https://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–28
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Alternating and Parallel Proximal Gradient Methods for Nonsmooth, Nonconvex Minimax: A Unified Convergence Analysis

Eyal Cohen,[a] **Marc Teboulle**[a,*]

[a] School of Mathematical Sciences, Tel Aviv University, Ramat Aviv 69978, Israel
*Corresponding author
Contact: cneyal@gmail.com (EC); teboulle@tauex.tau.ac.il, https://orcid.org/0000-0002-4228-131X (MT)

Received: October 10, 2022
Revised: August 27, 2023
Accepted: January 1, 2024
Published Online in Articles in Advance: February 8, 2024

MSC2020 Subject Classifications: Primary: 90C26, 90C47, 49K35, 65K05

https://doi.org/10.1287/moor.2022.0294

Abstract. There is growing interest in nonconvex minimax problems that is driven by an abundance of applications. Our focus is on nonsmooth, nonconvex-strongly concave minimax, thus departing from the more common weakly convex and smooth models assumed in the recent literature. We present proximal gradient schemes with either parallel or alternating steps. We show that both methods can be analyzed through a single scheme within a unified analysis that relies on expanding a general convergence mechanism used for analyzing nonconvex, nonsmooth optimization problems. In contrast to the current literature, which focuses on the complexity of obtaining nearly approximate stationary solutions, we prove subsequence convergence to a critical point of the primal objective and global convergence when the latter is semialgebraic. Furthermore, the complexity results we provide are with respect to approximate stationary solutions. Lastly, we expand the scope of problems that can be addressed by generalizing one of the steps with a Bregman proximal gradient update, and together with a few adjustments to the analysis, this allows us to extend the convergence and complexity results to this broader setting.

Funding: The research of E. Cohen was partially supported by a doctoral fellowship from the Israel Science Foundation [Grant 2619-20] and Deutsche Forschungsgemeinschaft [Grant 800240]. The research of M. Teboulle was partially supported by the Israel Science Foundation [Grant 2619-20] and Deutsche Forschungsgemeinschaft [Grant 800240].

Keywords: nonconvex nonsmooth minimax • nonsmooth minimization-maximization • proximal gradient method • Kurdyka–Lojasiewicz property • Bregman distance • global convergence • convergence rate

1. Introduction

The research on minimax optimization problems has gone a long way since von Neumann's seminal work (von Neumann [44]), which presented the *minimax theorem* and practically established the field of game theory. These problems can be found in various additional realms, including robust optimization in Ben-Tal and Nemirovski [7] and Ben-Tal et al. [8], signal processing in Razaviyayn et al. [39], and machine learning with applications such as robust deep learning and generative adversarial networks (GANs) in Goodfellow et al. [25]. A general minimax problem can be stated as

$$\min_{u\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} \{K(u,v) := f(u) + c(u,v) - g(v)\}. \tag{M}$$

Much research was devoted to the *convex-concave* setting, where it is assumed that $f(\cdot)$ and $g(\cdot)$ are convex and $c(\cdot,\cdot)$ convex in u and concave in v (see, for instance, the works of Auslender and Teboulle [2], Korpelevich [26], Mokhtari et al. [32], and Nemirovski [34] and references therein), and to the primal-dual methods (Chambolle and Pock [15], Cohen et al. [18], Drori et al. [23], Pock et al. [37]), which require $c(\cdot,\cdot)$ to be bilinear. In such a setting and subject to appropriate constraints qualification conditions (see, e.g., Rockafellar [40, section 37]), there exists a saddle point (also referred to as a Nash equilibrium); i.e., a point (u^*, v^*) that satisfies the condition

$$K(u^*,v) \le K(u^*,v^*) \le K(u,v^*), \quad \forall u \in \mathbb{R}^n,\ \forall v \in \mathbb{R}^m, \tag{1.1}$$

and thus, also ensures having a zero duality gap.

Recently, there is a great interest in nonconvex minimax optimization problems, where the driving force originates in modern applications, such as robust deep learning, GANs, fair statistical inference, and distributed processing (see the recent review in Razaviyayn et al. [39] and references therein). The difficulty in the nonconvex setting

is that the zero duality gap is not guaranteed anymore; so, a saddle point may not exist, or otherwise, its computation may be intractable (see, e.g., Daskalakis et al. [20]).

The recent literature (see, e.g., Barazandeh and Razaviyayn [3], Diakonikolas et al. [22], Fiez et al. [24], Lin et al. [28], Liu et al. [29], Lu et al. [31], Nouiehed et al. [35], Ostrovskii et al. [36], Rafique et al. [38], Razaviyayn et al. [39], Thekumparampil et al. [43]) considers *weakly convex-weakly concave* problems, where the functions $f(\cdot)$ and $g(\cdot)$ are still assumed to be convex and $c(\cdot,\cdot)$ is a continuous weakly convex-weakly concave function (i.e., $c(\cdot,v)$ is weakly convex[1] for every $v \in \operatorname{dom} g$, and $-c(u,\cdot)$ is weakly convex for every $u \in \operatorname{dom} f$). The function $c(\cdot,\cdot)$ is also assumed to be *smooth* (i.e., continuously differentiable (C^1) with a Lipschitz continuous gradient (this by itself implies the weak convexity/concavity)), except in Liu et al. [29] and Rafique et al. [38], where the lack of smoothness is, however, balanced by other restrictive assumptions. Finally, we note that most of the works cited, with the exception of Barazandeh and Razaviyayn [3], Lu et al. [31], and Rafique et al. [38], consider the following specialization of (M):

$$\min_{u\in\mathcal{U}} \max_{v\in\mathcal{V}} c(u,v), \tag{C}$$

where $f(\cdot)$ and $g(\cdot)$ are set to be the indicator functions $\delta_{\mathcal{U}}(\cdot)$ and $\delta_{\mathcal{V}}(\cdot)$ of the nonempty, closed, and convex sets $\mathcal{U} \subseteq \mathbb{R}^n$ and $\mathcal{V} \subseteq \mathbb{R}^m$, respectively. This model is restrictive as it does not allow for more general nonsmooth (i.e., beyond indicator functions) and nonconvex objectives.

Because of the intractability of finding a saddle point, the focus is on finding stationary solutions that satisfy certain first-order conditions. There are two main approaches for defining such stationary points; for a detailed review, see Razaviyayn et al. [39] and references therein. The first approach views Model (M) as a game where the order of the minimization and maximization is ignored (despite the lack of a zero duality gap) (see, for instance, Barazandeh et al. [4], Diakonikolas et al. [22], Fiez et al. [24], Liu et al. [29], Lu et al. [31], Nouiehed et al. [35], Ostrovskii et al. [36]), and so, stationarity at a point $(\overline{u},\overline{v})$ is defined with respect to the function $K(u,v)$, separately for u and for v (with the exception of Fiez et al. [24], which also considers second-order conditions). The second approach, which will be adopted in this work, views Model (M) as an optimization problem of minimizing the function $\max_{v\in\mathbb{R}^m} K(\cdot,v)$ and thus, seeks to define and obtain stationary points for the function $\max_v K(\cdot,v)$; see Lin et al. [28], Lu et al. [31], Ostrovskii et al. [36], Rafique et al. [38], and Thekumparampil et al. [43]. More precisely, we seek to find a critical point of the induced *minimization problem* emerging from Model (M):

$$\min_{u\in\mathbb{R}^n} \left\{ \Theta(u) := \max_{v\in\mathbb{R}^m} K(u,v) = f(u) + \phi(u) \right\}, \tag{P}$$

where $\phi : \mathbb{R}^n \to (-\infty,\infty]$ is defined by $\phi(u) := \max_{v\in\mathbb{R}^m}\{c(u,v) - g(v)\}$. That is, we look for a critical point of $\Theta(\cdot)$, $\overline{u} \in \mathbb{R}^n$, such that $0 \in \partial\Theta(\overline{u})$ (cf. Definition 1).

The most popular methods for solving minimax problems are the gradient descent-ascent (GDA)-based algorithms, which in their simplest form, are single-loop iterative schemes with either simultaneous (parallel) or alternating updates of the minimax variables (u, v) (see, e.g., Diakonikolas et al. [22], Fiez et al. [24], Lin et al. [28], Lu et al. [31]). Many of the current suggested methods, however, require nested loops, and as such, they are more complex and require a dedicated analysis. Indeed, some employ another optimization scheme for updating the variables (u, v) (see, e.g., Barazandeh and Razaviyayn [3], Nouiehed et al. [35], Ostrovskii et al. [36]), whereas others solve a sequence of regularized problems (see, e.g., Liu et al. [29], Rafique et al. [38], Thekumparampil et al. [43]).

The focus of this work is on the analysis of the proximal gradient descent-ascent (PGDA) Algorithm PGDA, which is a classical extension of GDA, when applied to a broader nonsmooth, nonconvex setting. We consider two simple minimization-maximization schemes that contain no inner loops and only require two proximal gradient steps at each iteration, with respect to u and with respect to v (cf. Section 2). The schemes differ in the fact that one allows us to execute the steps in parallel, whereas the other is confined to an alternating execution of the two steps. This can be simply described through one unified scheme that uses an auxiliary variable; see Algorithm 1 (PGDA) in Section 2. Assuming that the proximal steps are tractable, then both algorithms are easily implementable. Alternating schemes are often considered as more challenging to analyze than parallel schemes. Nevertheless, we are able to conduct a mostly joint analysis of both algorithms and obtain similar convergence results up to a constant factor.

We begin with a simple question. What are the consequences of relaxing the weak convexity requirement of Model (M) alluded to? Namely, we seek to analyze PGDA when applied to a nonconvex-strongly concave Model (M), which is presented in detail in Section 2. Basically, for $u \in \operatorname{dom} g$, $K(u,\cdot)$ is required to be strongly convex, whereas $K(\cdot,v)$ is a general extended valued proper and lower-semicontinuous (lsc) function, for every $v \in \operatorname{dom} g$.

Like most of the works cited, the coupling function $c(\cdot,\cdot)$ is assumed to be smooth (i.e., C^1 with a Lipschitz continuous gradient). Such a model allows us to consider genuine nonconvex, nonsmooth problems in the minimizing variable u. For example and without delving into the details of any specific application, we can impose a sparsity constraint or regularization with respect to the primal variable by setting $f(\cdot) \equiv \delta_{\{x:\|x\|_0 \le \lambda\}}(\cdot)$ or $f(\cdot) \equiv \lambda\|\cdot\|_0$, respectively, with $\lambda > 0$ and recalling that the l_0-norm $\|x\|_0$ is defined as the number of the nonzero elements of the vector x.

