Journal of Optimization Theory and Applications (2022) 193:324–353
https://doi.org/10.1007/s10957-021-01929-5

A Dynamic Alternating Direction of Multipliers for Nonconvex Minimization with Nonlinear Functional Equality Constraints

Eyal Cohen[1] · Nadav Hallak[2] · Marc Teboulle[1]

Received: 28 April 2021 / Accepted: 16 August 2021 / Published online: 13 September 2021

Abstract

This paper studies the minimization of a broad class of nonsmooth nonconvex objective functions subject to nonlinear functional equality constraints, where the gradients of the differentiable parts in the objective and the constraints are only locally Lipschitz continuous. We propose a specific proximal linearized alternating direction method of multipliers in which the proximal parameter is generated dynamically, and we design an explicit and tractable backtracking procedure to generate it. We prove subsequent convergence of the method to a critical point of the problem, and global convergence when the problem's data are semialgebraic. These results are obtained with no dependency on the explicit manner in which the proximal parameter is generated. As a byproduct of our analysis, we also obtain global convergence guarantees for the proximal gradient method with a dynamic proximal parameter under local Lipschitz continuity of the gradient of the smooth part of the nonlinear sum composite minimization model.

Keywords Augmented Lagrangian-based methods · Nonconvex and nonsmooth minimization · Proximal gradient method · Kurdyka-Lojasiewicz property · Global convergence

Mathematics Subject Classification 90C26 · 90C30 · 65K05 · 90C46

Communicated by Boris S. Mordukhovich.

✉ Marc Teboulle
teboulle@tauex.tau.ac.il

Eyal Cohen
eyalcohen1@mail.tau.ac.il

Nadav Hallak
ndvhllk@technion.ac.il

[1] School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel

[2] Faculty of Industrial Engineering and Management, The Technion, Haifa, Israel

1 Introduction

Over the last few years, in many modern applications ranging from machine learning, statistics to image/signal processing (see, e.g., [15,20,27] and references therein), it has been observed that nonconvex models are often more faithful than their convex relaxed/approximate counterparts and provide better quality solutions, even in spite of the obvious difficulties in deriving meaningful convergence guarantees for a nonconvex problem. This led to a strong renewed interest in nonconvex optimization methods and their theoretical properties.

This study addresses a broad class of nonsmooth nonconvex minimization problems with nonlinear functional equality constraints via the fundamental (augmented) Lagrangian-based optimization methodology. The augmented Lagrangian (AL) approach is a centerpiece in algorithmic optimization literature with a long history that can be traced back to the classical works of [18] and [24]. To this day, it provides fertile ground for algorithmic frameworks, with prominent fundamental algorithms such as the Proximal Method of Multipliers (PMM) [25] and the Alternating Direction Method of Multipliers (ADMM) [16,17]. We refer the reader to the monographs [3,5] for a comprehensive survey.

The recent literature on Lagrangian-based methods is quite voluminous and is almost entirely devoted to the convex setting, see e.g., [3–5,11,17,29] and references therein. Unlike the convex setup, the situation in the nonconvex setting is far from being well-understood, especially global analysis of Lagrangian-based methods for general nonlinear models remains scarce and very challenging.

The genuine nonconvex optimization model we study has been barely touched in the literature. In fact, only very recently some progress has been initiated for certain types of nonconvex models with *linear* constraints. Even in this simpler linearly constraints setup, the analysis is challenging and is forced to rely on various types of assumptions; see e.g., [19] and [10] for recent results, as well as relevant references, therein.

To our knowledge, the class of nonsmooth nonconvex optimization models with nonlinear equality constraints studied here within the Lagrangian-based scheme we propose, has not been addressed in the literature. The only closely related model we are aware of is the universal Lagrangian framework studied by [9], in which the authors introduce a novel methodology to conduct global analysis of a nonsmooth nonlinear composite optimization problem, within the framework of a broad Lagrangian-based scheme with global convergence guarantees. The results developed in [9] have revealed the fundamental theoretical difficulties involved in the global analysis of Lagrangian-based methods when studying such an all-embracing nonsmooth nonconvex composite minimization model; see, e.g., [28] for a brief summary of this approach, and the underlying nontrivial assumptions needed in the analysis of [9]. Besides the theoretical front, another central issue is on the tractability side, namely, how to handle the minimization step efficiently while preserving the convergence properties. In this context, the approach suggested by [9] reduces in many scenarios to the optimization of very difficult composite nonlinear problems.

Motivated by the recent general framework developed by [9], we seek to bypass the limitations mentioned above by considering a more specific class of problems, whereby one can refine the theoretical analysis and design tractable minimization steps by further exploiting structures and specific data information. Indeed, here we focus our efforts on a class of nonconvex nonsmooth optimization models with data assumptions which, on the one hand, are sufficiently broad to cover many applications, and on the other hand, allow for the global convergence analysis of more tractable schemes.

The paper is structured as follows. In the next section, we introduce the class of nonconvex nonsmooth problems with nonlinear constraints as well as some mathematical preliminaries on our approach. Section 3 introduces the dynamic alternating directions scheme (described by Algorithm 1), and Sect. 4 analyzes its global and subsequence convergence guarantees. In Sect. 5, we provide an explicit and tractable backtracking procedure (cf. Algorithm 2) to generate the proximal parameter used in Algorithm 1 and prove that it satisfies all the properties required by our analysis. Finally, as an interesting byproduct of our analysis, in Sect. 6, we derive new results on global convergence for the proximal gradient method with a dynamic proximal parameter when applied to the nonconvex sum composite minimization model where the smooth part is locally Lipschitz. Section 7 concludes this work, including a brief discussion on additional algorithmic variants that could be obtained within our approach.

Notation. Unless otherwise specified, our notations are standard and can be found in the classical monograph of [26].

2 Problem Formulation and Mathematical Preliminaries

We consider the nonconvex nonsmooth model under the following blanket assumptions on the problem's data:

$$\min_{u\in\mathbb{R}^n, v\in\mathbb{R}^m} \left\{\Phi(u,v):=f(u)+g(u)+h(v) : F(u)=Gv\right\}, \tag{M}$$

where

- $f:\mathbb{R}^n\to(-\infty,\infty]$ is a proper and lower-semicontinuous (lsc) function and $\inf_{u\in\mathbb{R}^n} f(u) > -\infty$;
- $g:\mathbb{R}^n\to\mathbb{R}$ is a continuously differentiable (C^1) function with a *locally* Lipschitz continuous gradient;
- $h:\mathbb{R}^m\to\mathbb{R}$ is a C^1 function with an L_h-Lipschitz continuous gradient, $L_h>0$;
- $F:\mathbb{R}^n\to\mathbb{R}^q$ is a C^1 mapping with a *locally* Lipschitz continuous Jacobian;
- $G\in\mathbb{R}^{q\times m}$ has full row rank, i.e., the minimal eigenvalue $\lambda_{\min}(GG^T)>0$.

Model (M) encompasses a wide variety of problems arising in disparate statistical and engineering applications see e.g., [15,20,21,27], as well as problems ranging from nonlinear least squares to bi-level optimization. We briefly describe two specific examples.

Example 2.1 (Nonconvex Stackelberg competition) In a Stackelberg competition (see, e.g., [31]), firms optimize their utility function in a predetermined order. In the common scenario, there is a leader firm which optimizes first and a follower firm that optimizes second based on the choices of the leader firm. Consequently, the leader firm takes into account the implications of its decisions on the follower firm. Model (M) captures Stackelberg competitions having the general form: $H_{\text{leader}}(u; v) = f(u)+g(u)+h(v)$, and $H_{\text{follower}}(v; u) = \frac{1}{2}v^T Gv + F(u)^T v$; our model is obtained by applying the first-order optimality conditions to the follower firm's problem. Note that this model can be considered as a bi-level optimization problem.

Example 2.2 (Sparsity-related approximations) Model (M) also captures the interesting problem obtained from setting h to be a continuously differentiable sparsity inducing function, such as the famous Smoothly Clipped Absolute Deviation (SCAD) function proposed by [14]. This leads to a variety of possible sparsity-related problems, and nonlinear regression mimicking the ℓ_1-norm regression. Moreover, noting that G can be chosen as a low row rank matrix, compressed sensing-type instances are also captured by the model.

Model (M) also includes the so-called minimization of the sum composite model.

Example 2.3 (Nonlinear composite model) Letting $g = 0$, and G be the identity map $I_{q\times m}$, the model (M) reduces to the nonconvex sum composite minimization problem:

$$\min\left\{f(u) + h(F(u)) : \; u \in \mathbb{R}^n\right\}.$$

This model was studied in [9], but with the more general h being nonsmooth extended-valued, through a general Lagrangian-based framework.[1] Here, we assume that h is a C^1 function with an L_h-Lipschitz continuous gradient. As we shall see, as a byproduct of our analysis, this allows us to tackle the sum composite problem directly, namely via the so-called proximal gradient scheme [1], yet under the mild locally Lipschitz assumptions and with a dynamic proximal parameter, and obtain global convergence guarantees; see Sect. 6 for details.

Based on a variational analysis of the problem (see [26, Chapters 8 and 10]), we define the following constraint qualification conditions for (M). We denote by $\nabla F(\cdot) \in \mathbb{R}^{q\times n}$ the Jacobian of F.

Definition 2.1 (Constraint qualification) A point $u \in \mathbb{R}^n$ is said to satisfy the constraint qualification [CQ] conditions for problem (M) if the following hold true:

(i) The function f is subdifferential regular at u.
(ii) $\partial^\infty f(u) \cap \operatorname{ran} \nabla F(u)^T = \{0\}$.

For more details regarding regularity, see Definition 7.25 and Corollaries 8.11 and 8.19 in [26]. In Appendix A, we also give a more detailed account on subdifferentials of nonconvex functions which are used here.

The [CQ] conditions provide both smoothness and regularity of the constraint set, and the necessary subdifferential calculus rules which allow us to derive first-order necessary optimality conditions for (M).

[1] It should be mentioned that our general model (M) could formally be reformulated through this more general model. However, by doing so, the specific data information and structure of our model (M) above will be lost and cannot anymore be beneficially exploited.

Lemma 2.1 *(First-order optimality conditions) Let $(u^*, v^*) \in \mathbb{R}^n \times \mathbb{R}^m$ be a local minimizer of (M), where u^* satisfies the [CQ] conditions of Definition 2.1. Then,*

(i) (u^*, v^*) is a feasible point, i.e., $0 = F(u^*) - Gv^*$, and
(ii) there exists $y^* \in \mathbb{R}^q$ such that

$$0 \in \partial f(u^*) + \nabla g(u^*) + \nabla F(u^*)^T y^* \quad \textit{and} \quad 0 = \nabla h(v^*) - G^T y^*. \tag{2.1}$$

The Lagrangian $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^q \to (-\infty, \infty]$ associated with problem (M) is defined by

$$\mathcal{L}(u, v, y) := f(u) + g(u) + h(v) + \langle y, F(u) - Gv \rangle. \tag{2.2}$$

We note that

$$\begin{aligned} \partial \mathcal{L}(u, v, y) = \Big(\partial f(u) + \{\nabla g(u) + \nabla F(u)^T y\} \Big) \\ \times \{\nabla h(v) - G^T y\} \times \{F(u) - G(v)\}, \end{aligned} \tag{2.3}$$

where the subdifferential calculus rules which justify (2.3) always hold and do not depend on the [CQ] conditions (see [26, Exercise 8.8]). Thus, it can be easily verified that the first-order optimality conditions of Lemma 2.1 are equivalent to requiring the point (u^*, v^*, y^*) to be a critical point of the Lagrangian, i.e., $(u^*, v^*, y^*) \in \operatorname{crit} \mathcal{L} := \{(u^*, v^*, y^*) : 0 \in \partial \mathcal{L}(u^*, v^*, y^*)\}$. The main purpose of this work is to devise an algorithm that converges to such a point.

