^2}{2}}{2} - \frac{2\|\mathcal{M}\|^2 + 2\left(\ell + \|\mathcal{M}\|\right)^2}{\rho\gamma^2} > 0. \tag{6.16}$$

Rewriting this inequality yields the following equivalent one

$$16\|\mathcal{M}\|^2 + \rho\gamma^2\left(\ell + \rho\|F\|^2 - 2\mu\right) + 8\ell^2 + 16\ell\|\mathcal{M}\| < 0.$$

Since $\mathcal{M} = \mu I - \rho F^T F$, and symmetric we have

$$\|\mathcal{M}\| = \lambda_{\max}(\mathcal{M}) = \lambda_{\max}(\mu I - \rho F^T F) = \lambda_{\max}(\mu I) - \rho\lambda_{\min}(F^T F) = \mu - \rho\gamma^2,$$

where the last equality uses the fact $\lambda_{\min}(F^T F) = \lambda_{\min}(FF^T) = \gamma^2$.

Therefore, defining $t := \mu - \rho\gamma^2 \equiv \|\mathcal{M}\|$, and rearranging terms, the above inequality reduces to show that

$$\psi(t) := 16t^2 - 2\left(\rho\gamma^2 - 8\ell\right)t + \rho\gamma^2\left(\ell + \rho\|F\|^2 - 2\rho\gamma^2\right) + 8\ell^2 < 0. \tag{6.17}$$

Computing the (reduced) discriminant Δ_ψ of the above quadratic function $\psi(\cdot)$ yields

$$\Delta_\psi := \left(\rho\gamma^2 - 8\ell\right)^2 - 16\left(\rho\gamma^2\left(\ell + \rho\|F\|^2 - 2\rho\gamma^2\right) + 8\ell^2\right) = \rho^2\gamma^2\eta - 32\rho\gamma^2\ell - 64\ell^2,$$

where thanks to our assumption $\kappa(FF^T) < 2$, we have $\eta := \left(33\gamma^2 - 16\|F\|^2\right) > 0$. Therefore, (6.17) holds (and hence so does (6.16)), if $\Delta_\psi > 0$ and $t_1 < t < t_2$ where t_1 and t_2 are the zeroes of $\psi(t)$. Some algebra then shows that the latter is satisfied with

$$\rho > \bar{\rho} := \frac{8\ell}{\eta\gamma}\left(2\gamma + \sqrt{4\gamma^2 + \eta}\right),$$

and

$$t_1 \equiv \frac{\left(\rho\gamma^2 - 8\ell\right) - \sqrt{\Delta_\psi}}{16} < t < \frac{\left(\rho\gamma^2 - 8\ell\right) + \sqrt{\Delta_\psi}}{16} \equiv t_2. \tag{6.18}$$

Moreover, since $\|\mathcal{M}\| = \mu - \rho\gamma^2 = t$, we must have $t \geq 0$, and indeed it is easy to check that $t_1 > 0$. Using the relation $\mu = t + \rho\gamma^2$, we can rewrite (6.18) as follows

$$\mu_1 \equiv \frac{\left(17\rho\gamma^2 - 8\ell\right) - \sqrt{\Delta_\psi}}{16} < \mu < \frac{\left(17\rho\gamma^2 - 8\ell\right) + \sqrt{\Delta_\psi}}{16} \equiv \mu_2.$$

and the proof is completed. □

Acknowledgments.

The research of Jérôme Bolte is sponsored by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA9550-15-1-0500 & the FMJH Program Gaspard Monge in optimization and operations research. The research of Shoham Sabach was partially supported by the German Israel Foundation, GIF Grant G-1243-304.6/2014. The research of Marc Teboulle was partially supported by the Israel Science Foundation, ISF Grants 998/12 and 1844/16, and the German Israel Foundation, GIF Grant G-1243-304.6/2014.

References

[1] A. Auslender. An exact penalty method for nonconvex problems covering, in particular, nonlinear programming, semidefinite programming, and second-order cone programming. *SIAM J. Optim.* 25(3): 1732–1759, 2015.

[2] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116(1):5–16, 2009.

[3] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Math. Oper. Res.*, 35(2): 438–457, 2010.

[4] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forwardbackward splitting, and regularized GaussSeidel methods. *Mathematical Programming*, 137(1-2), 91-129, 2013.

[5] A. Beck and Y. C. Eldar. Sparsity constrained nonlinear optimization: optimality conditions and algorithms. *SIAM J. Optim.*, 23(3): 1480–1509, 2013.

[6] A. Beck and N. Hallak. On the minimization over sparse symmetric sets: projections, optimality conditions, and algorithms. *Math. Oper. Res.*, 41(1): 196–223, 2016.

[7] S. Becker, J. Bobin, and E. Candès. NESTA: a fast and accurate first-order method for sparse recovery. *SIAM J. Imaging Sci.*, 4(1): 1–39, 2011.

[8] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods.* Prentice-Hall, Enlewood Cliffs, N. J., 1989.

[9] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Belmont MA: Athena Scientific, originally published by academic press, inc., in 1982 edition, 1996.

[10] A. Björck. *Numerical Methods for Least Squares Problems.* SIAM, Philadelphia, 1996.

[11] J. Bolte, and A. Daniilidis, and A. S. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.*, 17(4):1205–1223, 2007.

[12] J. Bolte and E. Pauwels. Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs. *Math. Oper. Res.*, 41(2):442–465, 2016.

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

[14] G. Chavent. *Nonlinear Least Squares for Inverse Problems.* Scientific Computation, 317. Springer, New York, 2009.

[15] A. Edelman, and T. A. Arias and S. T. Steven. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1998.

[16] M. Fortin and R. Glowinski. *Augmented Lagrangian Methods Applications to the Solution of Boundary Valued Problems.* Elsevier, 1983.

[17] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*. Society for Industrial Mathematics, 1989.

[18] R. Hesse and D. R. Luke. Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. *SIAM J. Optim.*, 23(4): 2397–2419, 2013.

[19] P. C. Haarhoff and J. D. Buys. A new method for the optimization of a nonlinear function subject to nonlinear constraints. *Computer J.*, 13 178–184, 1970.

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

[21] S. Kim and M. Kojima, Solving polynomial least squares problems via semidefinite programming relaxations. *J. Global Optim.*, 46(1):1–23, 2009.

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

[23] A. S. Lewis, D. R. Luke, and J. Malick. Local linear convergence for alternating and averaged nonconvex projections. Foundations of Computational Mathematics, 9(4), 485-513, 2009.

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

[25] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels, in : Les Équations aux Derivées Partielles, pp. 87–89. Éditions du Centre National de la Recherche Scientifique, Paris 1963.

[26] R. Luss and M. Teboulle. Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint. *SIAM Rev.*, 55(1): 65–98, 2013.

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

[28] J. M. Ortega and W. C. Rheinbold. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, Inc. 1970.

[29] M. J. D. Powell, A method for nonlinear constraints in minimization problems, in Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.

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

[31] R. T. Rockafellar, and R. J.-B. Wets. *Variational Analysis*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 317. Springer-Verlag, Berlin, 1998.

[32] 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.

[33] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*, The MIT Press, Cambridge, 2011.