Our analysis adopts the *minimization problem* perspective of the second approach, but unfortunately, the tools used by the works cited are not adequate for addressing the broader setting in which the function $f(\cdot)$ is neither convex nor weakly convex. To understand this last point, we note that the recent works that consider Model (P) do not prove convergence of algorithms but focus on the *complexity* of obtaining approximate solutions, and this approach relies on an assumption that $K(\cdot,\cdot)$ is weakly convex-concave and that dom g is compact. Indeed, in such a case, $\Theta(\cdot)$, the objective of (P), is l-weakly convex, with $l>0$, and it follows that the corresponding Moreau envelope $\Theta_{1/2l}(\cdot)$ (Moreau [33]) is a smooth function. Thus, following the stationary notions of Davis and Drusvyatskiy [21], the current literature seeks to obtain an ϵ-approximate stationary solution of $\Theta_{1/2l}(\cdot)$ (i.e., a point $\hat{u} \in \mathbb{R}^n$ for which $\|\nabla\Theta_{1/2l}(\hat{u})\| \le \epsilon$). Such a point is a *nearly ϵ-approximate stationary* solution of $\Theta(\cdot)$ as there exists $u \in \mathbb{R}^n$ such that $\operatorname{dist}(0,\partial\Theta(u)) \le \epsilon$ and $\|u-\hat{u}\| \le \epsilon/2l$ (for further details, see, e.g., Lin et al. [28] and Thekumparampil et al. [43]). We emphasize that the behavior of $\Theta(\cdot)$ is characterized near but not at the obtained solution. The structure of the aforementioned analysis collapses when the weak-convexity assumption of Model (M) is removed, and so, a fresh approach is required.

The structure of (P) leads us to consider the convergence mechanism developed in Bolte et al. [11]. However, at this point, we faced a hurdle. The minimax structure of PGDA implies oscillation of the sequences produced by the algorithm and hence, the lack of the crucial sufficient descent property required in Bolte et al. [11]. Consequently, we cannot directly apply the convergence mechanism. Furthermore, overcoming the difficulty by defining an appropriate Lyapunov function, as done in Bolte et al. [11] and Cohen et al. [17], is also not applicable to (P). To tackle this inherent difficulty, in Section 3 we adopt and extend the approach and proof techniques of Bolte et al. [11] with a set of novel conditions that still guarantee global convergence under appropriate conditions. Interestingly, although this approach was motivated by the analysis of PGDA in the minimax context, it should be stressed that the derived result is a general convergence mechanism that could be useful in other contexts as well.

The novel general convergence mechanism allows us to derive under mild assumptions two main convergence-type results for both versions of PGDA, which are absent in the current literature. First and foremost, we prove both subsequence and global convergence of the sequence $\{u^k, v^k\}_{k\in\mathbb{N}}$ generated by either method, and second, we prove iteration complexity (cf. Section 4). More specifically, subject to a standard boundness assumption on the sequence, we prove that for any accumulation point $(\overline{u},\overline{v})$, $\overline{u}$ is a critical point of $\Theta(\cdot)$ (i.e., $0 \in \partial\Theta(\overline{u})$), and $\overline{v}$ maximizes $K(\overline{u},\cdot)$. Then, assuming that the problem data are semialgebraic allows us to obtain a global convergence result. We note that the class of semialgebraic functions is extensive (see Bolte et al. [11] and references therein for various examples), including the l_0-norm. Moreover, we obtain an $O(\epsilon^{-2})$ complexity result, with respect to an ϵ-approximate stationary solution (i.e., a point $\hat{u}$ such that $\operatorname{dist}(\hat{u},\partial\Theta(\hat{u})) \le \epsilon$) and not just to a *nearly* ϵ-approximate stationary solution. We emphasize that the choice of the step sizes for PGDA is driven by the analysis but is not dependent on the choice of the desired accuracy ϵ. The latter only determines the number of sufficient iterations. This is not the case in some of the recent works, where relaxing the strong concavity assumption and addressing weakly convex-concave problems come with a price of direct dependence of the algorithm parameters on the desired accuracy (see, e.g., Lin et al. [28], Nouiehed et al. [35], Ostrovskii et al. [36]).

We stress the fact that we analyze both the parallel and alternating versions of PGDA through the single scheme defined in Section 2. As a consequence, our analysis of both cases is mostly unified; see Section 4. Beyond the important aspect of aesthetics, the analysis within the unified framework allows us to assert a hypothesis with respect to how these two variants of PGDA compare. Parallel and alternating updates go back to the well-known Jacobi and Gauss–Seidel (GS) schemes in computational linear algebra. The GS-based schemes are known to be in general more efficient. In minimax optimization, there is an analogous trade-off between the parallel versus alternate PGDA, and this expectation is corroborated by the results of our analysis (cf. Theorem 2) as the acceptable step size for the u-proximal gradient step of the alternating algorithm is larger than that of the parallel one.

We conclude this work with an additional expansion of the scope of problems that can be addressed by PGDA or more precisely, by a generalized version of PGDA (cf. Section 5). Using the recent framework developed in Bauschke et al. [5] and Bolte et al. [13], we partially relax the smoothness assumption with respect to $c(\cdot,\cdot)$ by replacing the requirement that the function $c(\cdot,v)$ has a Lipschitz continuous gradient, for any $v \in$ dom g, with a more generalized condition expressed by the notion of *L-smooth adaptability* of Bolte et al. [13].

Accordingly, the proximal gradient step with respect to u is replaced by an appropriate Bregman proximal gradient (BPG) step. This may also cure cases were the classical Euclidean proximal step is intractable. By carrying out few and well-targeted adjustments to our analysis, we are able to extend the complexity and convergence results to this broader setting.

Remark 1 (Comparison with a Recent Related Work by Chen et al. [16]). We end this introduction with a discussion on clarifying the significant differences between our contribution and a recent related work by Chen et al. [16] that was kindly brought to our attention by a referee. In the latter work, the authors have studied the convergence properties of the PGDA for the minimax Model (M), although only with respect to the parallel version of PGDA. First and foremost, here we present, under mild assumptions, a unified convergence analysis of both versions of the PGDA, both the parallel algorithm (parallel proximal gradient descent-ascent (PPGDA)) and the more involved alternating scheme (alternating proximal gradient descent-ascent (APGDA)), where the update of one iterate depends on the other. The paper by Chen et al. [16], on the other hand, only handles the parallel version (PPGDA). Moreover, as further explained, it also requires quite stringent assumptions that we do not make here. Indeed, with the mapping $v^*(u) := \arg\max_v(c(u,v) - g(v))$, the analysis in Chen et al. [16] relies on an assumption that the function $\|v^*(\cdot) - v\|^2$ has a nonempty subdifferential. We, on the other hand, forgo this assumption. In fact, this actually sheds light on a fundamental distinction between the approach of the two works. The convergence analysis in Chen et al. [16] is based on an application of the convergence mechanism developed in Bolte et al. [11]. However, within this approach, Chen et al. [16] are confronted with the difficulty that the term $\|v^*(\cdot) - v\|^2$ appears in their analysis, and hence, this forces them to bypass it by making the assumption that the limiting subdifferential of $\|v^*(\cdot) - v\|^2$ exists (note that we do not know if $v^*(u)$ is convex or not; hence, the classical subdifferential cannot be used here). This already appears to be quite a restrictive assumption or at least one that is not easy to check. However, there is even one more crucial assumption. Indeed, to establish their main convergence result (Chen et al. [16, theorem 2]; and likewise, for the asymptotic rate result given in Chen et al. [16, theorem 4), Chen et al. [16] further require that the function H (cf. Chen et al. [16, proposition 2]) defined by

$$H(u,v) := \phi(u) + f(u) + \alpha\|v^*(\cdot) - v\|^2, \ (\alpha > 0), \text{ with } \phi(u) := \max_v(c(u,v) - g(v)), \tag{1.2}$$

has the Kurdyka–Łojasiewicz (KL) property (Bolte et al. [10], Kurdyka [27], Lojasiewicz [30]), with the desingularizing function $\varphi(t) = c/\theta t^\theta, c > 0, \theta \in (0,1]$. This is a very restrictive assumption that we do not make in our work, and it is not clear how such a restrictive assumption can at all be verified for the function H defined. We, on the other hand, are able to forgo this stringent assumption by completely departing from the approach of Chen et al. [16]. We overcome the hurdles of the analysis by developing in Section 3 a novel general methodology for establishing convergence guarantees, which in our case, support the unified analysis of both PPGDA and APGDA. Furthermore, unlike Chen et al. [16] and as stated, as an additional beneficial consequence of our approach, we are also capable to extend both algorithms when $c(u, v)$ lacks the usual smoothness of the gradient with respect to v, which allows for appropriately altering PGDA by including non-Euclidean Bregman-based proximal steps, thus significantly enlarging the scope of the two schemes. All these results are absent in Chen et al. [16]. Finally, we note that the paper by Chen et al. [16] assumes that the feasible set of the inner maximization problem is compact, whereas we do not need this assumption in this work.

1.1. Notation

Unless otherwise specified, our notations are standard and can be found in the classical monograph of Rockafellar and Wets [41].

2. The Nonconvex Minimax Model and the Two Algorithms

We consider the following nonsmooth, nonconvex-strongly concave minimax model,

$$\min_{u\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} \{K(u,v) := f(u) + c(u,v) - g(v)\}, \tag{M}$$

under the following standing assumption.

Assumption 1 (Model (M)).

i. $\inf_{u\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} K(u,v) > -\infty$.

ii. $f : \mathbb{R}^n \to (-\infty, \infty]$ *is a proper and lsc function*.

iii. $c:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ *is a continuously differentiable* (C^1) *function and* $c(u,\cdot)$ *is a concave function for every* $u\in\mathbb{R}^n$. *In addition, there exist* $L_{vv}>0$, $L_{vu}>0$, $L_{uv}>0$, *and* $L_{uu}>0$ *such that* $\forall u,\overline{u}\in\mathbb{R}^n$, $\forall v,\overline{v}\in\mathbb{R}^m$,

$$\|\nabla_u c(u,v)-\nabla_u c(\overline{u},v)\|\le L_{uu}\|u-\overline{u}\|, \tag{2.1}$$

$$\|\nabla_u c(u,v)-\nabla_u c(u,\overline{v})\|\le L_{uv}\|v-\overline{v}\|, \tag{2.2}$$

$$\|\nabla_v c(u,v)-\nabla_v c(\overline{u},v)\|\le L_{vu}\|u-\overline{u}\|, \tag{2.3}$$

$$\|\nabla_v c(u,v)-\nabla_v c(u,\overline{v})\|\le L_{vv}\|v-\overline{v}\|. \tag{2.4}$$

iv. $g:\mathbb{R}^m\to(-\infty,\infty]$ *is a proper, lsc, and convex function.*
v. *For every* $u\in\mathbb{R}^n$, $g(\cdot)-c(u,\cdot)$ *is* σ*-strongly convex, with* $\sigma>0$.

We use the following notation to denote the ratio between the Lipschitz constant and the strong convexity parameter related to the inner maximization problem:

$$\kappa=L_{vv}/\sigma. \tag{2.5}$$

Next, we present the proximal gradient descent-ascent algorithms. Basically, PGDA is a simple single-loop unified scheme that encompasses two types of algorithms, both of which are composed of the following two proximal gradient steps (one with respect to v and the other with respect to u (see a more detailed description in Algorithm 1 (PGDA))):

$$v^{k+1}=\operatorname{prox}_{\beta g}(v^k+\beta\nabla_v c(u^k,v^k)), \tag{2.6}$$

$$u^{k+1}\in\operatorname{prox}_{\alpha f}(u^k-\alpha\nabla_u c(u^k,w^k)). \tag{2.7}$$

Setting $w^k=v^k$ leads to the PPGDA algorithm, a *Jacobi*-like method where both steps can be executed in parallel. Alternatively, by setting $w^k=v^{k+1}$, we obtain the APGDA algorithm, a *Gauss–Seidel*-like method where the u-step depends on the result of the preceding v-step.