To achieve our goal, we rely on an augmented Lagrangian approach (see, e.g., [2, Chapter 4.2]), in which the Lagrangian associated with (M) (cf. (2.2)) is penalized by a quadratic term, with a penalty parameter $\rho > 0$, to obtain the augmented Lagrangian function

$$\mathcal{L}_\rho(u, v, y) := f(u) + g(u) + h(v) + \langle y, F(u) - Gv \rangle + \frac{\rho}{2} \|F(u) - Gv\|^2, \tag{2.4}$$

where $\|\cdot\|$ denotes the usual Euclidean l_2-norm. It will be convenient in the sequel to distinguish between the smooth and nonsmooth elements of $\mathcal{L}_\rho$. For this purpose, let $\varphi : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^q \to \mathbb{R}$ be defined by

$$\begin{aligned} \varphi(u, v, y) &:= g(u) + \langle y, F(u) - Gv \rangle + \frac{\rho}{2} \|F(u) - Gv\|^2 \\ &= g(u) + \frac{1}{2\rho} \|y + \rho(F(u) - Gv)\|^2 - \frac{1}{2\rho} \|y\|^2, \end{aligned} \tag{2.5}$$

so that

$$\mathcal{L}_\rho(u, v, y) = f(u) + h(v) + \varphi(u, v, y).$$

A key motivation behind the Augmented Lagrangian approach is that any critical point (u^*, v^*, y^*) of $\mathcal{L}_\rho$ satisfies that (u^*, v^*) is a critical point of (M); this can easily be established via the characterization of the critical set of Lagrangian function (2.2); see also Proposition 4.1 in Sect. 4.3.

Algorithm 1 Dynamic linearized alternating direction method of multipliers (DAM)

```
Input: $(u^0, v^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^q, \theta > 0$.
For $k = 0, 1, 2, \ldots$
    1. Generate a proximal parameter $\sigma_{k+1} > 0$;
    2. Set
       $$u^{k+1} \in \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \Phi_k(u) := f(u) + \langle \nabla_u \varphi(u^k, v^k, y^k), u \rangle + \frac{\sigma_{k+1}}{2} \left\| u - u^k \right\|^2 \right\}; \tag{3.1}$$
       $$v^{k+1} = \operatorname{argmin}_{v \in \mathbb{R}^m} \left\{ \Psi_k(v) := \varphi(u^{k+1}, v, y^k) + \langle \nabla h(v^k), v \rangle + \frac{\theta}{2} \left\| v - v^k \right\|^2 \right\}; \tag{3.2}$$
       $$y^{k+1} = y^k + \rho \left( F(u^{k+1}) - G v^{k+1} \right); \tag{3.3}$$
```

3 A Dynamic Alternating Directions of Multipliers

For the purpose of obtaining a critical point of $\mathcal{L}_\rho$, we propose the *Dynamic linearized Alternating direction method of Multipliers (DAM)* described by Algorithm 1. As its name suggests, this method relies on an ADMM-based scheme whereby we use here a specific linearization of the augmented Lagrangian and proximal regularization.

Several comments regarding Algorithm 1 are in order.

Remark 3.1 (On the proximal parameter σ_k) The proximal parameter σ_{k+1} is generated *dynamically* according to the information obtained at the current iterate; hence the dynamic nature of Algorithm 1. The analysis of Algorithm 1 requires that the sequence $\{\sigma_k\}_{k \geq 1}$ will satisfy two conditions: (i) that it will facilitate a descent property with respect to the u-update step; and, (ii) that it will be bounded. In Sect. 5, we provide a detailed backtracking procedure to generate σ_{k+1} in a tractable manner (cf. Algorithm 2) so that both of these conditions are met under our model's assumptions and boundedness of the generated sequence of iterates. However, we emphasize that our convergence guarantees (stated in Sect. 4) are *independent* of the explicit manner in which the proximal parameter is generated.

Remark 3.2 (On the u-update) u-update step (3.1) is in fact a proximal gradient step:

$$\begin{aligned} u^{k+1} &\in \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ f(u) + \langle \nabla_u \varphi(u^k, v^k, y^k), u \rangle + \frac{\sigma_{k+1}}{2} \left\| u - u^k \right\|^2 \right\} \\ &= \operatorname{prox}_{\sigma_{k+1}^{-1} f} \left(u^k - \sigma_{k+1}^{-1} \nabla_u \varphi(u^k, v^k, y^k) \right). \end{aligned}$$

A preliminary requirement for the validity of the u-update step is that the solution set of the minimization problem in (3.1) is nonempty. The results in [26, p. 21], concerning the Moreau envelopes and proximal mappings of proper and l.s.c (but not necessarily convex) functions, together with the fact that $\inf_{u \in \mathbb{R}^n} f(u) > -\infty$, guarantee that the proximal mapping is nonempty and compact; this is stated formally below in Lemma 4.2.

Remark 3.3 (On the v-update) Noting that Ψ_k is a strongly convex function, due to the fact that $\rho G^T G + \theta I \succ 0$ for any $\theta > 0$, we can derive from the first-order optimality conditions of (3.2) an *explicit expression* for the v-update step:

$$v^{k+1} = (\rho G^T G + \theta I)^{-1} \left(\theta v^k + G^T (y^k + \rho F(u^{k+1})) - \nabla h(v^k) \right). \tag{3.4}$$

Since the matrix $G^T G + \theta I$ remains unchanged throughout the execution of Algorithm 1, its inversion needs to be calculated only once.

In the next section we establish the subsequence and global convergence guarantees of Algorithm 1.

4 Convergence Analysis of DAM

4.1 Preliminaries

Our analysis will revolve around the function φ and its derivatives; these are recorded below for ease of use.

$$\nabla_u \varphi(u, v, y) = \nabla g(u) + \nabla F^T(u)(y + \rho(F(u) - Gv)), \tag{4.1}$$

$$\nabla_v \varphi(u, v, y) = -G^T(y + \rho(F(u) - Gv)), \tag{4.2}$$

$$\nabla_y \varphi(u, v, y) = F(u) - Gv. \tag{4.3}$$

Mainly, we utilize the *local Lipschitz* continuity of the gradients above; the proof is rather technical and so is deferred to Appendix C.

Lemma 4.1 *(Local Lipschitz continuity of $\nabla\varphi$) The function*

$$\varphi(u, v, y) = g(u) + \langle y, F(u) - Gv \rangle + \frac{\rho}{2} \|F(u) - Gv\|^2$$

satisfies that for every nonempty and compact set $C \subseteq \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^q$ there exists $L_C \geq 0$ such that

$$\|\nabla\varphi(x) - \nabla\varphi(z)\| \leq L_C \|x - z\|, \quad \forall x, z \in C. \tag{4.4}$$

That is, $\nabla\varphi$ is locally Lipschitz continuous.

Let us recall the well-definedness of the proximal mapping operator with respect to a nonconvex function (cf. [26, page 21]).

Lemma 4.2 *(Properties of the proximal mapping) Let* $\zeta : \mathbb{R}^d \to (-\infty, \infty]$ *be a proper lower semi-continuous function with* $\inf_{\mathbb{R}^n} \zeta > -\infty$, $x \in \mathbb{R}^d$, *and* $t > 0$. *The proximal map associated to* ζ/t *given by*

$$\operatorname{prox}_{\zeta/t}(x) = \operatorname{argmin}\left\{\zeta(z) + \frac{t}{2}\|z - x\|^2 : z \in \mathbb{R}^d\right\},$$

is a nonempty and compact set.

The infrastructure for the forthcoming analysis of Algorithm 1 comprises the following assumptions related to the proximal parameter.

Assumption A (Proximal parameter implies descent) The proximal parameters $\{\sigma_k\}_{k\geq 1}$ are chosen so that there exists $\nu > 0$ such that for every $k \geq 1$,

$$\begin{aligned}&\varphi(u^{k+1}, v^k, y^k) - \varphi(u^k, v^k, y^k) - \langle \nabla_u \varphi(u^k, v^k, y^k), u^{k+1} - u^k\rangle \\ &\leq \frac{\sigma_{k+1} - \nu}{2} \left\|u^{k+1} - u^k\right\|^2.\end{aligned} \tag{4.5}$$

Assumption B (Proximal parameter sequence is bounded) The proximal parameter's sequence $\{\sigma_k\}_{k\geq 1}$ is bounded, i.e.,

$$\bar{\sigma} := \sup_{k\geq 1} \sigma_k < \infty. \tag{4.6}$$

We will also utilize the celebrated descent lemma.

Lemma 4.3 *(Descent lemma [2]) Let* $\phi : \mathbb{R}^n \to \mathbb{R}$ *be a function with an* L_S*-Lipschitz continuous gradient* $\nabla\phi$ *over a given convex set S, with* $L_S \in \mathbb{R}_+$. *Then,*

$$|\phi(x) - \phi(z) - \langle \nabla\phi(z), x - z\rangle| \leq \frac{L_S}{2}\|x - z\|^2, \quad \forall x, z \in S. \tag{4.7}$$

Our analysis is based on the methodology and terminology introduced by [7], and refined in the more recent [8]; see [30] for a summary of these results. Accordingly, we begin by defining the notion of *gradient-like descent sequence* [8, Definition 6.1].

Definition 4.1 (Gradient-like descent sequence) A sequence $\{x_k\}_{k\geq 1} \subseteq \mathbb{R}^d$ is called a gradient-like descent sequence with respect to a proper and lsc function $\Gamma : \mathbb{R}^d \to (-\infty, \infty]$ if it satisfies the following three conditions:

(C1) **Sufficient decrease property.** There exists $c_1 > 0$ such that

$$\Gamma(x^{k+1}) - \Gamma(x^k) \leq -c_1 \left\|x^{k+1} - x^k\right\|^2, \quad \forall k \geq 1.$$

(C2) **A subgradient lower bound for the iterates gap.** There exists $c_2 > 0$ such that, for every $k \geq 1$, there exists $\xi^{k+1} \in \partial\Gamma(x^{k+1})$ which satisfies

$$\left\|\xi^{k+1}\right\| \leq c_2 \left\|x^{k+1} - x^k\right\|.$$

(C3) **Continuity condition.** Let $\{x^k\}_{k\in\mathcal{K}}$ be a subsequence converging to a point $\bar{x}$. Then,

$$\limsup_{k\in\mathcal{K}\subseteq\mathbb{N}} \Gamma(x^k) \le \Gamma(\bar{x}).$$

A bounded gradient-like descent sequence has the following convergence guarantees (cf. [7] and [8]).

Lemma 4.4 *(Subsequence convergence) Let $\{x_k\}_{k\ge 1}$ be a bounded gradient-like descent sequence with respect to the proper and lsc function Γ. Denote Ω as the set of cluster points of $\{x^k\}_{k\ge 1}$. Then, Ω is a nonempty and compact subset of* crit Γ*; $\lim_{k\to\infty} \operatorname{dist}(x^k, \Omega) = 0$; and Γ is finite and constant on Ω.*

In order to obtain a global convergence result, we assume that Γ satisfies the nonsmooth Kurdyka-Łojasiewicz (KL) property [6], which is in particular true when Γ is a semi-algebraic function. For further details, see [7] and references therein.

Theorem 4.1 *(Global convergence) [8, Theorem 6.2] Let $\{x_k\}_{k\ge 1}$ be a bounded gradient-like descent sequence with respect to the proper and lsc function Γ. If Γ satisfies the KL property, then the sequence $\{x^k\}_{k\ge 1}$ has a finite length, i.e., $\sum_{k=1}^{\infty} \left\|x^{k+1} - x^k\right\| < \infty$, and it converges to a critical point $x^* \in$* crit Γ.

In light of Lemma 4.4 and Theorem 4.1, our goal boils down to establishing that the sequence generated by Algorithm 1 is a gradient-like descent sequence.

4.2 The Descent-ascent and *v*-*y* Relations

Before we establish that the sequence generated by Algorithm 1 is indeed a gradient-like descent sequence, two results are required: (i) links between the primal variable v and the multiplier y (Lemma 4.5); (ii) a descent-ascent relation (Lemma 4.6).

Lemma 4.5 *(v and y relation) It holds that*

$$G^T y^{k+1} = \nabla h(v^k) + \theta(v^{k+1} - v^k), \qquad \textit{for all } k \ge 0 \tag{4.8}$$

$$\left\|y^{k+1} - y^k\right\|^2 \le \frac{2\theta^2}{\lambda_{\min}(GG^T)} \left\|v^{k+1} - v^k\right\|^2 + \frac{2(\theta + L_h)^2}{\lambda_{\min}(GG^T)} \left\|v^k - v^{k-1}\right\|^2, \quad \textit{for all } k \ge 1. \tag{4.9}$$

Proof We begin by proving Equation 4.8. Recall that (cf. (4.2))

$$\nabla_v \varphi(u^{k+1}, v^{k+1}, y^k) = -G^T(y^k + \rho(F(u^{k+1}) - Gv^{k+1})) = -G^T y^{k+1},$$

where the last equality is due to multiplier update step (3.3). Applying the first-order optimality condition to the minimization problem of v-update step (3.2), we obtain that

$$\begin{aligned}0 &= \nabla_v \varphi(u^{k+1}, v^{k+1}, y^k) + \nabla h(v^k) + \theta(v^{k+1} - v^k)\\ &= -G^T y^{k+1} + \nabla h(v^k) + \theta(v^{k+1} - v^k),\end{aligned}$$

meaning that Equation 4.8 holds true for any $k \geq 0$.

We continue with proving relation (4.9). The matrix GG^T is positive definite and hence $\lambda{:=}\lambda_{\min}(GG^T) > 0$. Additionally, for every $y \in \mathbb{R}^q$, we have

$$\lambda^{-1} \left\| G^T y \right\|^2 = \lambda^{-1} \langle G^T y, G^T y\rangle = \lambda^{-1}\langle y, GG^T y\rangle \geq \lambda^{-1}\langle y, \lambda y\rangle = \|y\|^2 .$$

Thus, for every $k \geq 1$,

$$\left\| y^{k+1} - y^k \right\| \leq \lambda^{-1/2} \left\| G^T(y^{k+1} - y^k)\right\| . \tag{4.10}$$

Combining (4.10) with (4.8), we obtain for every $k \geq 1$ that

$$\begin{aligned}\left\| y^{k+1} - y^k \right\|^2 &\leq \frac{1}{\lambda} \left\| \nabla h(v^k) - \nabla h(v^{k-1}) + \theta(v^{k+1} - v^k) - \theta(v^k - v^{k-1}) \right\|^2\\ &\leq \frac{1}{\lambda}\left(\left\|\nabla h(v^k) - \nabla h(v^{k-1})\right\| + \theta \left\| v^{k+1} - v^k \right\| + \theta \left\| v^k - v^{k-1}\right\|\right)^2\\ &\leq \frac{1}{\lambda}\left(\theta \left\| v^{k+1} - v^k\right\| + (\theta + L_h)\left\| v^k - v^{k-1}\right\|\right)^2,\end{aligned}$$

where for the second inequality we apply the triangle inequality, and the subsequent inequality is due to the L_h-Lipschitz continuity of ∇h . We finally complete the proof by utilizing the relation $(s+t)^2 \leq 2s^2 + 2t^2$. □

Lemma 4.6 *(Descent-ascent relation) Suppose that Assumption A holds. Then, for every $k \geq 1$, it holds that*

$$\begin{aligned}\Delta_{\mathcal{L}_\rho,k} &{:=} \mathcal{L}_\rho(u^{k+1}, v^{k+1}, y^{k+1}) - \mathcal{L}_\rho(u^k, v^k, y^k)\\ &\leq -\frac{\nu}{2}\left\| u^{k+1} - u^k\right\|^2 - \frac{(2\theta - L_h)}{2}\left\| v^{k+1} - v^k\right\|^2 + \frac{1}{\rho}\left\| y^{k+1} - y^k\right\|^2 .\end{aligned} \tag{4.11}$$

Proof We will prove the required by establishing bounds following from the three update steps in Algorithm 1, starting with y-update step (3.3). From (3.3), we trivially have that

$$\begin{aligned}\delta_{y,k} &{:=} \mathcal{L}_\rho(u^{k+1}, v^{k+1}, y^{k+1}) - \mathcal{L}_\rho(u^{k+1}, v^{k+1}, y^k)\\ &= \langle y^{k+1} - y^k, F(u^{k+1}) - Gv^{k+1}\rangle = \frac{1}{\rho}\left\| y^{k+1} - y^k\right\|^2 .\end{aligned} \tag{4.12}$$

By the u-update step (3.1),

$$u^{k+1} \in \operatorname{argmin}_{u\in\mathbb{R}^n} f(u) + \langle \nabla_u \varphi(u^k, v^k, y^k), u\rangle + \frac{\sigma_{k+1}}{2}\left\| u - u^k\right\|^2,$$

we have that

$$f(u^{k+1}) + \langle \nabla_u \varphi(u^k, v^k, y^k), u^{k+1} - u^k \rangle \\ + \frac{\sigma_{k+1}}{2} \left\| u^{k+1} - u^k \right\|^2 \leq f(u^k). \tag{4.13}$$

Assumption A then ensures that decrease property (4.5) holds true, which combined with (4.13) results in

$$f(u^{k+1}) + \varphi(u^{k+1}, v^k, y^k) - f(u^k) - \varphi(u^k, v^k, y^k) \leq -\frac{\nu}{2} \left\| u^{k+1} - u^k \right\|^2 .$$

Hence, setting $\delta_{u,k} := \mathcal{L}_\rho(u^{k+1}, v^k, y^k) - \mathcal{L}_\rho(u^k, v^k, y^k)$, yields

$$\delta_{u,k} \leq -\frac{\nu}{2} \left\| u^{k+1} - u^k \right\|^2 . \tag{4.14}$$

Now, let us consider the implications of the v-update (3.2). For any $k \geq 1$, we have from the y update step (3.3) and the definition of $\lambda_{\min}(\cdot)$ that

$$\begin{aligned} \delta_{v,k} &:= \mathcal{L}_\rho(u^{k+1}, v^{k+1}, y^k) - \mathcal{L}_\rho(u^{k+1}, v^k, y^k) \\ &= h(v^{k+1}) - h(v^k) - \langle G^T y^k, v^{k+1} - v^k \rangle \\ &\quad + \frac{\rho}{2} \left(\left\| G v^{k+1} - F(u^{k+1}) \right\|^2 - \left\| G v^k - F(u^{k+1}) \right\|^2 \right) \\ &= h(v^{k+1}) - h(v^k) - \left\langle y^k + \rho(F(u^{k+1}) - G v^{k+1}) + \frac{\rho}{2} G(v^{k+1} - v^k), G(v^{k+1} - v^k) \right\rangle \\ &\leq h(v^{k+1}) - h(v^k) - \langle G^T y^{k+1}, v^{k+1} - v^k \rangle - \frac{\rho \lambda_{\min}(G^T G)}{2} \left\| v^{k+1} - v^k \right\|^2 . \end{aligned}$$

Applying identity (4.8) from Lemma 4.5, we obtain that

$$\delta_{v,k} \leq h(v^{k+1}) - h(v^k) - \langle \nabla h(v^k), v^{k+1} - v^k \rangle \\ - \frac{2\theta + \rho \lambda_{\min}(G^T G)}{2} \left\| v^{k+1} - v^k \right\|^2 . \tag{4.15}$$

Since h has an L_h-Lipschitz continuous gradient, it holds that (cf. Lemma 4.3)

$$h(v^{k+1}) - h(v^k) - \langle \nabla h(v^k), v^{k+1} - v^k \rangle \leq \frac{L_h}{2} \left\| v^{k+1} - v^k \right\|^2 . \tag{4.16}$$

Thus, by summing (4.15) and (4.16), and noting that $\lambda_{\min}(G^T G) \geq 0$, we obtain that

$$\begin{aligned} \delta_{v,k} &\leq -\frac{\rho \lambda_{\min}(G^T G) + 2\theta - L_h}{2} \left\| v^{k+1} - v^k \right\|^2 \\ &\leq -\frac{2\theta - L_h}{2} \left\| v^{k+1} - v^k \right\|^2 . \end{aligned} \tag{4.17}$$

Finally, by summing our three bounds (4.12), (4.14), and (4.17), and using the relation

$$\Delta_{\mathcal{L}_\rho,k} = \delta_{u,k} + \delta_{v,k} + \delta_{y,k},$$

we obtain stated inequality (4.11). □

Remark 4.1 (On the descent-ascent of the scheme) A prominent difficulty in the Lagrangian-based approach is the ascent property that is induced by the multiplier update, as expressed by Lemma 4.6. In the nonconvex setup, this nullifies a fundamental tool of descent methods for nonconvex optimization—the sufficient decrease property. We overcome this challenge in the forthcoming section.

By combining Lemma 4.5 and Lemma 4.6, we can derive a bound on $\Delta_{\mathcal{L}_\rho,k}$ given in terms of the primal variables u and v.

Corollary 4.1 *Suppose that Assumption A holds. Then, for every $k \geq 1$,*

$$\begin{aligned}\Delta_{\mathcal{L}_\rho,k} \leq -\frac{\nu}{2}\left\|u^{k+1}-u^k\right\|^2 - \left(\frac{2\theta - L_h}{2} - \frac{2\theta^2}{\rho\lambda_{\min}(GG^T)}\right)\left\|v^{k+1}-v^k\right\|^2 \\ +\frac{2(\theta+L_h)^2}{\rho\lambda_{\min}(GG^T)}\left\|v^k - v^{k-1}\right\|^2. \qquad (4.18)\end{aligned}$$

The bound in Corollary 4.1 is utilized in the next section to establish a sufficient decrease property for a Lyapunov function.

4.3 The Lyapunov Function and Convergence Guarantees

To reach our goal of convergence to a critical point, we adopt the approach formulated by [9] and introduce the Lyapunov function $\mathcal{E}_\beta : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^q \times \mathbb{R}^m \to (-\infty, \infty]$, with $\beta > 0$, defined by

$$\mathcal{E}_\beta(u, v, y, w) := \mathcal{L}_\rho(u, v, y) + \frac{\beta}{2}\|v - w\|^2. \qquad (4.19)$$

The following relationship between the critical points of $\mathcal{E}_\beta$, $\mathcal{L}_\rho$, and $\mathcal{L}$ justifies referring to $\mathcal{E}_\beta$ as a Lyapunov function for (M).

Proposition 4.1 *(Critical points relationships) For any $u \in \mathbb{R}^n$, $v, w \in \mathbb{R}^m$, and $y \in \mathbb{R}^q$, it holds that*

$$\begin{aligned}(u, v, y, w) \in \operatorname{crit}\mathcal{E}_\beta &\Longrightarrow (u, v, y) \in \operatorname{crit}\mathcal{L}_\rho \text{ and } \mathcal{E}_\beta(u, v, y, w) = \mathcal{L}_\rho(u, v, y), \\ (u, v, y) \in \operatorname{crit}\mathcal{L}_\rho &\Longrightarrow (u, v, y) \in \operatorname{crit}\mathcal{L} \text{ and } \mathcal{L}_\rho(u, v, y) = \mathcal{L}(u, v, y).\end{aligned}$$

Proof Let $(u, v, y, w) \in \operatorname{crit}\mathcal{E}_\beta$, then by the definitions of $\mathcal{E}_\beta$ and $\mathcal{L}_\rho$, we have

$$0 \in \partial_u \mathcal{E}_\beta(u, v, y, w) = \partial_u \mathcal{L}_\rho(u, v, y), \qquad (4.20)$$

$$0 = \nabla_v \mathcal{E}_\beta(u, v, y, w) = \nabla_v \mathcal{L}_\rho(u, v, y) + \beta(v - w), \tag{4.21}$$

$$0 = \nabla_y \mathcal{E}_\beta(u, v, y, w) = \nabla_y \mathcal{L}_\rho(u, v, y), \tag{4.22}$$

$$0 = \nabla_w \mathcal{E}_\beta(u, v, y, w) = \beta(w - v). \tag{4.23}$$

Thus, we obtain that $w = v$ by (4.23), and together with (4.21), it follows that $\nabla_v \mathcal{L}_\rho(u, v, y) = 0$. Combined with (4.20) and (4.22), we obtain that $(u, v, y) \in \operatorname{crit} \mathcal{L}_\rho(u, v, y)$. Having $w = v$ and the definition of $\mathcal{E}_\beta$ also imply that $\mathcal{E}_\beta(u, v, y, w) = \mathcal{L}_\rho(u, v, y)$.