We recall that for any proper lsc function $\varphi:\mathbb{R}^d\to(-\infty,\infty]$ and for every $t>0$ and every $u\in\mathbb{R}^n$, the Moreau proximal map associated with φ, which is assumed nonempty (see Rockafellar and Wets [41, p. 20]), is defined by

$$\operatorname{prox}_{t\varphi}(z):=\arg\min_{x\in\mathbb{R}^d}\left\{\varphi(x)+\frac{1}{2t}\|x-z\|^2\right\}. \tag{2.8}$$

Note that as the function $f(\cdot)$ in nonconvex, its proximal map is a set-valued map. Consequently, from the computational side, both algorithms will be efficient when the functions f, g are prox friendly (i.e., admit nonempty tractable proximal steps).

Algorithm 1 (PGDA) – A Unified Scheme for Parallel/Alternating Proximal Gradient Descent-Ascent (PPGDA/APGDA)
Input: $(u^0,v^0)\in\mathbb{R}^n\times\mathbb{R}^m$, $\alpha>0$, and $\beta=1/L_{vv}$.
For $k=0,1,2,\ldots$

$$v^{k+1}=\arg\max_{v\in\mathbb{R}^m}\left\{\langle\nabla_v c(u^k,v^k),v\rangle-g(v)-\frac{1}{2\beta}\|v-v^k\|^2\right\}, \tag{2.9}$$

$$w^k=\begin{cases}v^k & \text{for PPGDA,}\\ v^{k+1} & \text{for APGDA,}\end{cases} \tag{2.10}$$

$$u^{k+1}\in\arg\min_{u\in\mathbb{R}^n}\left\{f(u)+\langle\nabla_u c(u^k,w^k),u\rangle+\frac{1}{2\alpha}\|u-u^k\|^2\right\}. \tag{2.11}$$

The choice for the proximal parameter $\alpha>0$ will be made precise in Section 4.

2.1. A Composite Minimization Problem Perspective

Let $F:\mathbb{R}^n\times\mathbb{R}^m\to[-\infty,\infty)$ be defined by

$$F(u,v):=c(u,v)-g(v). \tag{2.12}$$

Then, for every $u \in \mathbb{R}^n$, the function $F(u,\cdot)$ is σ-strongly concave (i.e., $-F(u,\cdot)$ is σ-strongly convex), and so,

$$v^*(u) := \arg\max_{v\in\mathbb{R}^m} F(u,v) = \arg\min_{v\in\mathbb{R}^m}\{-F(u,v)\} \tag{2.13}$$

is a singleton (i.e., the mapping $v^* : \mathbb{R}^n \to \mathbb{R}^m$ is well defined). Thus, we can define the function $\phi : \mathbb{R}^n \to \mathbb{R}$ by

$$\phi(u) := \max_{v\in\mathbb{R}^n} F(u,v) = F(u,v^*(u)) = c(u,v^*(u)) - g(v^*(u)), \tag{2.14}$$

which allows us to state Model (M) as the following composite minimization problem:

$$\text{(P)} \qquad \min_{u\in\mathbb{R}^n}\{\Theta(u) := f(u) + \phi(u)\}.$$

We continue with an analysis of the function $\phi(\cdot)$. First, we prove that the mapping $v^*(\cdot)$ is Lipschitz continuous.

Lemma 1 (Lipschitz Continuity of v^*). *The mapping $v^\star : \mathbb{R}^n \to \mathbb{R}^m$ is (L_{vu}/σ)-Lipschitz continuous: that is,*

$$\|v^*(u) - v^*(\overline{u})\| \le \frac{L_{vu}}{\sigma}\|u - \overline{u}\|, \quad \forall u, \overline{u} \in \mathbb{R}^n. \tag{2.15}$$

Proof. For every $u \in \mathbb{R}^n$, $v^*(u) = \arg\max_{v\in\mathbb{R}^m}\{c(u,v) - g(v)\}$, and so, by the first-order optimality condition, we have that $0 \in \nabla_v c(u,v^*(u)) - \partial g(v^*(u))$: that is,

$$\nabla_v c(u,v^*(u)) \in \partial g(v^*(u)). \tag{2.16}$$

The σ-strong convexity of $g(\cdot) - c(u,\cdot)$ implies that $\partial g - \nabla_v c(u,\cdot)$ is σ-strongly monotone (see Rockafellar and Wets [41, p. 565]): that is, for every $v, \overline{v} \in \mathbb{R}^m$, $\xi \in \partial g(v)$, and $\overline{\xi} \in \partial g(\overline{v})$, we have

$$\sigma\|v - \overline{v}\|^2 \le \langle(\xi - \nabla_v c(u,v)) - (\overline{\xi} - \nabla_v c(u,\overline{v})), v - \overline{v}\rangle. \tag{2.17}$$

Thus, for every $u, \overline{u} \in \mathbb{R}^n$, by applying (2.16) for both u and $\overline{u}$ and setting $v = v^*(u)$ and $\overline{v} = v^*(\overline{u})$, we obtain that

$$\begin{aligned}\sigma\|v^*(u) - v^*(\overline{u})\|^2 &\le \langle\nabla_v c(u,v^*(\overline{u})) - \nabla_v c(\overline{u},v^*(\overline{u})), v^*(u) - v^*(\overline{u})\rangle\\ &\le \|\nabla_v c(u,v^*(\overline{u})) - \nabla_v c(\overline{u},v^*(\overline{u}))\|\|v^*(u) - v^*(\overline{u})\|\\ &\le L_{vu}\|u - \overline{u}\|\,\|v^*(u) - v^*(\overline{u})\|,\end{aligned}$$

where the second inequality is because of the Cauchy–Schwartz inequality and the third is because of the L_{vu}-Lipschitz continuity of $\nabla_v c(\cdot, v^*(\overline{u}))$ (cf. (2.3)). Finally, we divide by $\sigma\|v^*(u) - v^*(\overline{u})\|$ and obtain Inequality (2.15). □

Next, we prove a variant of Danskin's theorem (Danskin [19]; see also Bertsekas et al. [9, proposition 4.5.1]), which states that $\phi(\cdot)$ is a C^1 function with a gradient explicitly expressed in the terms of $v^*(\cdot)$. The proof follows similar arguments to those used in proving Danskin's theorem and is given here for completeness. We note that, unlike in the works by Barazandeh et al. [4], Chen et al. [16], Lin et al. [28], Lu et al. [31], Ostrovskii et al. [36], and Thekumparampil et al. [43], we do not assume that dom g is bounded.

Proposition 1 (Differentiability of ϕ). *The function $\phi : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable (C^1), and its gradient is given by*

$$\nabla\phi(u) = \nabla_u c(u,v^*(u)), \tag{2.18}$$

where $v^*(u) = \arg\min\{g(v) - c(u,v) : v \in \mathbb{R}^m\}$.

Proof. The goal is to determine the existence of the following limit:

$$\lim_{\|d\|\to 0} \frac{\phi(u+d) - \phi(u) - \langle\nabla_u c(u,v^*(u)), d\rangle}{\|d\|} = 0. \tag{2.19}$$

With $\Delta(u,d) := \phi(u+d) - \phi(u) - \langle\nabla_u c(u,v^*(u)), d\rangle$, we obtain this goal by proving that for every $u \in \mathbb{R}^n$,

$$\liminf_{\|d\|\to 0}\frac{\Delta(u,d)}{\|d\|} \ge 0 \quad \text{and} \quad \limsup_{\|d\|\to 0}\frac{\Delta(u,d)}{\|d\|} \le 0. \tag{2.20}$$

Let $u \in \mathbb{R}^n$ and $d \in \mathbb{R}^n \setminus \{0\}$. By the definition of the function ϕ (cf. (2.14)), we have

$$\begin{aligned}\Delta(u,d) &= \phi(u+d) - \phi(u) - \langle \nabla_u c(u, v^*(u)), d\rangle \\ &= F(u+d, v^*(u+d)) - F(u, v^*(u)) - \langle \nabla_u c(u, v^*(u)), d\rangle \\ &= F(u+d, v^*(u+d)) - F(u+d, v^*(u)) + F(u+d, v^*(u)) - F(u, v^*(u)) - \langle \nabla_u c(u, v^*(u)), d\rangle \\ &\geq F(u+d, v^*(u)) - F(u, v^*(u)) - \langle \nabla_u c(u, v^*(u)), d\rangle \\ &= c(u+d, v^*(u)) - c(u, v^*(u)) - \langle \nabla_u c(u, v^*(u)), d\rangle,\end{aligned}$$

where the inequality is because of the definition of v^* (2.13), which implies that $F(u+d, v^*(u+d)) \geq F(u+d, v^*(u))$. It follows that for all $u \in \mathbb{R}^n$,

$$\liminf_{\|d\|\to 0} \frac{\Delta(u,d)}{\|d\|} \geq \lim_{\|d\|\to 0} \frac{c(u+d, v^*(u)) - c(u, v^*(u)) - \langle \nabla_u c(u, v^*(u)), d\rangle}{\|d\|} = 0,$$

where the last limit result is because of the definition of the gradient of the differentiable function $c(\cdot, v^*(u))$.

Next, for every $u \in \mathbb{R}^n$ and $d \in \mathbb{R}^n \setminus \{0\}$, we have

$$\begin{aligned}\Delta(u,d) &= F(u+d, v^*(u+d)) - F(u, v^*(u)) - \langle \nabla_u c(u, v^*(u)), d\rangle \\ &= F(u+d, v^*(u+d)) - F(u, v^*(u+d)) + F(u, v^*(u+d)) - F(u, v^*(u)) - \langle \nabla_u c(u, v^*(u)), d\rangle \\ &\leq F(u+d, v^*(u+d)) - F(u, v^*(u+d)) - \langle \nabla_u c(u, v^*(u)), d\rangle \\ &= c(u+d, v^*(u+d)) - c(u, v^*(u+d)) - \langle \nabla_u c(u, v^*(u)), d\rangle,\end{aligned}$$

where the inequality is because of the definition of v^* (2.13), which implies that $F(u, v^*(u+d)) \leq F(u, v^*(u))$.

Noting that the function $c(\cdot, v^*(u+d))$ is differentiable, we apply the *mean value theorem* and obtain that there exists $\alpha \in [0,1]$ such that

$$c(u+d, v^*(u+d)) - c(u, v^*(u+d)) = \langle \nabla_u c(u+\alpha d, v^*(u+d)), d\rangle,$$

and as a result, we have

$$\begin{aligned}\Delta(u,d) &\leq \langle \nabla_u c(u+\alpha d, v^*(u+d)) - \nabla_u c(u, v^*(u)), d\rangle \\ &\leq \|\nabla_u c(u+\alpha d, v^*(u+d)) - \nabla_u c(u, v^*(u))\| \, \|d\|,\end{aligned}$$

where the second inequality is because of the Cauchy–Schwartz inequality. It follows that

$$\limsup_{\|d\|\to 0} \frac{\Delta(u,d)}{\|d\|} \leq \lim_{\|d\|\to 0} \|\nabla_u c(u+\alpha d, v^*(u+d)) - \nabla_u c(u, v^*(u))\| = 0,$$

where the last limit result is because of the fact that both $\nabla_u c(\cdot,\cdot)$ and v^* are continuous (cf. Lemma 1). □

Finally, we prove that the gradient $\nabla\phi$ is Lipschitz continuous.