Next, assume that $(u, v, y) \in \operatorname{crit} \mathcal{L}_\rho(u, v, y)$. Then, by the definitions of $\mathcal{L}_\rho$ and $\mathcal{L}$, we have

$$0 \in \partial_u \mathcal{L}_\rho(u, v, y) = \partial_u \mathcal{L}(u, v, y) + \rho \nabla F(u)^T (F(u) - Gv) \tag{4.24}$$

$$0 = \nabla_v \mathcal{L}_\rho(u, v, y) = \nabla_v \mathcal{L}(u, v, y) - \rho G^T (F(u) - Gv), \tag{4.25}$$

$$0 = \nabla_y \mathcal{L}_\rho(u, v, y) = \nabla_y \mathcal{L}(u, v, y) = F(u) - Gv. \tag{4.26}$$

So, by (4.26), we obtain that $\nabla_y \mathcal{L}(u, v, y) = 0$ and that $F(u) - Gv = 0$. Together, with (4.24) and (4.25), we also obtain that $0 \in \partial_u \mathcal{L}(u, v, y)$ and that $\nabla_v \mathcal{L}(u, v, y) = 0$, and it follows that $(u, v, y) \in \operatorname{crit} \mathcal{L}$. Having $F(u) = Gv$ and the definition of $\mathcal{L}_\rho$ also imply that $\mathcal{L}_\rho(u, v, y) = \mathcal{L}(u, v, y)$. □

At this point, we are ready to prove that the sequence $\{z^k := (u^k, v^k, y^k, v^{k-1})\}_{k \geq 1}$ is a gradient-like descent sequence with respect to the Lyapunov function $\mathcal{E}_\beta$, for appropriate choices of ρ, θ, and β. In the following three lemmas we prove that the conditions of Definition 4.1 are satisfied. We begin with proving the sufficient descent property of the sequence $\{z^k\}_{k \geq 1}$ with respect to $\mathcal{E}_\beta(\cdot)$.

Lemma 4.7 *(Sufficient decrease for $\mathcal{E}_\beta$) Suppose that Assumption A holds and that*

$$\rho = \frac{4\eta L_h}{\lambda_{\min}(GG^T)} \quad \text{and} \quad \beta = \theta = \frac{(\eta - 1)L_h}{2}, \tag{4.27}$$

for some $\eta > 2 + \sqrt{5}$. Then, there exists $c_1 > 0$ such that, for every $k \geq 1$, we have

$$\mathcal{E}_\beta(z^{k+1}) - \mathcal{E}_\beta(z^k) \leq -c_1 \left\| z^{k+1} - z^k \right\|^2. \tag{4.28}$$

Proof Let $k \geq 1$ and ν be as defined in Assumption A. Recalling the bound of $\Delta_{\mathcal{L}_\rho, k}$ (cf. (4.18)), we observe that

$$\begin{aligned} \Delta_{\mathcal{E}_\beta, k} &:= \mathcal{E}_\beta(u^{k+1}, v^{k+1}, y^{k+1}, v^k) - \mathcal{E}_\beta(u^k, v^k, y^k, v^{k-1}) \\ &= \Delta_{\mathcal{L}_\rho, k} + \frac{\beta}{2} \left(\left\| v^{k+1} - v^k \right\|^2 - \left\| v^k - v^{k-1} \right\|^2 \right) \\ &\leq -\frac{\nu}{2} \left\| u^{k+1} - u^k \right\|^2 - \alpha_1 \left\| v^{k+1} - v^k \right\|^2 - \alpha_2 \left\| v^k - v^{k-1} \right\|^2, \end{aligned}$$

with

$$\lambda = \lambda_{\min}(GG^T),\ \alpha_1 = \frac{2\theta - L_h - \beta}{2} - \frac{2\theta^2}{\rho\lambda} \quad \text{and} \quad \alpha_2 = \frac{\beta}{2} - \frac{2(\theta + L_h)^2}{\rho\lambda}.$$

Substituting ρ, θ, and β according to (4.27) followed by simple algebra, we obtain that

$$\alpha_1 = \alpha_2 = \frac{L_h}{8\eta}(\eta^2 - 4\eta - 1).$$

As the solutions for the quadratic equation $\eta^2 - 4\eta - 1 = 0$ are $\eta_{1,2} = 2 \pm \sqrt{5}$, it follows that $\alpha_1 > 0$ for every $\eta > 2 + \sqrt{5}$.

Following Lemma 4.5 (cf. (4.9)), we have

$$\left\| y^{k+1} - y^k \right\|^2 \le \frac{2(\theta + L_h)^2}{\lambda} \left(\left\| v^{k+1} - v^k \right\|^2 + \left\| v^k - v^{k-1} \right\|^2 \right).$$

Adding $\left\| v^{k+1} - v^k \right\|^2 + \left\| v^k - v^{k-1} \right\|^2$ to both sides and setting

$$\alpha_3 = 1 + 2(\theta + L_h)^2/\lambda = 1 + \frac{1}{2}\left((\eta + 1)L_h\right)^2/\lambda,$$

we obtain

$$\begin{aligned} &\left\| v^{k+1} - v^k \right\|^2 + \left\| v^k - v^{k-1} \right\|^2 + \left\| y^{k+1} - y^k \right\|^2 \\ &\le \alpha_3 \left(\left\| v^{k+1} - v^k \right\|^2 + \left\| v^k - v^{k-1} \right\|^2 \right). \end{aligned}$$

Therefore, with $c_1 = \min\{\nu/2, \alpha_1/\alpha_3\}$, we have

$$\Delta_{\mathcal{E}_\beta,k} \le -c_1 \left(\left\| u^{k+1} - u^k \right\|^2 + \left\| v^{k+1} - v^k \right\|^2 + \left\| y^{k+1} - y^k \right\|^2 + \left\| v^k - v^{k-1} \right\|^2 \right).$$

□

We continue with proving the subgradient lower bound and the continuity condition (cf. conditions (C2) and (C3) of Definition 4.1). This requires the boundedness assumption on the proximal parameter sequence stated by Assumption B .

Remark 4.2 The proximal parameter backtracking procedure we devise in Sect. 5 satisfies Assumption A, and it also satisfies Assumption B under a standard boundness assumption with respect to the sequence generated by DAM.

Lemma 4.8 *(Subgradient bound for $\mathcal{E}_\beta$) Suppose that Assumption B holds and assume that the sequence $\{\omega^k:=(u^k, v^k, y^k)\}_{k\geq 1}$ is bounded. Then, there exists $c_2 > 0$, such that, for every $k \geq 1$, there exists $\xi^{k+1} \in \partial\mathcal{E}_\beta(z^{k+1})$ satisfying*

$$\left\|\xi^{k+1}\right\| \leq c_2 \left\|z^{k+1} - z^k\right\|. \tag{4.29}$$

Proof Let $k \geq 1$. We note that for every $\xi^{k+1} = (\xi_u^{k+1}, \xi_v^{k+1}, \xi_y^{k+1}, \xi_w^{k+1}) \in \partial\mathcal{E}_\beta(z^{k+1})$, we have

$$\xi_u^{k+1} \in \partial_u\mathcal{E}(u^{k+1}, v^{k+1}, y^{k+1}, v^k) = \partial_u f(u^{k+1}) + \nabla_u\varphi(u^{k+1}, v^{k+1}, y^{k+1}), \tag{4.30}$$

$$\begin{aligned}\xi_v^{k+1} &= \nabla_v\mathcal{E}(u^{k+1}, v^{k+1}, y^{k+1}, v^k)\\ &= \nabla h(v^{k+1}) + \nabla_v\varphi(u^{k+1}, v^{k+1}, y^{k+1}) + \beta(v^{k+1} - v^k),\end{aligned} \tag{4.31}$$

$$\xi_y^{k+1} = \nabla_y\mathcal{E}(u^{k+1}, v^{k+1}, y^{k+1}, v^k) = F(u^{k+1}) - Gv^{k+1}, \tag{4.32}$$

$$\xi_w^{k+1} = \nabla_w\mathcal{E}(u^{k+1}, v^{k+1}, y^{k+1}, v^k) = -\beta(v^{k+1} - v^k). \tag{4.33}$$

Furthermore, we recall that u-step (3.1) is stated as the following minimization problem:

$$u^{k+1} \in \operatorname{argmin}_{u\in\mathbb{R}^n}\left\{f(u) + \langle\nabla_u\varphi(u^k, v^k, y^k), u\rangle + \frac{\sigma_{k+1}}{2}\left\|u - u^k\right\|^2\right\}.$$

Thus, we apply the first-order optimality condition and obtain that there exists $\zeta^{k+1} \in \partial f(u^{k+1})$ such that

$$\zeta^{k+1} = -\nabla_u\varphi(u^k, v^k, y^k) - \sigma_{k+1}(u^{k+1} - u^k).$$

It follows that there exists ξ_u^{k+1} as in (4.30), such that

$$\begin{aligned}\left\|\xi_u^{k+1}\right\| &= \left\|\nabla_u\varphi(u^{k+1}, v^{k+1}, y^{k+1}) - \nabla_u\varphi(u^k, v^k, y^k) - \sigma_{k+1}(u^{k+1} - u^k)\right\|\\ &\leq \left\|\nabla_u\varphi(u^{k+1}, v^{k+1}, y^{k+1}) - \nabla_u\varphi(u^k, v^k, y^k)\right\| + \sigma_{k+1}\left\|u^{k+1} - u^k\right\|\\ &\leq \left\|\nabla_u\varphi(u^{k+1}, v^{k+1}, y^{k+1}) - \nabla_u\varphi(u^k, v^k, y^k)\right\| + \sigma_{k+1}\left\|z^{k+1} - z^k\right\|.\end{aligned}$$

Recalling that $\nabla_u\varphi$ is locally Lipschitz continuous (cf. Lemma 4.1) and noting that the sequence $\{\omega^k:=(u^k, v^k, y^k)\}_{k\geq 1}$ is bounded, it follows (cf. Proposition B.1) that there exists $L_\varphi \in \mathbb{R}_+$ such that, for every $k \geq 1$, we have

$$\left\|\nabla_u\varphi(\omega^{k+1}) - \nabla_u\varphi(\omega^k)\right\| \leq L_\varphi\left\|\omega^{k+1} - \omega^k\right\| \leq L_\varphi\left\|z^{k+1} - z^k\right\|.$$

Together with Assumption B, it follows that

$$\left\|\xi_u^{k+1}\right\| \leq (L_\varphi + \bar{\sigma})\left\|z^{k+1} - z^k\right\|,$$

with $\bar{\sigma} := \sup_{k \geq 1} \sigma_k < \infty$.

Next, we note that

$$\begin{aligned} &\nabla_v \varphi(u^{k+1}, v^{k+1}, y^{k+1}) \\ &= -G^T \left(y^{k+1} + \rho(F(u^{k+1}) - Gv^{k+1}) \right) = -G^T(2y^{k+1} - y^k), \end{aligned} \tag{4.34}$$

where the second equality is due to multiplier update step (3.3). Together with identity (4.8) proven in Lemma 4.5, we obtain that,

$$\nabla_v \varphi(u^{k+1}, v^{k+1}, y^{k+1}) = -G^T(y^{k+1} - y^k) - \nabla h(v^k) - \theta(v^{k+1} - v^k). \tag{4.35}$$

So together with (4.31), we have

$$\begin{aligned} \left\| \xi_v^{k+1} \right\| &= \left\| \nabla h(v^{k+1}) + \nabla_v \varphi(u^{k+1}, v^{k+1}, y^{k+1}) + \beta(v^{k+1} - v^k) \right\| \\ &= \left\| \nabla h(v^{k+1}) - \nabla h(v^k) - G^T(y^{k+1} - y^k) + (\beta - \theta)(v^{k+1} - v^k) \right\| \\ &\leq \left\| h(v^{k+1}) - \nabla h(v^k) \right\| + \|G\| \left\| y^{k+1} - y^k \right\| + |\beta - \theta| \left\| v^{k+1} - v^k \right\| \\ &\leq (L_h + |\beta - \theta|) \left\| v^{k+1} - v^k \right\| + \|G\| \left\| y^{k+1} - y^k \right\| \\ &\leq (L_h + |\beta - \theta| + \|G\|) \left\| z^{k+1} - z^k \right\|, \end{aligned} \tag{4.36}$$

where the second inequality is due to the L_h-Lipschitz continuity of ∇h.