Lemma 2 (Lipschitz Continuity of $\nabla\phi$). *The function $\phi : \mathbb{R}^n \to \mathbb{R}$ has an L_ϕ-Lipschitz continuous gradient, with*

$$L_\phi = L_{uu} + \frac{L_{uv} \cdot L_{vu}}{\sigma} > 0. \tag{2.21}$$

Proof. Note that for every $u, \overline{u} \in \mathbb{R}^n$, we have

$$\begin{aligned}\|\nabla\phi(u) - \nabla\phi(\overline{u})\| &= \|\nabla_u c(u, v^*(u)) - \nabla_u c(\overline{u}, v^*(\overline{u}))\| \\ &= \|\nabla_u c(u, v^*(u)) - \nabla_u c(\overline{u}, v^*(u)) + \nabla_u c(\overline{u}, v^*(u) - \nabla_u c(\overline{u}, v^*(\overline{u}))\| \\ &\leq \|\nabla_u c(u, v^*(u)) - \nabla_u c(\overline{u}, v^*(u))\| + \|\nabla_u c(\overline{u}, v^*(u) - \nabla_u c(\overline{u}, v^*(\overline{u}))\| \\ &\leq L_{uu}\|u - \overline{u}\| + L_{uv}\|v^*(u) - v^*(\overline{u})\| \\ &\leq \left(L_{uu} + \frac{L_{uv} \cdot L_{vu}}{\sigma}\right)\|u - \overline{u}\|,\end{aligned}$$

where the first inequality is because of the triangle inequality, the second is because of the Lipschitz continuity of $\nabla_u c(\cdot, v^*(u))$ and $\nabla_u c(\overline{u}, \cdot)$ (cf. (2.1) and (2.2), respectively), and the third inequality is because of Lemma 1. □

In light of the results, it is natural to think of applying the proximal gradient algorithm (PGA) to Model (P): that is, repeatedly set

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^n}\left\{ f(u) + \langle \nabla\phi(u^k), u\rangle + \frac{1}{\alpha}\|u-u^k\|^2 \right\}$$
$$= \arg\min_{u\in\mathbb{R}^n}\left\{ f(u) + \langle \nabla_u c(u^k, v^*(u^k)) + \frac{1}{\alpha}\|u-u^k\|^2 \right\}. \tag{2.22}$$

The difficulty in this approach is that finding $v^*(u^k)$ (i.e., solving the strongly convex minimization Problem (2.13) with $u = u^k$) requires for each iteration k to use another inner-loop algorithm, thus leading to an overall more computationally demanding and complicated nested algorithm. In this context, PPGDA and APGDA can be viewed as approximating PGA, where instead of solving (2.13), we perform *one* proximal gradient step (step (2.9)) and replace $v^*(u^k)$ with either v^k or v^{k+1}.

3. A General Mechanism for Convergence Analysis

The minimax structure of the proposed algorithm, which computes a proximal *gradient ascent* step followed by (or simultaneously with) a proximal *gradient descent*, implies oscillation of the produced sequences and hence, the lack of the crucially required sufficient descent property for the objective function $\Theta(\cdot)$. Indeed, as proven in Section 4 in Lemma 7, the right-hand side of Inequality (4.7) contains an additional positive term that precludes our ability to obtain the required descent property. Thus, unfortunately, we cannot directly apply the convergence mechanism developed in Bolte et al. [11] to prove global convergence to a critical point. To tackle this inherent difficulty, we adopt and extend the approach and proof techniques of Bolte et al. [11] (here, we follow the more recent work by Bolte et al. [13]) with a set of novel conditions that still guarantee global convergence. Because the forthcoming results can be useful in contexts other than minimax algorithms, we present here a general mechanism that will then be applied in Section 4 for proving global convergence of the respective algorithms PPGDA and APGDA in a unified manner. Before doing so, given that the model (P) is nonconvex and nonsmooth, we first recall some basic variational analysis concepts; see Rockafellar and Wets [41] for more details.

Definition 1 (Subdifferentials). Let $\psi : \mathbb{R}^d \to (-\infty, \infty]$ 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 $v \in \mathbb{R}^d$ satisfying

$$\psi(x) \geq \psi(z) + \langle v, x-z\rangle + o(\|x-z\|), \tag{3.1}$$

and for $z \notin \operatorname{dom}\psi$, we set $\hat{\partial}\psi(z) = \partial\psi(z) = \emptyset$.

ii. The *(limiting) subdifferential* of ψ at z, denoted by $\partial\psi(z)$, is the set of all vectors $v \in \mathbb{R}^d$ 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 $v^k \to v$.

The set of critical points of ψ is defined by $\operatorname{crit}\psi := \{x : 0 \in \partial\psi(x)\}$.

When ψ is convex, the limiting subdifferential reduces to the classical subdifferential of the convex function ψ. Futhermore, note that in the context of our Model (P), namely with $\phi \in C^1$, we also have the following useful subdifferential rule (cf. Rockafellar and Wets [41, p. 304]):

$$\partial(\psi+\phi)(x) = \partial\psi(x) + \nabla\phi(x), \quad \forall x \in \mathbb{R}^d, \tag{3.2}$$

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

Remark 2 (Closedness of the Graph of $\partial\psi$). We note, as in Bolte et al. [11, remark 1(ii)], that given a convergent sequence $(x^k, w^k) \underset{k\to\infty}{\to} (x, w)$, such that $w^k \in \partial\psi(x^k)$ and $\lim_{k\to\infty}\psi(x^k) = \psi(x)$, it holds that $w \in \partial\psi(x)$.

Following Bolte et al. [11] and Bolte et al. [13], we are now ready to extend the definition of gradient-like descent sequences to include a *perturbation sequence* in the following way. The convergence framework begins by defining a set of conditions that will guarantee that every cluster point x^* of a generated sequence $\{(x^k\}_{k\geq 0}$ is a critical point of a proper and lsc function Ψ.

Definition 2 (Perturbed Gradient-Like Descent Sequence). Let $\Psi : \mathbb{R}^d \to (-\infty, \infty]$ be a proper and lsc function. Then, a sequence $\{(x^k, v_k)\}_{k\geq 0} \subseteq \operatorname{dom}\Psi \times \mathbb{R}_+$ is called *a perturbed gradient-like descent sequence* with respect to Ψ if it satisfies the following three conditions.

Condition 1 (Perturbed Sufficient Decrease Property). *There exists $c_1 > 0$ such that for every $k \in \mathbb{N}$,*

$$c_1(\|x^{k+1} - x^k\|^2 + v_k^2) \leq \left(\Psi(x^k) + \frac{1}{2}v_k^2\right) - \left(\Psi(x^{k+1}) + \frac{1}{2}v_{k+1}^2\right). \tag{3.3}$$

Condition 2 (Perturbed Subgradient Lower Bound on Iterates Gap). *There exists $c_2 > 0$ such that for every $k \geq 0$, there exists $\xi^{k+1} \in \partial\Psi(x^{k+1})$, which satisfies*

$$\|\xi^{k+1}\| \leq c_2(\|x^{k+1} - x^k\| + v_k). \tag{3.4}$$

Condition 3. *Let $\{x^k\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$ be a subsequence that converges to a point $\overline{x}$. Then,*

$$\limsup_{k\in\mathcal{K}\subseteq\mathbb{N}} \Psi(x^k) \leq \Psi(\overline{x}).$$

Clearly, with $v_k \equiv 0$ for all k (i.e., there is no perturbation), then the novel Conditions 1 and 2 reduce to the classical conditions introduced in Bolte et al. [11] and Bolte et al. [13], namely where Condition 1 appears with the usual sufficient decrease property for the function ψ measured in terms of the squared norm of the iterates gap and Condition 2 appears with a subgradient lower bound for the norm of the iterates gap, whereas the continuity-like Condition 3 remains untouched. The role of the sequence v_k introduced in the novel set of conditions will be crucial in the forthcoming analysis, in particular allowing us to conduct the convergence analysis of PPGDA and APGDA within a unified framework; see Section 4.

The next lemma plays a key role in establishing both convergence and iteration complexity.

Lemma 3 (Function Values and Subgradients Convergence). *Let $\Psi : \mathbb{R}^d \to (-\infty, \infty]$ be a proper and lsc function, consider the sequence $\{(x^k, v_k)\}_{k\in\mathbb{N}} \subseteq \mathbb{R}^d \times \mathbb{R}_+$, and assume that $\underline{\Psi} := \inf_{k\in\mathbb{N}}\Psi(x^k) > -\infty$ (e.g., when $\{x^k\}_{k\in\mathbb{N}}$ is bounded or when $\inf_{x\in\mathbb{R}^d}\Psi(x) > -\infty$).*

i. *Suppose that $\{(x^k, v_k)\}_{k\in\mathbb{N}}$ satisfies Condition 1 with respect to Ψ. Then,*

$$\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0, \quad \lim_{k\to\infty} v_k = 0, \tag{3.5}$$

and the sequence $\{\Psi(x^k)\}_{k\in\mathbb{N}}$ is convergent (i.e., the limit $\overline{\Psi} := \lim_{k\in\mathbb{N}} \Psi(x^k) \in \mathbb{R}$ exists).

ii. *Set $K_1 = (\Psi(x^0) + v_0^2/2 - \underline{\Psi})/c_1$, with c_1 as in Condition 1. Then, $K_1 < \infty$, and for every $N \in \mathbb{N}$,*

$$\min_{1\leq k\leq N} (\|x^{k+1} - x^k\|^2 + v_k^2) \leq \frac{K_1}{N}. \tag{3.6}$$

iii. *Suppose in addition that Condition 2 is satisfied. Then, for every $k \in \mathbb{N}$, there exists $\xi^k \in \partial\Psi(x^k)$ such that*

$$\lim_{k\to\infty} \xi^k = 0, \tag{3.7}$$

and for every $N \in \mathbb{N}$,

$$\min_{1\leq k\leq N} \operatorname{dist}(0, \partial\Psi(x^k)) \leq \sqrt{\frac{K_2}{N}}, \tag{3.8}$$

where $K_2 := 2c_2^2K_1 \equiv 2c_2^2(\Psi(x^0) + v_0^2/2 - \underline{\Psi})/c_1$ with c_1 and c_2 as in Conditions 1 and 2. Furthermore, the minimizing index k is the same in (3.6) and in (3.8).

Proof. Summing the inequality of Condition 1 over $N \in \mathbb{N}$, we obtain that

$$\begin{aligned}\sum_{k=0}^{N-1}(\|x^{k+1} - x^k\|^2 + v_k^2) &\leq \frac{1}{c_1}\sum_{k=0}^{N-1}\left(\Psi(x^k) + \frac{1}{2}v_k^2 - \Psi(x^{k+1}) - \frac{1}{2}v_{k+1}^2)\right) \\ &= \frac{1}{c_1}\left(\Psi(x^0) + \frac{1}{2}v_0^2 - \Psi(x^N) - \frac{1}{2}v_N^2\right) \\ &\leq \frac{1}{c_1}\left(\Psi(x^0) + \frac{1}{2}v_0^2 - \underline{\Psi}\right) < \infty. \end{aligned} \tag{3.9}$$

Thus, the nonnegative series $\sum_{k=0}^{\infty}(\|x^{k+1}-x^k\|^2+v_k^2)$ is convergent, and it follows that

$$\lim_{k\to\infty}\|x^{k+1}-x^k\|=0 \quad \text{and} \quad \lim_{k\to\infty} v_k=0.$$

Furthermore, note that (C1) together with the lower boundness of $\{\Psi(x^k)\}_{k\in\mathbb{N}}$ implies that the sequence $\{\Psi(x^k)+v_k/2\}_{k\in\mathbb{N}}$ is nonincreasing and bounded from below. It follows that it is a convergent sequence with a limit $\overline{\Psi}\in\mathbb{R}$, and as $\lim_{k\to\infty}v_k=0$, then

$$\lim_{k\to\infty}\Psi(x^k)=\overline{\Psi}.$$

In addition, as

$$N\cdot\min_{1\le k\le N}(\|x^{k+1}-x^k\|^2+v_k^2)\le\sum_{k=0}^{N-1}(\|x^{k+1}-x^k\|^2+v_k^2),$$

we obtain Inequality (3.6) by dividing (3.9) by N.

Finally, we assume Condition 2, namely that Inequality (3.4) holds. Then, together with (3.5), we obtain that $\lim_{k\to\infty}\xi^k=0$. Moreover, as

$$(\|x^{k+1}-x^k\|+v_k)^2\le 2(\|x^{k+1}-x^k\|^2+v_k^2),$$

then combining (3.6) and (3.4) proves the bound (3.8). □

We are now ready to establish subsequence convergence.

Lemma 4 (Subsequence Convergence). *Let $\{x^k,v_k\}_{k\ge0}$ be a gradient-like descent sequence with regard to the proper and lsc function Ψ. In addition, assume that $\{x^k\}_{k\in\mathbb{N}}$ is bounded. Denote ω as the set of cluster points of $\{x^k\}_{k\in\mathbb{N}}$. Then, ω is a nonempty, compact, and connected set; $\omega\subset\operatorname{crit}\Psi$; $\lim_{k\to\infty}\operatorname{dist}(x^k,\omega)=0$; and Ψ is finite and constant on ω.*

Proof. The proof is basically the same as that of Bolte et al. [13, lemma 6.1]. As $\{x^k\}_{k\in\mathbb{N}}$ is bounded, then there exist a cluster point $\overline{x}\in\mathbb{R}^d$ and a subsequence $\{x^k\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$ that converges to $\overline{x}$, and hence, ω is nonempty. Furthermore, ω is compact as it can be viewed as an intersection of compact sets. Next, by Condition 3 together with Ψ being lsc, we have that for any such cluster point and corresponding subsequence,

$$\Psi(\overline{x})=\lim_{\substack{k\to\infty\\ k\in\mathbb{K}\subseteq\mathbb{N}}}\Psi(x^k)=\lim_{k\to\infty}\Psi(x^k):=\overline{\Psi}\in\mathbb{R}, \tag{3.10}$$

where the existence of the second limit is because of Lemma 3. Finally, Lemma 3 together with the closedness property of $\partial\Psi$ (cf. Remark 2) implies that $0\in\partial\Psi(\overline{x})$. □

In order to obtain a global convergence result, we first need the so-called KL property (Bolte et al. [10], Kurdyka [27], Lojasiewicz [30]), which appears as a central property for proving global convergence results in the nonconvex setting (see, e.g., Bolte et al. [11], Bolte et al. [12], and references therein).

We recall here the basic definition of the nonsmooth KL property and in Lemma 5, a key result regarding the uniformized KL property (see Bolte et al. [11, lemma 6, p. 478]), which will be useful for the proof of Theorem 1. Before doing so, we define the following class of functions:

$$\Phi_\eta:=\{\varphi\in C^0[0,\eta)\cap C^1(0,\eta):\varphi(0)=0,\varphi \text{ is concave, and } \varphi'(t)>0,\quad \forall t\in(0,\eta))\}.$$

Definition 3 (The Nonsmooth KL Property). A proper and lsc function $h:\mathbb{R}^d\to(-\infty,\infty]$ has the KL property locally at $\overline{x}\in\operatorname{dom}h$ if there exist $\eta>0$, $\varphi\in\Phi_\eta$, and a neighborhood $\mathcal{N}(\overline{x})$ such that

$$\varphi'(h(x)-h(\overline{x}))\operatorname{dist}(0,\partial h(x))\ge 1,$$

for all $x\in\mathcal{N}(\overline{x})\cap\{x\in\mathbb{R}^d:h(\overline{x})<h(x)<h(\overline{x})+\eta\}$.

Lemma 5 (The Uniformized KL Property). *Let $\Omega\subseteq\mathbb{R}^d$ be a compact set, and let $h:\mathbb{R}^d\to(-\infty,\infty]$ be a proper and lsc function. Assume that h 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 $\overline{x}\in\Omega$, we have*

$$\varphi'(h(x)-h(\overline{x}))\operatorname{dist}(0,\partial h(x))\ge 1, \tag{3.11}$$

for all $x\in\{x\in\mathbb{R}^d:\operatorname{dist}(x,\Omega)<\varepsilon\}\cap\{x\in\mathbb{R}^d:h(\overline{x})<h(x)<h(\overline{x})+\eta\}$.

Recall that the important class of semialgebraic functions is known to satisfy the nonsmooth Kurdyka–Łojasiewicz property (see, e.g., Bolte et al. [11] and references therein).

In the sequel, we also use the following simple technical result whose proof is immediate.

Proposition 2. *For every $a, b \in \mathbb{R}$ and $\gamma > 0$, we have*
i. $a \cdot b \le (\gamma/2)a^2 + (1/2\gamma)b^2$ *and*
ii. $(a+b)^2 \le (1+\gamma)a^2 + (1+\gamma^{-1})b^2 = (1+\gamma)(a^2 + b^2/\gamma)$.

We can now state and prove the desired global convergence result, which extends Bolte et al. [11, theorem 6.2].

Theorem 1 (Global Convergence). *Let $\{x^k, \nu_k\}_{k\ge 0} \subseteq \mathbb{R}^d \times \mathbb{R}_+$ be a perturbed gradient-like descent sequence with respect to the proper and lsc function Ψ, and assume that $\{x^k\}_{k\ge 0}$ is bounded. In addition, assume that ψ is a semialgebraic function. Then, $\{x^k\}_{k\in\mathbb{N}}$ has a finite length (i.e., $\sum_{k=1}^{\infty} \|x^{k+1} - x^k\| < \infty$), and it converges to a critical point $x^* \in \operatorname{crit}\Psi$. Furthermore, we have $\lim_{k\to\infty} \nu_k = 0$.*

Proof. For convenience, we define the function $\mathcal{L}_\Psi : \mathbb{R}^d \times \mathbb{R} \to (-\infty, \infty])$ by

$$\mathcal{L}_\Psi(x, \nu) := \Psi(x) + \frac{1}{2}\nu^2.$$

Then, $\mathcal{L}_\Psi$ is a semialgebraic function, and in particular, it satisfies the KL property at each point of $\operatorname{dom}\Psi \times \mathbb{R}$. For every $k \in \mathbb{N}$, set $z^k = (x^k - x^{k-1}, \nu_{k-1})$. As $\|z^{k+1}\|^2 = \|x^{k+1} - x^k\|^2 + \nu_k^2$, we can write Condition 1 as

$$\mathcal{L}_\Psi(x^k, \nu_k) - \mathcal{L}_\Psi(x^{k+1}, \nu_{k+1}) \ge c_1 \|z^{k+1}\|^2. \tag{3.12}$$

Next, we note that $\partial\mathcal{L}_\Psi(x, \nu) = \partial\Psi(x) \times \{\nu\}$ (see Rockafellar and Wets [41, proposition 10.5]), and so, there exists $\zeta^{k+1} \in \partial\mathcal{L}_\Psi(x^{k+1}, \nu_k)$ that satisfies the inequality of Condition 2, where $\zeta^{k+1} = (\xi^{k+1}, \nu_k)$, with $\xi^{k+1} \in \partial\Psi(x^{k+1})$. Thus, recalling that $\nu_k \ge 0$ and together with the fact that

$$\|\zeta^{k+1}\| = \sqrt{\|\xi^{k+1}\|^2 + \nu_k^2} \le \|\xi^{k+1}\| + \nu_k,$$

we obtain

$$\|\zeta^{k+1}\| \le c_2(\|x^{k+1} - x^k\| + \nu_k) + \nu_k \le (c_2 + 1)(\|x^{k+1} - x^k\| + \nu_k).$$

Furthermore, as

$$\|x^{k+1} - x^k\| + \nu_k \le \sqrt{2\|x^{k+1} - x^k\|^2 + 2\nu_k^2} = \sqrt{2}\|z^{k+1}\|$$

(cf. Proposition 2(ii)), then

$$\|\zeta^{k+1}\| \le \sqrt{2}(1 + c_2)\|z^{k+1}\|. \tag{3.13}$$

The proof continues similarly to that of Bolte et al. [13, theorem 6.2]. As $\{x^k\}_{k\in\mathbb{N}}$ is bounded, there exists a subsequence $\{x^k\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$ that converges to $\overline{x} \in \mathbb{R}^d$, and as proved in Lemma 4 (cf. (3.10)), we have

$$\lim_{k\to\infty} \Psi(x^k) = \Psi(\overline{x}).$$

In addition, $\lim_{k\to\infty} \nu_k = 0$, as stated by Lemma 3, and it follows that

$$\underline{\mathcal{L}}_\Psi := \lim_{k\to\infty} \mathcal{L}_\Psi(x^k, \nu_k) = \mathcal{L}_\Psi(\overline{x}, 0) = \Psi(\overline{x}).$$

Consider the case where there exists an integer k for which $\mathcal{L}_\Psi(x^k, \nu_k) = \underline{\mathcal{L}}_\Psi$. The perturbed sufficient decrease Condition 1 (cf. (3.3)) implies that the sequence $\{\mathcal{L}_\Psi(x^k, \nu_k)\}_{k\in\mathbb{N}}$ is a decreasing sequence; so,

$$\underline{\mathcal{L}}_\Psi \le \mathcal{L}_\Psi(x^{k+1}, \nu_{k+1}) \le \mathcal{L}_\Psi(x^k, \nu_k) = \underline{\mathcal{L}}_\Psi,$$

and that is,

$$\Psi(x^{k+1}) + \frac{1}{2}\nu_{k+1} = \Psi(x^k) + \frac{1}{2}\nu_k.$$

We immediately obtain by (3.3) that $x^{k+1} = x^k$, $\nu_k = 0$, and it follows that $\Psi(x^{k+1}) = \Psi(x^k)$ and $\nu_{k+1} = \nu_k = 0$. A trivial induction shows that the sequence $\{x^k, \nu_k\}_{k\in\mathbb{N}}$ is stationary.