Finally, we note that

$$\left\| \xi_y^{k+1} \right\| = \left\| F(u^{k+1}) - Gv^{k+1} \right\| = \rho^{-1} \left\| y^{k+1} - y^k \right\| \leq \rho^{-1} \left\| z^{k+1} - z^k \right\| \tag{4.37}$$

and

$$\left\| \xi_w^{k+1} \right\| = \beta \left\| v^{k+1} - v^k \right\| \leq \beta \left\| z^{k+1} - z^k \right\|, \tag{4.38}$$

where the second equality in (4.37) is due to multiplier update step (3.3). Thus, by summing the bounds for $\|\xi_u^{k+1}\|$, $\|\xi_v^{k+1}\|$, $\left\|\xi_y^{k+1}\right\|$, and $\|\xi_w^{k+1}\|$, we obtain that

$$\left\| \xi^{k+1} \right\| \leq \left\| \xi_u^{k+1} \right\| + \left\| \xi_v^{k+1} \right\| + \left\| \xi_y^{k+1} \right\| + \left\| \xi_w^{k+1} \right\| \leq c_2 \left\| z^{k+1} - z^k \right\|,$$

with $c_2 = L_\varphi + \bar{\sigma} + L_h + |\beta - \theta| + \|G\| + \rho^{-1} + \beta > 0$. □

Lemma 4.9 *(Continuity condition for $\mathcal{E}_\beta$) Suppose that Assumptions A and B hold and that ρ, θ, and β are as in Lemma 4.7. Let $\{z^{k+1}\}_{k \in \mathcal{K}}$ be a subsequence of $\{z^k\}_{k \geq 1}$ which converges to a point $\bar{z} = (\bar{u}, \bar{v}, \bar{y}, \bar{w})$. Then,*

$$\limsup_{\substack{k \to \infty \\ k \in \mathcal{K}}} \mathcal{E}_\beta(z^{k+1}) \leq \mathcal{E}_\beta(\bar{z}). \tag{4.39}$$

Proof The Lyapunov function $\mathcal{E}_\beta$ can be written as $\mathcal{E}_\beta(u, v, y, w) = f(u) + \Psi(u, v, y, w)$, where the function Ψ is defined by

$$\Psi(u, v, y, w) := g(u) + h(v) + \langle y, F(u) - Gv\rangle + \frac{\rho}{2}\|F(u) - Gv\|^2 + \frac{\beta}{2}\|v - w\|^2.$$

Obviously, Ψ is a continuous function and, as f is proper and lsc, it follows that $\mathcal{E}_\beta$ is also proper and lsc. Furthermore, we have $\lim_{\substack{k\to\infty\\ k\in\mathcal{K}}} \Psi(u^{k+1}, v^{k+1}, y^{k+1}, v^k) = \Psi(\bar{u}, \bar{v}, \bar{y}, \bar{w})$, and hence inequality (4.39) can be stated as

$$\begin{aligned}\limsup_{\substack{k\to\infty\\ k\in\mathcal{K}}} \mathcal{E}_\beta(u^{k+1}, v^{k+1}, y^{k+1}, v^k) &= \limsup_{\substack{k\to\infty\\ k\in\mathcal{K}}} f(u^{k+1}) + \lim_{\substack{k\to\infty\\ k\in\mathcal{K}}} \Psi(u^{k+1}, v^{k+1}, y^{k+1}, v^k)\\ &= \limsup_{\substack{k\to\infty\\ k\in\mathcal{K}}} f(u^{k+1}) + \Psi(\bar{u}, \bar{v}, \bar{y}, \bar{w})\\ &\le f(\bar{u}) + \Psi(\bar{u}, \bar{v}, \bar{y}, \bar{w}) = \mathcal{E}(\bar{u}, \bar{v}, \bar{y}, \bar{w}).\end{aligned}$$

Therefore, it remains to prove the following inequality

$$\limsup_{\substack{k\to\infty\\ k\in\mathcal{K}}} f(u^{k+1}) \le f(\bar{u}). \tag{4.40}$$

Let $k \in \mathcal{K}$. We recall that u-step (3.1) is given by the following minimization problem:

$$u^{k+1} \in \operatorname{argmin}_{u\in\mathbb{R}^n} \left\{ f(u) + \langle \nabla_u \varphi(u^k, v^k, y^k), u\rangle + \frac{\sigma_{k+1}}{2}\left\|u - u^k\right\|^2 \right\},$$

which immediately implies that

$$\begin{aligned}&f(u^{k+1}) + \langle \nabla_u \varphi(u^k, v^k, y^k), u^{k+1}\rangle + \frac{\sigma_{k+1}}{2}\left\|u^{k+1} - u^k\right\|^2\\ &\le f(\bar{u}) + \langle \nabla_u \varphi(u^k, v^k, y^k), \bar{u}\rangle + \frac{\sigma_{k+1}}{2}\left\|\bar{u} - u^k\right\|^2.\end{aligned}$$

After some algebra we can state the above inequality as

$$\begin{aligned}f(u^{k+1}) &\le f(\bar{u}) + \left\langle \nabla_u \varphi(u^k, v^k, y^k) + \frac{\sigma_{k+1}}{2}(\bar{u} - u^k + u^{k+1} - u^k), \bar{u} - u^{k+1}\right\rangle\\ &\le f(\bar{u}) + \left\|\nabla_u \varphi(u^k, v^k, y^k) + \frac{\sigma_{k+1}}{2}(\bar{u} - u^k + u^{k+1} - u^k)\right\| \left\|\bar{u} - u^{k+1}\right\| \\ &\le f(\bar{u}) + \alpha_k \left\|\bar{u} - u^{k+1}\right\|,\end{aligned} \tag{4.41}$$

with

$$\alpha_k := \left\|\nabla_u \varphi(u^k, v^k, y^k)\right\| + \frac{\sigma_{k+1}}{2}\left(\left\|\bar{u} - u^k\right\| + \left\|u^{k+1} - u^k\right\|\right) \ge 0. \tag{4.42}$$

The second and third inequalities in (4.41) are due to the Cauchy–Schwarz and triangle inequalities, respectively.

Taking lim sup over $\mathcal{K}$, we obtain that

$$\limsup_{\substack{k\to\infty\\ k\in\mathcal{K}}} f(u^{k+1}) \leq f(\bar{u}) + \limsup_{\substack{k\to\infty\\ k\in\mathcal{K}}} \left(\alpha_k \left\|\bar{u} - u^{k+1}\right\|\right).$$

Thus, proving inequality (4.40) reduces to proving that $\lim_{\substack{k\to\infty\\ k\in\mathcal{K}}} \left(\alpha_k \left\|\bar{u} - u^{k+1}\right\|\right) = 0$. We assume that $\lim_{\substack{k\to\infty\\ k\in\mathcal{K}}} u^{k+1} = \bar{u}$, and so it is sufficient to prove that

$$\limsup_{\substack{k\to\infty\\ k\in\mathcal{K}}} \alpha_k < \infty,$$

which requires that we analyze the subsequence $\{(u^k, v^k, y^k\}_{k\in\mathcal{K}}$. To this end, we recall the sufficient descent result of Lemma 4.7 (cf. (4.28)), i.e., that there exists $c_1 > 0$ such that

$$\mathcal{E}_\beta(z^{k+1}) - \mathcal{E}_\beta(z^k) \leq -c_1 \left\|z^{k+1} - z^k\right\|^2, \quad \forall k \geq 1.$$

Thus, summing $\left\|z^{k+1} - z^k\right\|^2$ over $\mathcal{K}$, we obtain that

$$\begin{aligned}\sum_{k\in\mathcal{K}} \left\|z^{k+1} - z^k\right\|^2 &\leq \lim_{\substack{N\to\infty\\ N\in\mathcal{K}}} \sum_{k=1}^{N} \left\|z^{k+1} - z^k\right\|^2 \leq \limsup_{\substack{N\to\infty\\ N\in\mathcal{K}}} c_1^{-1} \sum_{k=1}^{N} \left(\mathcal{E}_\beta(z^k) - \mathcal{E}_\beta(z^{k+1}))\right) \\ &= c_1^{-1} \left(\mathcal{E}_\beta(z^1) - \liminf_{\substack{N\to\infty\\ N\in\mathcal{K}}} \mathcal{E}_\beta(z^{N+1})\right) \leq c_1^{-1} \left(\mathcal{E}_\beta(z^1) - \mathcal{E}_\beta(\bar{z})\right) < \infty,\end{aligned} \tag{4.43}$$

where the first inequality is due to the sufficient descent and the second inequality is justified by the fact that $\lim_{\substack{k\to\infty\\ k\in\mathcal{K}}} \left\|z^{k+1}\right\| = \bar{z}$ and that $\mathcal{E}_\beta$ is proper and lsc, which implies that

$$\liminf_{\substack{k\to\infty\\ k\in\mathcal{K}}} \mathcal{E}_\beta(z^{k+1}) \geq \mathcal{E}_\beta(\bar{z}) > -\infty.$$

As the nonnegative series $\sum_{k\in\mathcal{K}} \left\|z^{k+1} - z^k\right\|^2$ is finite, it follows that $\lim_{\substack{k\to\infty\\ k\in\mathcal{K}}} \left\|z^{k+1} - z^k\right\| = 0$, which implies that the subsequence $\{z^k\}_{k\in\mathcal{K}}$ also converges to $\bar{z}$. This result has several consequences. First, as φ is C^1, it follows that $\nabla_u\varphi$ is continuous over $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^q$ and therefore

$$\lim_{\substack{k\to\infty\\ k\in\mathcal{K}}} \left\|\nabla_u\varphi(u^k, v^k, y^k)\right\| = \|\nabla_u\varphi(\bar{u}, \bar{v}, \bar{y})\| < \infty. \tag{4.44}$$

Second, noting that $\lim_{\substack{k\to\infty\\k\in\mathcal{K}}} u^k = \lim_{\substack{k\to\infty\\k\in\mathcal{K}}} u^{k+1} = \bar{u}$, we have

$$\lim_{\substack{k\to\infty\\k\in\mathcal{K}}} \left\| u^{k+1} - u^k \right\| = 0 \quad\text{and}\quad \lim_{\substack{k\to\infty\\k\in\mathcal{K}}} \left\| \bar{u} - u^k \right\| = 0.$$

Together with Assumption B which states that $\sup_{k\geq 1} \sigma_k < \infty$, it follows that

$$\lim_{\substack{k\to\infty\\k\in\mathcal{K}}} \frac{\sigma_{k+1}}{2} \left(\left\| \bar{u} - u^k \right\| + \left\| u^{k+1} - u^k \right\| \right) = 0. \tag{4.45}$$

Summing (4.44) and (4.45) and recalling the definition of α_k (cf. (4.42)) we obtain that

$$\limsup_{\substack{k\to\infty\\k\in\mathcal{K}}} \alpha_k = \lim_{\substack{k\to\infty\\k\in\mathcal{K}}} \alpha_k < \infty.$$

□

Equipped with Lemma 4.7, Lemma 4.8, and Lemma 4.9, we can now apply Lemma 4.4 and Theorem 4.1 followed by Proposition 4.1, to establish that Algorithm 1 achieves criticality, given that the parameters ρ and θ are chosen accordingly to (4.27).