Otherwise, let ω be the set of cluster points of $\{x^k\}_{k\in\mathbb{N}}$. Then, by Lemma 4, ω is a nonempty, compact set; $\omega \subseteq \operatorname{crit}\Psi$; and Ψ is constant on ω. Also note that for every $\delta > 0$, there exists $k_\delta \in \mathbb{N}$ such that $\operatorname{dist}(x^k, \omega) < \delta$, for

every $k \geq k_\delta$. Then, $\mathcal{L}_\Psi$ is constant on $\omega \times \{0\}$, and by applying Lemma 5, we obtain that there exist $\delta > 0$, $\eta > 0$, and $\varphi \in \Phi_\eta$ such that for all $\overline{x} \in \omega$ and for every $k \geq k_\delta$,

$$\varphi'(\mathcal{L}_\Psi(x^k, v_k) - \mathcal{L}_\Psi(\overline{x}, 0))\text{dist}(0, \partial\mathcal{L}_\Psi(x^k, v_k)) \geq 1. \tag{3.14}$$

Let $\zeta^k \in \partial\mathcal{L}_\Psi(x^k, v_k)$, which satisfies Inequality (3.13). Then, as $\|\zeta^k\| \geq \text{dist}(0, \partial\mathcal{L}_\Psi(x^k, v_k))$, it follows that

$$\varphi'(\mathcal{L}_\Psi(x^k, v_k) - \mathcal{L}_\Psi(\overline{x}, \overline{y})) \geq \frac{1}{\text{dist}(0, \partial\mathcal{L}_\Psi(x^k, v_k)} \geq \frac{1}{\|\zeta^k\|} \geq \frac{1}{\sqrt{2}(1 + c_2)\|z^k\|}, \tag{3.15}$$

where the last inequality is because of (3.13). Next, define

$$\Delta_{p,q} := \varphi(\mathcal{L}_\Psi(x^p, v_p) - \mathcal{L}_\Psi(\overline{x}, 0)) - \varphi(\mathcal{L}_\Psi(x^q, v_q) - \mathcal{L}_\Psi(\overline{x}, 0)).$$

From the concavity of φ, we have

$$\Delta_{k,k+1} \geq \varphi'(\mathcal{L}_\Psi(x^k, v_k) - \mathcal{L}_\Psi(\overline{x}, 0))(\mathcal{L}_\Psi(x^k, v_k) - \mathcal{L}_\Psi(x^{k+1}, v_{k+1})),$$

and so, together with (3.12) and (3.15), we have

$$\Delta_{k,k+1} \geq \frac{\|z^{k+1}\|^2}{c\|z^k\|}, \tag{3.16}$$

with $c = c_1/(\sqrt{2}(1 + c_2))$. Multiplying by c and adding $\|z^k\|$, we obtain that

$$\|z^k\| + c\Delta_{k,k+1} \geq \frac{\|z^{k+1}\|^2 + \|z^k\|^2}{\|z^k\|} \geq 2\frac{\|z^{k+1}\|\,\|z^k\|}{\|z^k\|} = 2\|z^{k+1}\|, \tag{3.17}$$

where the last inequality is because of Proposition 2(i).

Next, given $N \in \mathbb{N}$, we sum (3.17) for $k = 1, \ldots, N-1$ and obtain that

$$\begin{aligned} 2\sum_{k=1}^{N}\|z^{k+1}\| &\leq \sum_{k=1}^{N-1}\|z^k\| + c\sum_{k=1}^{N}\Delta_{k,k+1} \\ &\leq \|z^1\| + \sum_{k=1}^{N}\|z^{k+1}\| + c\sum_{k=+1}^{N}\Delta_{k,k+1} \\ &= \|z^1\| + \sum_{k=1}^{N}\|z^{k+1}\| + c\Delta_{1,N+1}, \end{aligned}$$

where the last equality is because of the fact the $\Delta_{p,q} + \Delta_{q,r} = \Delta_{p,r}$. It follows that

$$\sum_{k=1}^{N}\|x^{k+1} - x^k\| \leq \sum_{k=1}^{N}\|z^{k+1}\| \leq \|z^1\| + c\Delta_{1,N+1} \leq \|z^1\| + \varphi(\mathcal{L}_\psi(x^1, v_1) - \mathcal{L}_\psi(\overline{x}, 0)),$$

where the last inequality is because of the nonnegativity of φ. As a result, we obtain that

$$\sum_{k=1}^{\infty}\|x^{k+1} - x^k\| < \infty,$$

and that is, $\{x^k\}$ is a Cauchy sequence, hence a convergent sequence, which together with Lemma 4, completes the proof that x^k converges to a critical point of Ψ. □

4. Convergence Analysis of PPGDA and APGDA

We analyze the convergence of PPGDA and APGDA in light of the formulation of Model (M) as the composite minimization problem

$$\min_{u \in \mathbb{R}^n}\{\Theta(u) := f(u) + \phi(u)\}. \tag{P}$$

Specifically, let $\{u^k, v^k\}_{k\in\mathbb{N}}$ be the sequence generated by the algorithm. Then, we prove that $\{u^k\}_{k\in\mathbb{N}}$ converges to a critical point of Θ: that is, to a point

$$\overline{u} \in \text{crit}\ \Theta := \{u : 0 \in \partial\Theta(u)\}.$$

In addition, we prove that $\{v^k\}_{k\in\mathbb{N}}$ converges to $v^*(\overline{u})$, where we recall that given $u \in \mathbb{R}^n$,

$$v^*(u) = \arg\max_{v\in\mathbb{R}^m}\{c(u,v) - g(v)\} = \arg\min_{v\in\mathbb{R}^m}\{g(v) - c(u,v)\}$$

is the solution of the inner maximization problem of Model (M).

In the following analysis, we use the notations defined in (2.5), and (2.21):

$$\kappa = L_{vv}/\sigma \quad \text{and} \quad L_\phi = L_{uu} + \frac{L_{uv}\cdot L_{vu}}{\sigma}.$$

4.1. Three Key Estimates for Algorithm 1 (PGDA)

Our main goal is to establish some key inequalities estimates for the sequences $\{u^k, v^k, w^k\}$ generated by the unified algorithm PGDA, whereby we recall that the auxiliary sequence $\{w^k\}$ is used to distinguish between the two proposed schemes PPGDA and APGDA through the following choice (cf. (2.10)):

$$w^k = \begin{cases} v^k & \text{for PPGDA,} \\ v^{k+1} & \text{for APGDA.} \end{cases}$$

The forthcoming three key lemmas that hold for the unified Algorithm 1 (PGDA) will serve as our main tool to analyze the PPGDA and APGDA algorithms.

We begin with analyzing the v-step (cf. (2.9)) and establish the following relation between v^{k+1}, v^k, and the corresponding maximizers $v^*(u^k)$ and $v^*(u^{k+1})$.

Lemma 6 (Relation Between v^{k+1} and v^k). *Let $\kappa = L_{vv}/\sigma$. Then, for every $k \in \mathbb{N}$, we have*

$$\|v^{k+1} - v^*(u^{k+1})\| \le \sqrt{\kappa/(\kappa+1)}\|v^k - v^*(u^k)\| + \frac{L_{vu}}{\sigma}\|u^{k+1} - u^k\|, \tag{4.1}$$

$$\|v^{k+1} - v^*(u^k)\| \le \sqrt{\kappa/(\kappa+1)}\left(\|v^k - v^*(u^{k-1})\| + \frac{L_{vu}}{\sigma}\|u^k - u^{k-1}\|\right), \tag{4.2}$$

$$\|v^{k+1} - v^*(u^{k+1})\|^2 \le \frac{\kappa+1/2}{\kappa+1}\|v^k - v^*(u^k)\|^2 + \frac{(2\kappa+1)L_{vu}^2}{\sigma^2}\|u^{k+1} - u^k\|^2, \tag{4.3}$$

$$\|v^{k+1} - v^*(u^k)\|^2 \le \frac{\kappa+1/2}{\kappa+1}\left(\|v^k - v^*(u^{k-1})\|^2 + \frac{2\kappa L_{vu}^2}{\sigma^2}\|u^k - u^{k-1}\|^2\right). \tag{4.4}$$

Proof. Fix $k \in \mathbb{N}$. The v-step, as defined in (2.9), reads

$$v^{k+1} = \operatorname{prox}_{\frac{1}{L_{vv}}g}\left(v^k + \frac{1}{L_{vv}}\nabla_v c(u^k, v^k)\right). \tag{4.5}$$

It is a proximal gradient step that addresses the model's inner maximization, which can be stated as the following minimization problem:

$$\min_{v\in\mathbb{R}^m}\{\Gamma_k(v) := g(v) - c(u^k, v)\},$$

where we recall that $g(\cdot)$ is proper, lsc, and convex; $-c(u^k,\cdot)$ is a convex C^1 function with an L_{vv}-Lipschitz continuous gradient (cf. (2.4)); $\Gamma_k(\cdot)$ is σ-strongly convex, with $\sigma > 0$; and its unique minimizer is given by

$$v^*(u^k) := \arg\min_{v\in\mathbb{R}^m}\Gamma_k(v).$$

Therefore, $0 \in \partial\Gamma_k(v^*(u^k))$, and by the subgradient inequality for the σ-strongly convex function Γ_k, we obtain

$$\Gamma_k(v^{k+1}) - \Gamma_k(v^*(u^k)) \ge \langle 0, v^{k+1} - v^*(u^k)\rangle + \frac{\sigma}{2}\|v^{k+1} - v^*(u^k)\|^2 = \frac{\sigma}{2}\|v^{k+1} - v^*(u^k)\|^2.$$

Applying the well-known proximal gradient inequality (cf. Beck and Teboulle [6, lemma 2.6, p. 56] or Teboulle [42, lemma 4.1]) for (4.5), we also have

$$\Gamma_k(v^{k+1}) - \Gamma_k(v^*(u^k)) \le \frac{L_{vv}}{2}\left(\|v^k - v^*(u^k)\|^2 - \|v^{k+1} - v^*(u^k)\|^2\right).$$

Combining both inequalities (after taking the square root of both sides of the resulting inequality), we obtain that

$$\|v^{k+1} - v^*(u^k)\| \leq \sqrt{L_{vv}/(L_{vv}+\sigma)}\|v^k - v^*(u^k)\| = \sqrt{\kappa/(\kappa+1)}\|v^k - v^*(u^k)\|. \tag{4.6}$$

Together with the triangle inequality and the Lipschitz continuity of $v^*(\cdot)$, as stated by Lemma 1, we prove both Inequalities (4.1) and (4.2) as follows:

$$\begin{aligned}
\|v^{k+1} - v^*(u^{k+1})\| &= \|v^{k+1} - v^*(u^k) + v^*(u^k) - v^*(u^{k+1})\| \\
&\leq \|v^{k+1} - v^*(u^k)\| + \|v^*(u^k) - v^*(u^{k+1})\| \\
&\leq \sqrt{\kappa/(\kappa+1)}\|v^k - v^*(u^k)\| + \frac{L_{vu}}{\sigma}\|u^{k+1} - u^k\|, \\
\|v^{k+1} - v^*(u^k)\| &\leq \sqrt{\kappa/(\kappa+1)}\|v^k - v^*(u^k)\| \\
&= \sqrt{\kappa/(\kappa+1)}\|v^k - v^*(u^{k-1}) + v^*(u^{k-1}) - v^*(u^k)\| \\
&\leq \sqrt{\kappa/(\kappa+1)}(\|v^k - v^*(u^{k-1})\| + \|v^*(u^k) - v^*(u^{k-1})\|) \\
&\leq \sqrt{\kappa/(\kappa+1)}\left(\|v^k - v^*(u^{k-1})\| + \frac{L_{vu}}{\sigma}\|u^k - u^{k-1}\|\right).
\end{aligned}$$

Finally, to obtain Inequalities (4.3) and (4.4), we take the square of both sides of (4.1) and (4.2), respectively, and apply Proposition 2(ii) with $\gamma \equiv 1/2\kappa$, noting that

$$\frac{\kappa(1+1/2\kappa)}{\kappa+1} = \frac{\kappa+1/2}{\kappa+1}. \quad \square$$

Next, we turn to the u-step (cf. (2.11)). The first lemma considers the function values gap of consecutive iterates.

Lemma 7 (Function Values Gap for the u-Step). *Let $\{u^k, w^k\}_{k\in\mathbb{N}}$ be the sequence generated by PGDA. Then, for every $k \geq 0$, we have*

$$\Theta(u^{k+1}) - \Theta(u^k) \leq -\frac{1}{2}\left(\frac{1}{\alpha} - L_{uv}^2 - L_\phi\right)\|u^{k+1} - u^k\|^2 + \frac{1}{2}\|w^k - v^*(u^k)\|^2. \tag{4.7}$$

Proof. The u-step can be stated as

$$u^{k+1} \in \arg\min_{u\in\mathbb{R}^n}\left\{f(u) + \langle \nabla_u c(u^k, w^k), u - u^k\rangle + \frac{1}{2\alpha}\|u - u^k\|^2\right\},$$

and so, we have that

$$f(u^{k+1}) - f(u^k) \leq -\frac{1}{2\alpha}\|u^{k+1} - u^k\|^2 - \langle \nabla_u c(u^k, w^k), u^{k+1} - u^k\rangle. \tag{4.8}$$

Having $\nabla\phi(u) = \nabla_u c(u, v^*(u)),\ \forall u \in \mathbb{R}^n$, together with the fact that $\nabla\phi(\cdot)$ is L_ϕ-Lipschitz continuous (cf. Theorem 1 and Lemma 2) allows us to apply the *descent lemma* and obtain that

$$\phi(u^{k+1}) - \phi(u^k) \leq \langle \nabla_u c(u^k, v^*(u^k)), u^{k+1} - u^k\rangle + \frac{L_\phi}{2}\|u^{k+1} - u^k\|^2. \tag{4.9}$$

Summing Inequalities (4.8) and (4.9), we have

$$\begin{aligned}
\Theta(u^{k+1}) - \Theta(u^k) &= f(u^{k+1}) - f(u^k) + \phi(u^{k+1}) - \phi(u^k) \\
&\leq \frac{1}{2}\left(L_\phi - \frac{1}{\alpha}\right)\|u^{k+1} - u^k\|^2 + \langle \nabla_u c(u^k, v^*(u^k)) - \nabla_u c(u^k, w^k), u^{k+1} - u^k\rangle.
\end{aligned} \tag{4.10}$$

Finally, we derive the desired result using the Cauchy–Schwartz inequality, the L_{uv}-Lipschitz continuity of $\nabla_u c(u^k,\cdot)$ (cf. (2.2)), and Proposition 2(i) applied with $\gamma \equiv L_{uv}$:

$$\begin{aligned}\Theta(u^{k+1})-\Theta(u^k) &\le \frac{1}{2}\left(L_\phi-\frac{1}{\alpha}\right)\|u^{k+1}-u^k\|^2+\|\nabla_u c(u^k,v^*(u^k))-\nabla_u c(u^k,w^k)\|\,\|u^{k+1}-u^k\|\\ &\le \frac{1}{2}\left(L_\phi-\frac{1}{\alpha}\right)\|u^{k+1}-u^k\|^2+L_{uv}\|v^*(u^k)-w^k)\|\,\|u^{k+1}-u^k\|\\ &\le \frac{1}{2}\left(L_\phi-\frac{1}{\alpha}\right)\|u^{k+1}-u^k\|^2+L_{uv}\left(\frac{L_{uv}}{2}\|u^{k+1}-u^k\|^2+\frac{1}{2L_{uv}}\|v^*(u^k)-w^k\|^2\right)\\ &= \frac{1}{2}\left(L_{uv}^2+L_\phi-\frac{1}{\alpha}\right)\|u^{k+1}-u^k\|^2+\frac{1}{2}\|v^*(u^k)-w^k\|^2. \quad \square\end{aligned}$$

We end by analyzing the subgradient of the objective function Θ in terms of the u-step. Before proceeding, recall that thanks to the differentiability of ϕ (cf. Theorem 1), we can use the subdifferential rule (3.2), which justifies the following identity:

$$\partial\Theta(u)=\partial f(u)+\nabla\phi(u), \quad \forall u\in \operatorname{dom} f. \tag{4.11}$$

We prove the following subgradient bound for PGDA.

Lemma 8 (Subgradient Bound for the u-Step). *Let $\{u^k,v^k\}_{k\in\mathbb{N}}$ be the sequence generated by PGDA. Then, there exists $M>0$ such that for every $k\ge 0$, there exists $\xi^{k+1}\in\partial\Theta(u^{k+1})$, which satisfies the following inequality:*

$$\|\xi^{k+1}\| \le M(\|u^{k+1}-u^k\|+\|w^k-v^*(u^k)\|). \tag{4.12}$$

Proof. The u-step can be stated as

$$u^{k+1}\in \arg\min_{u\in\mathbb{R}^n}\left\{f(u)+\frac{1}{2\alpha}\|u-(u^k-\alpha\nabla_u c(u^k,w^k))\|^2\right\}.$$

Writing the optimality condition, we have

$$0\in\partial f(u^{k+1})+\frac{1}{\alpha}(u^{k+1}-u^k)+\nabla_u c(u^k,w^k). \tag{4.13}$$

On the other hand, thanks to (4.11), we have

$$\partial\Theta(u^{k+1})=\partial f(u^{k+1})+\nabla\phi(u^{k+1})=\partial f(u^{k+1})+\nabla_u c(u^{k+1},v^*(u^{k+1})). \tag{4.14}$$

Therefore, setting

$$\xi^{k+1}=-\frac{1}{\alpha}(u^{k+1}-u^k)-\nabla_u c(u^k,w^k)+\nabla_u c(u^{k+1},v^*(u^{k+1})),$$

it follows from (4.13) and (4.14) that $\xi^{k+1}\in\partial\Theta(u^{k+1})$.

Next, note that

$$\begin{aligned}\|\xi^{k+1}\| &= \left\|-\frac{1}{\alpha}(u^{k+1}-u^k)+\nabla_u c(u^{k+1},v^*(u^{k+1}))-\nabla_u c(u^k,w^k)\right\|\\ &= \left\|-\frac{1}{\alpha}(u^{k+1}-u^k)+\nabla_u c(u^{k+1},v^*(u^{k+1}))-\nabla_u c(u^{k+1},w^k)+\nabla_u c(u^{k+1},w^k)-\nabla_u c(u^k,w^k)\right\|\\ &\le \frac{1}{\alpha}\|u^{k+1}-u^k\|+\|\nabla_u c(u^{k+1},v^*(u^{k+1}))-\nabla_u c(u^{k+1},w^k)\|+\|\nabla_u c(u^{k+1},w^k)-\nabla_u c(u^k,w^k)\|\\ &\le \left(\frac{1}{\alpha}+L_{uu}\right)\|u^{k+1}-u^k\|+L_{uv}\|v^*(u^{k+1})-w^k\|\\ &= \left(\frac{1}{\alpha}+L_{uu}\right)\|u^{k+1}-u^k\|+L_{uv}\|v^*(u^{k+1})-v^*(u^k)+v^*(u^k)-w^k\|\\ &\le \left(\frac{1}{\alpha}+L_{uu}\right)\|u^{k+1}-u^k\|+L_{uv}\|v^*(u^{k+1})-v^*(u^k)\|+L_{uv}\|v^*(u^k)-w^k\|,\end{aligned}$$

where the first inequality and the third inequality are because of the triangle inequality and where the second inequality is because of the Lipschitz continuity properties of $\nabla_u c$ as stated by Inequalities (2.1) and (2.2).

Finally, we apply Lemma 1, and together with the definition of L_ϕ (cf. (2.21)), we have

$$\begin{aligned}\|\xi^{k+1}\| &\le \left(\frac{1}{\alpha}+L_{uu}+\frac{L_{uv}\cdot L_{vu}}{\sigma}\right)\|u^{k+1}-u^k\|+L_{uv}\|v^*(u^k)-w^k\|\\ &= \left(\frac{1}{\alpha}+L_\phi\right)\|u^{k+1}-u^k\|+L_{uv}\|v^*(u^k)-w^k\|\\ &\le M\Big(\|u^{k+1}-u^k\|+\|v^*(u^k)-w^k\|\Big),\end{aligned}$$

with

$$M=\max\left\{\frac{1}{\alpha}+L_\phi, L_{uv}\right\}. \quad \square$$

Equipped with the results, we can now utilize the general convergence analysis framework that was presented in Section 3. This requires that we define a perturbation sequence $\{v_k\}_{k\in\mathbb{N}}\subseteq\mathbb{R}_+$ such that $\{u^k,v_k\}_{k\in\mathbb{N}}$ is a *perturbed gradient-like descent sequence* with respect to Θ (i.e., Conditions 1–3 are satisfied).

We begin by defining an adequate perturbed sequence $\{v_k\}_{k\in\mathbb{N}}$ for each of the algorithms (PPGDA and APGDA, respectively) and then prove Conditions 1 and 2.

4.2. PPGDA: Perturbed Sufficient Descent and Subgradient Bound

For PPGDA, we define the perturbation sequence $\{v_k\}_{k\in\mathbb{N}}$ as follows:

$$v_k=\sqrt{s}\|v^k-v^*(u^k)\|, \tag{4.15}$$

with $s>0$ to be specified in the proof of Lemma 9. We continue with proving Conditions 1 and 2.