Theorem 4.2 *(Subsequence and global convergence) Suppose that Assumption A and Assumption B hold true, and that the sequence $\{\omega^k{:=}(u^k, v^k, y^k)\}_{k\geq 1}$ is bounded. Let ρ and θ be chosen such that*

$$\rho = (4\eta L_h)/\lambda_{\min}(GG^T) \textit{ and } \theta = (\eta - 1)L_h/2 \textit{ for some } \eta > 2 + \sqrt{5}.$$

Then the following implications hold:

- *(i) Let Ω be the set of cluster points of the sequence $\{\omega^k\}_{k\geq 1}$. Then, Ω is a nonempty and compact set; $\Omega \subset \operatorname{crit} \mathcal{L}$; $\lim_{k\to\infty} \operatorname{dist}(\omega^k, \Omega) = 0$; and $\mathcal{L}$ is finite and constant on Ω.*
- *(ii) Assume, in addition, that the functions f, g, and h, and the mapping F are semi-algebraic. Then, $\{\omega^k\}_{k\geq 1}$ has a finite length, i.e., $\sum_{k=1}^{\infty} \left\| \omega^{k+1} - \omega^k \right\| < \infty$, and it converges to a critical point $\omega^* \in \operatorname{crit} \mathcal{L}$.*

Remark 4.3 Recall that as proved in [6], a semi-algebraic function satisfies the KL property, and that the semi-algebraic property is preserved for most useful operations e.g., finite sum and product of semi-algebraic functions, and their composition (see e.g., [7] for examples and references). The class of problems modeled by semi-algebraic functions is ubiquitous in applications, and hence the global convergence result in Theorem 4.2 is applicable to a wide variety of models.

Remark 4.4 Asymptotic convergence rates of the sequence $\{\omega^k\}_{k\geq 1}$ could also be established; see e.g., [8, Theorem 6.3].

Algorithm 2 Proximal parameter backtracking procedure

```
Parameters: $\nu > 0$, $\mu > 1$, and $\sigma_0 > 0$.
Input: $u^k$, $v^k$, $y^k$, and $\sigma_k$.
Compute:
  set $\sigma_{k,0} \in [\sigma_0, \sigma_k]$;    (5.1)
  repeat for $j = 1, 2, \ldots$
    $\sigma_{k,j} = \mu^{j-1}\sigma_{k,0}$;    (5.2)
    $u^{k,j} \in \operatorname{prox}_{\sigma_{k,j}^{-1} f}(u^k - \sigma_{k,j}^{-1}\nabla_u \varphi(u^k, v^k, y^k))$;    (5.3)
  until
  $\varphi(u^{k,j}, v^k, y^k) - \varphi(u^k, v^k, y^k) - \langle \nabla_u \varphi(u^k, v^k, y^k), u^{k,j} - u^k \rangle \le \frac{\sigma_{k,j} - \nu}{2} \left\| u^{k,j} - u^k \right\|^2$;    (5.4)
Output: $\sigma_{k+1} = \sigma_{k,J_k}$, $u^{k+1} = u^{k,J_k}$, where $J_k$ is the index $j$ of the last loop iteration.
```

5 The Proximal Parameter Backtracking Procedure

In this section we propose a tractable backtracking procedure to generate proximal parameters that meet the requirements posed by our analysis in Sect. 4.1 (note that (R2) is an implicit requirement):

(R1) The sequence $\{\sigma_k\}_{k \ge 1}$ satisfies the descent condition of Assumption A.
(R2) The backtracking procedure completes in a finite time.
(R3) The sequence $\{\sigma_k\}_{k \ge 1}$ is guaranteed to be bounded, i.e., Assumption B holds true whenever the sequence $\{(u^k, v^k, y^k)\}_{k \ge 1}$ is bounded.

The backtracking procedure is described by Algorithm 2, which comprises a loop that increases the candidate for the proximal parameter until a sufficient decrease property is met.

The generated proximal parameter satisfies (R1) by definition. We will focus now on proving that it also meets requirements (R2) and (R3), starting with (R2). Our derivations rely on the theory of Lipschitz functions, mainly on the fact that $\nabla \varphi$ is locally Lipschitz continuous (cf. Lemma 4.1); we defer the reader to some useful review/results on Lipschitz and local Lipschitz continuity in Appendix B.

Lemma 5.1 *(Finite termination) The backtracking procedure in Algorithm 2 terminates in a finite number of iterations.*

Proof Let $k \ge 1$. For every $j \ge 0$, define

$$\begin{aligned} S_{k,j} &:= \operatorname{prox}_{\sigma_{k,j}^{-1} f}(u^k - \sigma_{k,j}^{-1} \nabla_u \varphi(u^k, v^k, y^k)) \\ &= \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ f(x) + \langle \nabla_u \varphi(u^k, v^k, y^k), u \rangle + \frac{\sigma_{k,j}}{2} \left\| u - u^k \right\|^2 \right\}. \end{aligned}$$

Note that $\sigma_{k,j} \ge \sigma_0 > 0$, as can be easily verified by steps (5.1) and (5.2). Thus, $S_{k,j}$ is nonempty and compact (cf. Remark 3.2); specifically, with $r_{k,j} := \sup\{\left\| x - u^k \right\| : x \in$

$S_{k,j}\}$, we have $r_{k,j} < \infty$. We will now show that as we increase the proximal parameter $\sigma_{k,j} \to \sigma_{k,j+1}$, the distance from the solution set $S_{k,j}$ to u^k does not increase, i.e., $r_{k,j} \leq r_{k,1}$ for any $j \geq 0$.

For any $j > 1$, set $x^j \in S_{k,j}$. Then

$$\begin{aligned} & f(x^i) + \langle \nabla_u \varphi(u^k, v^k, y^k), x^i \rangle + \frac{\sigma_{k,i}}{2} \left\| x^i - u^k \right\|^2 \\ & \leq f(x^1) + \langle \nabla_u \varphi(u^k, v^k, y^k), x^1 \rangle + \frac{\sigma_{k,i}}{2} \left\| x^1 - u^k \right\|^2 \end{aligned}$$

and

$$\begin{aligned} & f(x^1) + \langle \nabla_u \varphi(u^k, v^k, y^k), x^1 \rangle + \frac{\sigma_{k,1}}{2} \left\| x^1 - u^k \right\|^2 \\ & \leq f(x^i) + \langle \nabla_u \varphi(u^k, v^k, y^k), x^i \rangle + \frac{\sigma_{k,1}}{2} \left\| x^i - u^k \right\|^2 . \end{aligned}$$

Consequently, noting that $\sigma_{k,j} > \sigma_{k,1}$, we can sum the equations above, divide by $(\sigma_{k,j} - \sigma_{k,1})/2$, and obtain that

$$\left\| x^j - u^k \right\| \leq \left\| x^1 - u^k \right\| \leq r_{k,1}.$$

Thus,

$$\{(u^{k,j}, v^k, y^k)\}_{j \geq 0} \subseteq \mathcal{B}[(u^k, v^k, y^k), r_{k,1}],$$

where $\mathcal{B}[(u^k, v^k, y^k), r_{k,1}]$ is the Euclidean norm ball centered at (u^k, v^k, y^k) with radius $r_{k,1}$. Since $\nabla \varphi$ is locally Lipschitz continuous (cf. Lemma 4.1), we in particular have that for the nonempty, convex, and compact set $\mathcal{B}[(u^k, v^k, y^k), r_{k,1}]$, there exists $L_k \geq 0$ such that

$$\|\nabla_u \varphi(x) - \nabla_u \varphi(z)\| \leq L_k \|x - z\|, \quad \forall x, z \in \mathcal{B}[(u^k, v^k, y^k), r_{k,1}].$$

Subsequently, the descent lemma (cf. Lemma 4.3) holds true over $\mathcal{B}[(u^k, v^k, y^k), r_{k,1}]$, meaning that the procedure terminates after at most $1 + \max\{0, \lceil (\log(L_k + \nu) - \log \sigma_{k,0}) / \log \mu \rceil\}$ iterations, with $\sigma_{k+1} \leq \max\{\sigma_k, \mu(L_k + \nu)\}$, as the iteration prior to the last must satisfy that $\sigma_{k,J_k}/\mu - \nu = \sigma_{k,J_k-1} - \nu < L_k$. □

Next we establish that the proximal parameters generated via Algorithm 2 satisfy (R3) whenever the sequence $\{u^k, v^k, y^k\}_{k\geq 1}$ is bounded.

Lemma 5.2 *(Boundedness of the proximal parameter) Suppose that the sequence $\{u^k, v^k, y^k\}_{k\geq 1}$ is bounded. Additionally, for $\mathcal{U}_k := \{u^{k,j} : j = 1, 2, \ldots, J_k\}$, assume that $\bigcup_{k=0}^{\infty} \mathcal{U}_k$ is bounded. Then, the proximal parameter sequence $\{\sigma_k\}_{k\geq 1}$ is bounded.*

Proof It is clear that $\inf_{k\geq 1}\sigma_k \geq \sigma_0 > 0$ (cf. (5.1) and (5.2)), and therefore it is enough to prove that

$$\sup_{k\geq 1}\sigma_k < \infty. \tag{5.5}$$

First note that $u^{k+1} \in \mathcal{U}_k$ as $u^{k+1} = u^{k,J_k}$, for every $k \geq 1$. Next, set $C_u := \mathrm{cl\,conv}\left(\{u^0\} \cup \bigcup_{k=0}^{\infty}\mathcal{U}_k\right)$, $C_v = \mathrm{cl}\,\{v^k\}_{k\geq 0}$ and $C_y = \mathrm{cl}\,\{y^k\}_{y\geq 0}$. Then, the set $C := C_u \times C_v \times C_y$ is nonempty and compact and as $\nabla_u\varphi$ is locally Lipschitz continuous (cf. Lemma 4.1) it follows (cf. Proposition B.1) that there exists $L_C \in \mathbb{R}_+$ such that $\nabla_u\varphi$ is L_C-Lipschitz continuous over C, and specifically, for every $k \geq 0$, $\nabla_u\varphi(\cdot, v^k, y^k)$ is L_C-Lipschitz continuous over the convex and compact set C_u. Hence, we can apply the descent Lemma (cf. Lemma 4.3) and obtain that, for every $\sigma \geq L_C + \nu$, every $k \geq 1$, and every $j \in \{1, 2, \ldots, J_k\}$, we have

$$\begin{aligned}&\varphi(u^{k,j}, v^k, y^k) - \varphi(u^k, v^k, y^k) - \langle \nabla_u\varphi(u^k, v^k, y^k), u^{k,j} - u^k\rangle \\ &\leq \frac{\sigma - \nu}{2}\left\|u^{k,j} - u^k\right\|^2.\end{aligned} \tag{5.6}$$

To obtain (5.5), we prove by induction that, for every $k \geq 1$, we have

$$\sigma_k \leq \max\{\sigma_0, \mu(L_C + \nu)\}. \tag{5.7}$$

For $k = 0$, we have $\sigma_k = \sigma_0$ and so (5.7) trivially holds. Next, consider the call to the backtracking procedure by DAM at iteration k and assume by contradiction that

$$\sigma_{k+1} > \max\{\sigma_0, \mu(L_C + \nu)\}. \tag{5.8}$$

Steps (5.1) and (5.2) together with the induction hypothesis imply that

$$\sigma_{k,1} \leq \sigma_k \leq \max\{\sigma_0, \mu(L_C + \nu)\}. \tag{5.9}$$

As $\sigma_{k+1} = \sigma_{k,J_k}$ then (5.8) and (5.9) imply that $J_k \geq 2$. In addition, as $\sigma_{k,J_k-1} = \sigma_{k,J_k}/\mu$ (cf. (5.2)), inequality (5.8) implies that $\sigma_{k,J_k-1} > L_C + \nu$. Then, together with (5.6), it follows that procedure's termination condition (5.4) holds true for $j = J_k - 1$ which is obviously a contradiction as it should hold true only for $j = J_k$. □

We conclude this section with the following remark when *global Lipschitz* constants are available.

Remark 5.1 (Global Lipschitz continuity) If the function g has an L_g-Lipschitz continuous gradient, and the mapping F is l_F-Lipschitz continuous with an L_F-Lipschitz continuous Jacobian. Then the backtracking procedure in Algorithm 2 can be replaced with

$$\sigma_{k+1} = \nu + L_g + L_F\left\|y^k + \rho(F(u^k) - Gv^k)\right\| + \rho l_F^2.$$

It can easily be verified that the above proximal parameter generator satisfies the descent condition of Assumption A, and that when the sequence $\{(u^k, v^k, y^k)\}_{k\geq 1}$ is bounded, then Assumption B also holds, i.e., $\{\sigma_k\}_{k\geq 1}$ is bounded.

6 Implication to the Proximal Gradient Method Under Locally Lipschitz Data

As an interesting byproduct of our analysis, we can immediately obtain convergence results of the proximal gradient scheme with locally Lipschitz data. More precisely, in this section we briefly review the special case when $h \equiv 0$, $F \equiv 0$, meaning that model described through (M) reduces to the minimization of the nonconvex sum composite minimization problem

$$\min_{u\in\mathbb{R}^n} f(u) + g(u),$$

where $f : \mathbb{R}^n \to (-\infty, \infty]$ is a proper and lsc function, and $g : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function with a *locally* Lipschitz continuous gradient. Alternatively, by setting G to be the identity map, the model (M) reduces to the nonlinear composite model $\min\{f(u) + g(u) + h(F(u)) : \ u \in \mathbb{R}^n\}$. In the later case, one can verify that under our local Lipschitz assumptions for (M), the gradient of the composite term $h \circ F$ is also locally Lipschitz. Thus in both special cases, we minimize the sum composite model:

$$\min\{f(u) + \psi(u) : \ u \in \mathbb{R}^n\},$$

where $\psi(u) := g(u)$ or $\psi(u) := g(u) + h(F(u))$, both having a locally Lipschitz gradient.

In this case, Algorithm 1 is effectively a proximal gradient scheme with a dynamic proximal parameter, described below.

Prox-Grad with locally Lipschitz gradient (PG-Loc).

1. Generate a proximal parameter $\sigma_{k+1} > 0$;
2. Set $u^{k+1} \in \operatorname{argmin}_{u\in\mathbb{R}^n} \left\{ f(u) + \langle \nabla\psi(u^k), u\rangle + \frac{\sigma_{k+1}}{2} \left\| u - u^k \right\|^2 \right\}$.

According to our convergence result in Theorem 4.2, this scheme has global convergence guarantees even though ψ only has a local Lipschitz continuous gradient.

Theorem 6.1 *(Subsequence and global convergence of PG-Loc) Suppose that Assumption A and Assumption B hold true, and that the sequence $\{u^k\}_{k\geq 1}$ generated by PG-Loc is bounded. Then the following implications hold.*

(i) Let Ω be the set of cluster points of the sequence $\{u^k\}_{k\geq 1}$. Then, Ω is a nonempty, compact, and connected set; $\Omega \subset \operatorname{crit}(f + \psi)$; $\lim_{k\to\infty} \operatorname{dist}(u^k, \Omega) = 0$; and $(f + \psi)$ is finite and constant on Ω.

(ii) Assume, in addition, that the functions f and ψ are semi-algebraic. Then, $\{u^k\}_{k\geq 1}$ has a finite length, i.e., $\sum_{k=1}^{\infty} \left\| u^{k+1} - u^k \right\| < \infty$, and it converges to a critical point of the objective function $u^* \in \operatorname{crit}(f + \psi)$.

We can then invoke Algorithm 2 as the proximal parameter generator in this case to derive a tractable, globally convergent, proximal gradient method with a backtracking procedure, for composite problems with locally Lipschitz continuous gradient.

7 Concluding Remarks

Under mild, yet specific, data assumptions and structure, we have established subsequence and global convergence results by using a dynamic backtracking mechanism to determine the proximal parameter. The suggested algorithm DAM, together with a minor refinement of the analysis, can also be applied to a model where G has only full column rank, i.e., $G^T G$ is positive definite, but GG^T is rank deficient. In this case, we should explicitly require having that the range of the mapping F is contained in the range of the matrix G. One may also consider other alternatives for the v-update step: (i) linearize $\varphi(u^{k+1}, \cdot, y^k)$ and leave h intact; (ii) linearize Ψ_k, i.e., linearize both h and $\varphi(u^{k+1}, \cdot, y^k)$. However, these alternatives are less attractive since they impose substantial additional restrictions on the model at hand and specifically on the matrix G. Finally, let us mention that if we do not linearize h in the optimization process, then h must be 'prox-friendly', a requirement that limits the cases where the step's minimization subproblem is tractable. Actually, in that scenario, it can be easily verified that the convergence analysis we conduct in Sect. 4 can even be further simplified.

Acknowledgements The research of E. Cohen was supported by a Ph.D fellowship under ISF Grants 1844-16 and 2619-20 and DFG grant 800240. The research of M. Teboulle was partially supported by the Israel Science Foundation, under ISF Grants 1844-16 and 2619-20, and by the German Research Foundation under DFG Grant 800240.

A Appendix A: Subdifferential Calculus and the Lagrangian

Model (M) is nonconvex and nonsmooth in general, and so we turn to subdifferential calculus for nonconvex functions. Specifically, we recall the definition of subdifferentials, critical points, and a few related results. For a more detailed presentation, see [22,23,26].

Definition A.1 (Subdifferentials) Let $\mathbb{E}$ be an Euclidean vector space, $\psi : \mathbb{E} \to (-\infty, \infty]$ be a proper and lsc function, and $z \in \operatorname{dom} \psi$.

(i) The *Fréchet (regular) subdifferential* of ψ at z, denoted by $\hat{\partial}\psi(z)$, is the set of all vectors $v \in \mathbb{E}$ satisfying

$$\psi(x) \geq \psi(z) + \langle v, x - z \rangle + o\left(\|x - z\|\right). \tag{A.1}$$

(ii) The *Mordukhovich (limiting) subdifferential* of ψ at z, denoted by $\partial\psi(z)$, is the set of all vectors $v \in \mathbb{E}$ such that there exist sequences $\{z^k\}_{k\in\mathbb{N}}$ and $\{v^k\}_{k\in\mathbb{N}}$ where $z^k \to z$, $\psi(z^k) \to \psi(z)$, $v^k \in \hat{\partial}\psi(z^k)$, and $v^k \to v$.

(iii) The *horizon subdifferential* of ψ at z, denoted by $\partial^{\infty}\psi z$, has a similar definition to that in (ii), but instead of $v^k \to v$ we have $t^k v^k \to v$ for some real sequence $t^k \searrow 0$.

For $z \notin \operatorname{dom}\psi$ we set $\hat{\partial}\psi(z) = \partial\psi(z) = \partial^{\infty}\psi(z) = \emptyset$.

Note that (A.1) is equivalent to

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

Remark A.1 (Closedness of graph of $\partial\psi$) We note, as in [7, Remark 1(ii)], that given a convergent sequence $(x^k, w^k) \xrightarrow[k\to\infty]{} (x, w)$, such that $w^k \in \partial\psi(x^k)$ and $\lim_{k\to\infty} \psi(x^k) = \psi(x)$, it holds that $w \in \partial\psi(x)$.

Proposition A.1 *(see Exercise 8.8(c), p.304 in [26]) Let $\psi : \mathbb{R}^d \to (-\infty, \infty]$ be an extended valued function and $\phi : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. Then,*

$$\partial(\psi + \phi)(x) = \partial\psi(x) + \nabla\phi(x), \quad \forall x \in \mathbb{R}^d,$$

where for $x \notin \operatorname{dom}\psi$, we note that $\emptyset + \nabla\phi(x) = \emptyset$.

Definition A.2 (Critical points) Let $\mathbb{E}$ be an Euclidean vector space, $\psi : \mathbb{E} \to (-\infty, \infty]$ be a proper and lsc function. Then, the set of *critical points* of ψ is defined by

$$\operatorname{crit}\psi := \{x : 0 \in \partial\psi(x)\}.$$

B Appendix B: Lipschitz and Local Lipschitz Continuity

We review the notion of Lipschitz and local Lipschitz continuity in the context of model (M). Thus, we restrict our discussion to Euclidean vector spaces and in the following $\mathbb{X}$ and $\mathbb{Y}$ denote such spaces.

We begin with the basic definition.

Definition B.1 (Lipschitz and locally Lipschitz continuity) Let $S \subseteq \mathbb{X}$ be a nonempty set and $\phi : S \to \mathbb{Y}$ be a continuous mapping over S. Then,

(i) ϕ is *L-Lipschitz continuous* over S with $L \geq 0$ if the following holds

$$\|\phi(x) - \phi(z)\| \leq L\,\|x - z\|, \quad \forall x, z \in S. \tag{B.1}$$

(ii) ϕ is *locally Lipschitz continuous* over S if for every $z \in S$ there exist $\varepsilon(z) > 0$, $L_\varepsilon(z) \geq 0$, and a neighborhood $\mathcal{N}_\varepsilon(y) := \{x \in S : \|x - z\| < \varepsilon(z)\}$, such that ϕ is $L_\varepsilon(z)$-Lipschitz continuous over $\mathcal{N}_\varepsilon(z)$, i.e.,

$$\|\phi(x) - \phi(z)\| \leq L_\varepsilon(z) \, \|x - z\| \,, \quad \forall x, z \in \mathcal{N}_\varepsilon(z). \tag{B.2}$$

When (i) or (ii) hold with $S \equiv \mathbb{X}$ then ϕ is referred to as *L-Lipschitz continuous* or *locally Lipschitz continuous*, respectively. In addition, note that when $\phi(x) - \phi(z)$ is a matrix then $\|\phi(x) - \phi(z)\|$ is the spectral norm.

We also recall the following characterization of locally Lipschitz continuous mappings.

Proposition B.1 *(Local Lipschitz continuity and compact sets (Theorem 2.1.6 in [13])) Let $S \subseteq \mathbb{X}$ be a nonempty set. Then, a mapping $\phi : \mathbb{X} \to \mathbb{Y}$ is locally Lipschitz continuous over S if and only if for every nonempty and compact set $C \subseteq S$ there exists $L_C \geq 0$ such that ϕ is L_C-Lipschitz continuous over C, i.e.,*

$$\|\phi(x) - \phi(z)\| \leq L_C \, \|x - z\| \,, \quad \forall x, z \in C. \tag{B.3}$$

We conclude with a proposition which deals with Lipschitz continuity properties of C^1 mappings.

Proposition B.2 *(Differential mappings and Lipschitz continuity) Let $\phi : \mathbb{X} \to \mathbb{Y}$ be a C^1 mapping. Then the following claims hold.*

(i) ϕ is locally Lipschitz continuous;

(ii) Let $B \subseteq \mathbb{X}$ be a closed ball, i.e., $B = \{x : \|x - z\| \leq r\}$, for some $z \in \mathbb{X}$ and $r \in (0, \infty]$, and assume that ϕ is L_B-Lipschitz continuous over B, with $L_B \geq 0$. Then,

$$\|\nabla \phi(x)\| \leq L_B, \ \forall x \in B.$$

Proof For the first claim, see, e.g., [12, Corollary, p. 32]. The second claim can be easily obtained using the definition of the Gâteaux derivative, see [12, p. 30], and the continuity of $\nabla \phi$. □

C Appendix C: Proof of Lemma 4.1

Proof of Lemma 4.1 We prove that $\nabla \varphi$ is locally Lipschitz continuous by utilizing the local Lipschitz continuity characterization stated in Proposition B.1, i.e., given a nonempty and compact set $C \subset \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^q$ we prove that there exists $L_C \in \mathbb{R}_+$ such that for every $\omega = (u, v, y)$, $\hat{\omega} = (\hat{u}, \hat{v}, \hat{y}) \in C$, we have

$$\left\| \nabla \varphi(\hat{\omega}) - \nabla \varphi(\omega) \right\| \leq L_C \left\| \hat{\omega} - \omega \right\|. \tag{C.1}$$

With $\nabla\varphi(\omega) = \big(\nabla_u\varphi(\omega), \nabla_v\varphi(\omega), \nabla_y\varphi(\omega)\big)$, we intend to prove that there exist $L_{C,u}$, $L_{C,v}$, and $L_{C,y} \in \mathbb{R}_+$, such that

$$\left\|\nabla_u\varphi(\hat{\omega}) - \nabla_u\varphi(\omega)\right\| \le L_{C,u} \left\|\hat{\omega} - \omega\right\|, \tag{C.2}$$

$$\left\|\nabla_v\varphi(\hat{\omega}) - \nabla_v\varphi(\omega)\right\| \le L_{C,v} \left\|\hat{\omega} - \omega\right\|, \tag{C.3}$$

$$\left\|\nabla_y\varphi(\hat{\omega}) - \nabla_y\varphi(\omega)\right\| \le L_{C,y} \left\|\hat{\omega} - \omega\right\|. \tag{C.4}$$

and as

$$\begin{aligned}\left\|\nabla\varphi(\hat{\omega}) - \nabla\varphi(\omega)\right\| &\le \left\|\nabla_u\varphi(\hat{\omega}) - \nabla_u\varphi(\omega)\right\| \\ &+ \left\|\nabla_v\varphi(\hat{\omega}) - \nabla_v\varphi(\omega)\right\| + \left\|\nabla_y\varphi(\hat{\omega}) - \nabla_y\varphi(\omega)\right\|,\end{aligned}$$

we can set $L_C = L_{C,u} + L_{C,v} + L_{C,y}$.

We begin with $\nabla_u\varphi$ (cf. (4.1)) and note that for every $(u, v, y), (\hat{u}, \hat{v}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^q$, we have

$$\begin{aligned}&\left\|\nabla_u\varphi(\hat{u}, \hat{v}, \hat{y}) - \nabla_u\varphi(u, v, y)\right\| \\ &= \left\|\nabla g(\hat{u}) + \nabla F(\hat{u})^T \big(y + \rho(F(\hat{u}) - G\hat{v})\big) - \nabla g(u) - \nabla F(u)^T (y + \rho(F(u) - Gv))\right\| \\ &\le \left\|\nabla g(\hat{u}) - \nabla g(u)\right\| + \left\|\nabla F(\hat{u})^T \big(\hat{y} + \rho(F(\hat{u}) - G\hat{v})\big) - \nabla F(u)^T (y + \rho(F(u) - Gv))\right\|,\end{aligned} \tag{C.5}$$

where the inequality is due to the triangle inequality.

Next, let C_u be the projection of C on $\mathbb{R}^n$ and B_u be a compact ball such that $C_u \subseteq B_u$. Then, as ∇g is locally Lipschitz continuous and B_u is compact, we can apply Proposition B.1 and obtain that there exists $L_g \ge 0$ such that

$$\left\|\nabla g(\hat{u}) - \nabla g(u)\right\| \le L_g \left\|\hat{u} - u\right\|, \quad \forall u, \hat{u} \in B_u.$$

Next, recall that F is C^1 and hence, by Proposition B.2(i), F is locally Lipschitz continuous. In addition, ∇F is also locally Lipschitz continuous. Thus, there exist $l_F \ge 0$ and $L_F \ge 0$, such that for every $u, \hat{u} \in B_u$, we have

$$\left\|F(\hat{u}) - F(u)\right\| \le l_F \left\|\hat{u} - u\right\|, \ \left\|\nabla F(\hat{u}) - \nabla F(u)\right\| \le L_F \left\|\hat{u} - u\right\|, \text{ and } \|\nabla F(u)\| \le l_F,$$

where the first two inequalities are due to Proposition B.1 and the last is due to Proposition B.2(ii). Applying the above inequalities, we obtain that, for every $u, \hat{u} \in B_u$, $v, \hat{v} \in \mathbb{R}^m$, and $y, \hat{y} \in \mathbb{R}^q$, we have

$$\begin{aligned}
&\left\|\nabla F(\hat{u})^T \left(\hat{y} + \rho(F(\hat{u}) - G\hat{v})\right) - \nabla F(u)^T \left(y + \rho(F(u) - Gv)\right)\right\| \\
&= \| \left(\nabla F(\hat{u})^T - \nabla F(u)^T\right)(y + \rho(F(u) - Gv)) \\
&\quad + \nabla F(\hat{u})^T \left(\hat{y} - y + \rho(F(\hat{u}) - F(u)) - \rho G(\hat{v} - v)\right) \| \\
&\le \left\|\nabla F(\hat{u}) - \nabla F(u)\right\| \left\|y + \rho(F(u) - Gv)\right\| \\
&\quad + \left\|\nabla F(\hat{u})\right\| \left(\left\|\hat{y} - y\right\| + \rho \left\|F(\hat{u}) - F(u)\right\| + \rho\left\|G\right\|\left\|\hat{v} - v\right\|\right) \\
&\le \left(L_F\left\|y{+}\rho(F(u) - Gv)\right\| + \rho l_F^2\right)\left\|\hat{u} - u\right\| {+} l_F \left(\left\|\hat{y} - y\right\| + \rho\left\|G\right\| \left\|\hat{v} - v\right\|\right),
\end{aligned}$$

Combining the results above, and noting that $C \subseteq B_u \times \mathbb{R}^m \times \mathbb{R}^q$, we obtain that, for every $\omega = (u, v, y)$, $\hat{\omega} = (\hat{u}, \hat{v}, \hat{y}) \in C$, we have

$$\begin{aligned}
\left\|\nabla_u\varphi(\hat{\omega}) - \nabla_u\varphi(\omega)\right\| &\le \left(L_g + L_F \left\|y + \rho(F(u) - Gv)\right\| + \rho l_F^2\right) \\
&\quad \left\|\hat{u} - u\right\| + l_F \left(\left\|\hat{y} - y\right\| + \rho \left\|G\right\| \left\|\hat{v} - v\right\|\right) \\
&\le L_\varphi(u, v, y) \left\|\hat{\omega} - \omega\right\|,
\end{aligned}$$

with $L_\varphi(u, v, y) := \left(L_g + L_F \left\|y + \rho(F(u) - Gv)\right\| + l_F(1 + \rho l_F + \rho \left\|G\right\|)\right) \left\|\hat{\omega} - \omega\right\|$. We complete the proof for $\nabla_u\varphi$ and obtain inequality (C.2) by setting $L_{C,u} = \sup_{(u,v,y)\in C} L_\varphi(u, v, y)$ and noting that $L_{C,u} < \infty$ as L_φ is continuous and C is compact.

Next, we examine $\nabla_v\varphi$ (cf. (4.2)). For every $\omega, \hat{\omega} \in C$, we have

$$\begin{aligned}
\left\|\nabla_v\varphi(\hat{\omega}) - \nabla_v\varphi(\hat{\omega})\right\| &= \left\|G^T \left(\hat{y} - y + \rho(F(\hat{u}) - F(u)) - \rho(G(\hat{v} - v))\right)\right\| \\
&\le \left\|G\right\| \left(\left\|\hat{y} - y\right\| + \rho \left\|F(\hat{u}) - F(u)\right\| + \rho \left\|G\right\| \left\|\hat{v} - v\right\|\right) \\
&\le \left\|G\right\| \left(\left\|\hat{y} - y\right\| + \rho l_F \left\|\hat{u} - u\right\| + \rho \left\|G\right\| \left\|\hat{v} - v\right\|\right) \\
&\le L_{C,v} \left\|\hat{\omega} - \omega\right\|,
\end{aligned}$$

with $L_{C,v} = \left\|G\right\| (1 + \rho(l_F + \left\|G\right\|))$.

Finally, with $\nabla_y\varphi$ (cf. (4.3)), for every $\omega, \hat{\omega} \in C$, we have

$$\begin{aligned}
\left\|\nabla_y\varphi(\hat{\omega}) - \nabla_y\varphi(\omega)\right\| &= \left\|F(\hat{u}) - F(u) - G(\hat{v} - v)\right\| \\
&\le \left\|F(\hat{u}) - F(u)\right\| + \left\|G(\hat{v} - v)\right\| \\
&\le l_F \left\|\hat{u} - u\right\| + \left\|G\right\| \left\|\hat{v} - v\right\| \\
&\le L_{C,y} \left\|\hat{\omega}) - \omega\right\|,
\end{aligned}$$

with $L_{C,y} = l_F + \left\|G\right\|$. □

References

1. Beck, A., Teboulle, M.: Gradient-based algorithms with applications to signal recovery problems. In: Palomar, D., Eldar, Y.C. (eds.) Convex Optimization in Signal Processing and Communications, pp. 139–162. Cambridge University Press, Cambridge (2009)
2. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Belmont, MA (1999)
3. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Athena Scientific, Belmont, MA (1996)
4. Bertsekas, D.P.: Convex Optimization Algorithms. Athena Scientific, Belmont, MA (2015)
5. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods. Athena Scientific (2003)
6. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Opt. **17**(4), 1205–1223 (2007)
7. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1–2), 459–494 (2014)
8. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. SIAM J. Opt. **28**, 2131–2151 (2018)
9. Bolte, J., Sabach, S., Teboulle, M.: Nonconvex Lagrangian-based optimization: monitoring schemes and global convergence. Math. Op. Res. **43**(4), 1210–1232 (2018)
10. Boţ, R.I., Nguyen, D.K.: The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. Math. Op. Res. **45**(2), 682–712 (2020)
11. Boyd, S., Parikh, N., Chu, E.: Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Now Publishers Inc (2011)
12. Clarke, F.H.: Optimization and Nonsmooth Analysis. SIAM (1990)
13. Cobzaş, Ş, Miculescu, R., Nicolae, A.: Lipschitz Functions, vol. 2241. Springer (2019)
14. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
15. Fessler, J.A.: Optimization methods for magnetic resonance image reconstruction: key models and optimization algorithms. IEEE Signal Process. Mag. **37**(1), 33–40 (2020)
16. Gabay, G., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)
17. Glowinski, R., Le Tallec, P.: Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics, vol. 9. SIAM (1989)
18. Hestenes, M.R.: Multiplier and gradient methods. J. Opt. Theory Appl. **4**(5), 303–320 (1969)
19. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Opt. **25**(4), 2434–2460 (2015)
20. Luke, D.R., Sabach, S., Teboulle, M.: Optimization on spheres: models and proximal algorithms with computational performance comparisons. SIAM J. Math. Data Sci. **1**(3), 408–445 (2019)
21. Mei, S., Bai, Y., Montanari, A.: The landscape of empirical risk for nonconvex losses. Annal. Stat. **46**(6A), 2747–2774 (2018)
22. Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation, I: Basic Theory, II: Applications. Springer, Berlin (2006)
23. Mordukhovich, B.S.: Variational Analysis and Applications. Springer, Cham, Switzerland (2018)
24. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: Fletcher, R. (ed.) Optimization, pp. 283–298. Academic Press, New York, NY (1969)
25. Rockafellar, R.T.: Augmented Lagrange multiplier functions and duality in nonconvex programming. SIAM J. Control **12**(2), 268–285 (1974)
26. Rockafellar, R.T., Wets, J.B.R.: Variational Analysis. Springer, Berlin (2004)
27. Royset, J. O.: Variational Analysis in Modern Statistics. Special Issue Mathematical Programming, Series B, Volume 174 (2019)
28. Sabach, S., Teboulle, M.: Lagrangian methods for composite optimization. In Handbook of Numerical Analyis. Edited by Ron Kimmel, Xue-Cheng Tai, Volume 20, 401–436. Elsevier (2019)
29. Shefi, R., Teboulle, M.: Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. SIAM J. Opt. **24**(1), 269–297 (2014)
30. Teboulle, M.: A simplified view of first order methods for optimization. Math. Program. **170**(1), 67–96 (2018)

31. Von Stackelberg, H.: Market Structure and Equilibrium. Springer Science & Business Media, Berlin (2010)