Lemma 9 (Perturbed Sufficient Descent for Θ—Condition 1 for PPGDA). *Let $\{u^k,v^k\}_{k\in\mathbb{N}}$ be the sequence generated by PPGDA, and assume that*

$$\alpha<\frac{1}{L} \text{ with } L:=L_{uv}^2+L_\phi+\frac{2(2\kappa^2+3\kappa+1)L_{vu}^2}{\sigma^2}. \tag{4.16}$$

Then, there exist $s>0$ and $c_1>0$ such that for every $k\in\mathbb{N}$, we have

$$c_1(\|u^{k+1}-u^k\|^2+v_k^2)\le\left(\Theta(u^k)+\frac{1}{2}v_k^2\right)-\left(\Theta(u^{k+1})+\frac{1}{2}v_{k+1}^2\right), \tag{4.17}$$

with v_k as defined in (4.15).

Proof. For every $k\in\mathbb{N}$, let

$$\Delta_k:=\left(\Theta(u^{k+1})+\frac{1}{2}v_{k+1}^2\right)-\left(\Theta(u^k)+\frac{1}{2}v_k^2\right).$$

Applying Lemma 7 with $w^k\equiv v^k$ (cf. (4.7)) and together with the definition of v_k (cf. (4.15)), we obtain that for every $s>0$,

$$\Delta_k\le\frac{1}{2}\left(L_{uv}^2+L_\phi-\frac{1}{\alpha}\right)\|u^{k+1}-u^k\|^2+\frac{1}{2s}v_k^2+\frac{1}{2}(v_{k+1}^2-v_k^2).$$

Next, we observe that (4.3) of Lemma 6 implies that

$$v_{k+1}^2\le\frac{(\kappa+1/2)}{\kappa+1}v_k^2+\frac{sL_{vu}^2(2\kappa+1)}{\sigma^2}\|u^{k+1}-u^k\|^2.$$

Combining this, we obtain that

$$\Delta_k\le-\frac{a_1}{2}\|u^{k+1}-u^k\|^2-\frac{a_2}{2}v_k^2,$$

with

$$a_1 = 1/\alpha - L_{uv}^2 - L_\phi - \frac{sL_{vu}^2(2\kappa+1)}{\sigma^2} \quad \text{and} \quad a_2 = 1 - \frac{\kappa+1/2}{\kappa+1} - \frac{1}{s} = \frac{s-2(\kappa+1)}{2s(\kappa+1)}.$$

Setting $c_1 = \min\{a_1, a_2\}/2$, it remains to determine a positive value for s such that both a_1 and a_2 are positive (i.e., with

$$\mu_1 = 2(\kappa+1) \quad \text{and} \quad \mu_2 = \frac{(1/\alpha - L_{uv}^2 - L_\phi)\sigma^2}{L_{vu}^2(2\kappa+1)},$$

we should prove that $0 \le \mu_1 < \mu_2$ and set $s \in (\mu_1, \mu_2)$). Indeed, $2(\kappa+1) > 0$, and by simple algebra, we obtain that

$$\begin{aligned}\mu_2 - \mu_1 &= \frac{(1/\alpha - L_{uv}^2 - L_\phi)\sigma^2 - 2(\kappa+1)L_{vu}^2(2\kappa+1)}{L_{vu}^2(2\kappa+1)} \\ &= \frac{\sigma^2}{L_{vu}^2(2\kappa+1)}\left(\frac{1}{\alpha} - L_{uv}^2 - L_\phi - \frac{2(2\kappa^2+3\kappa+1)L_{vu}^2)}{\sigma^2}\right) > 0,\end{aligned}$$

where the last inequality is because of (4.16). □

Lemma 10 (Perturbed Subgradient Lower Bound for the Iterates Gap—Condition 2 for PPGDA). *Let $\{u^k, v^k\}_{k\in\mathbb{N}}$ be the sequence generated by PPGDA and $\{v_k\}_{k\in\mathbb{N}}$ be as defined in (4.15), with $s>0$. Then, there exists $c_2 > 0$ such that for every $k \in \mathbb{N}$, there exists $\xi^{k+1} \in \partial\Theta(u^{k+1})$, which satisfies*

$$\|\xi^{k+1}\| \le c_2(\|u^{k+1} - u^k\| + v_k). \tag{4.18}$$

Proof. Applying Lemma 8 with $w^k \equiv v^k$, we obtain that there exists $M>0$ such that for every $k \ge 0$, there exists $\xi^{k+1} \in \partial\Theta(u^{k+1})$ bounded as follows:

$$\|\xi^{k+1}\| \le M\|u^{k+1} - u^k\| + M\|v^k - v^*(u^k)\| = M\|u^{k+1} - u^k\| + \frac{M}{\sqrt{s}}v_k \le c_2(\|u^{k+1} - u^k\| + v_k),$$

with

$$c_2 = \max\{M, M/\sqrt{s}\} \quad \square$$

4.3. APGDA: Perturbed Sufficient Descent and Subgradient Bound

Next, we turn to APGDA. In this case, we define the perturbation sequence $\{v_k\}_{k\in\mathbb{N}}$ by

$$v_k = \sqrt{t\|u^k - u^{k-1}\|^2 + s\|v^k - v^*(u^{k-1})\|^2}, \tag{4.19}$$

with $s>0$ and $t>0$ to be specified in the proof of Lemma 11. Conditions 1 and 2 are proved in the following lemmas.

Lemma 11 (Perturbed Sufficient Descent of Θ—Condition 1 for APGDA). *Let $\{u^k, v^k\}_{k\in\mathbb{N}}$ be the sequence generated by APGDA, and assume that*

$$\alpha < \frac{1}{L} \text{ with } L := L_{uv}^2 + L_\phi + \frac{2(2\kappa^2+\kappa)L_{vu}^2}{\sigma^2}. \tag{4.20}$$

Then, there exist $s>0$, $t>0$, and $c_1>0$ such that for every $k \in \mathbb{N}$, we have

$$c_1(\|u^{k+1} - u^k\|^2 + v_k^2) \le \left(\Theta(u^k) + \frac{1}{2}v_k^2\right) - \left(\Theta(u^{k+1}) + \frac{1}{2}v_{k+1}^2\right), \tag{4.21}$$

with v_k as defined in (4.19).

Proof. For every $k \in \mathbb{N}$, let

$$\Delta_k := \left(\Theta(u^{k+1}) + \frac{1}{2}v_{k+1}^2\right) - \left(\Theta(u^k) + \frac{1}{2}v_k^2\right).$$

Applying Lemma 7 with $w^k \equiv v^{k+1}$ (cf. (4.7)) and together with the definition of ν_k (cf. (4.19)), we obtain that for every $s>0$ and $t>0$,

$$\begin{aligned}\Delta_k &\le \frac{1}{2}\left(L_{uv}^2 + L_\phi - \frac{1}{\alpha}\right)\|u^{k+1}-u^k\|^2 + \frac{1}{2}\|v^{k+1}-v^*(u^k)\|^2 + \frac{1}{2}(\nu_{k+1}^2-\nu_k^2)\\ &= \frac{1}{2}\left(t + L_{uv}^2 + L_\phi - \frac{1}{\alpha}\right)\|u^{k+1}-u^k\|^2 + \frac{s+1}{2}\|v^{k+1}-v^*(u^k)\|^2 - \frac{1}{2}\nu_k^2.\end{aligned}$$

Next, we set $t = 2\kappa L_{vu}^2 s/\sigma^2$ and observe that by (4.4) of Lemma 6, we have

$$\|v^{k+1}-v^*(u^k)\|^2 \le \frac{\kappa+1/2}{\kappa+1}\left(\|v^k-v^*(u^{k-1})\|^2 + \frac{t}{s}\|u^k-u^{k-1}\|^2\right) = \frac{\kappa+1/2}{s(\kappa+1)}\nu_k^2.$$

Combining this, we obtain that

$$\Delta_k \le -\frac{a_1}{2}\|u^{k+1}-u^k\|^2 - \frac{a_2}{2}\nu_k^2,$$

with

$$a_1 = 1/\alpha - L_{uv}^2 - L_\phi - 2\kappa L_{vu}^2 s/\sigma^2 \quad \text{and} \quad a_2 = 1 - \frac{(s+1)(\kappa+1/2)}{s(\kappa+1)} = \frac{s-2\kappa-1}{2s(\kappa+1)}.$$

To complete the proof, we set $c_1 = \min\{a_1,a_2\}/2$, and it remains to determine a positive value for s such that both a_1 and a_2 are positive (i.e., with

$$\mu_1 = 2\kappa+1 \quad \text{and} \quad \mu_2 = \frac{(1/\alpha - L_{uv}^2 - L_\phi)\sigma^2}{2\kappa L_{vu}^2},$$

we should prove that $0 \le \mu_1 < \mu_2$ and set $s \in (\mu_1,\mu_2)$). Indeed, $2\kappa+1>0$, and by simple algebra, we obtain that

$$\begin{aligned}\mu_2 - \mu_1 &= \frac{(1/\alpha - L_{uv}^2 - L_\phi)\sigma^2 - 2\kappa L_{vu}^2(2\kappa+1)}{2\kappa L_{vu}^2}\\ &= \frac{\sigma^2}{2\kappa L_{vu}^2}\left(\frac{1}{\alpha} - L_{uv}^2 - L_\phi - \frac{2(2\kappa^2+\kappa)L_{vu}^2}{\sigma^2}\right) > 0,\end{aligned}$$

where the last inequality is because of (4.20). □

Lemma 12 (Perturbed Subgradient Lower Bound for the Iterates Gap—Condition 2 for APGDA). *Let $\{u^k,v^k\}_{k\in\mathbb{N}}$ be the sequence generated by APGDA and $\{\nu_k\}_{k\in\mathbb{N}}$ be as defined in (4.19), with $s>0$ and $t>0$. Then, there exists $c_2>0$ such that for every $k\in\mathbb{N}$, there exists $\xi^{k+1}\in\partial\Theta(u^{k+1})$, which satisfies*

$$\|\xi^{k+1}\| \le c_2(\|u^{k+1}-u^k\| + \nu_k). \tag{4.22}$$

Proof. Applying Lemma 8 with $w^k \equiv v^{k+1}$ followed by Lemma 6 (cf. (4.2)), we obtain that there exists $M>0$ such that for every $k\ge 0$, there exists $\xi^{k+1}\in\partial\Theta(u^{k+1})$ bounded as follows:

$$\begin{aligned}\|\xi^{k+1}\| &\le M\|u^{k+1}-u^k\| + M\|v^{k+1}-v^*(u^k)\|\\ &\le M\|u^{k+1}-u^k\| + M\sqrt{\kappa/(1+\kappa)}\left(\|v^k-v^*(u^{k-1})\| + \frac{L_{vu}}{\sigma}\|u^k-u^{k-1}\|\right)\\ &\le M\|u^{k+1}-u^k\| + \mu\left(\sqrt{s}\|v^k-v^*(u^{k-1})\| + \sqrt{t}\|u^k-u^{k-1}\|\right),\end{aligned}$$

with $\mu = M\sqrt{\kappa/(1+\kappa)}\cdot\max\{1/\sqrt{s}, L_{vu}/\sigma\sqrt{t}\}$.

Finally, note that by Proposition 2(ii) with $\gamma=1$, we have

$$\sqrt{s}\|v^k-v^*(u^{k-1})\| + \sqrt{t}\|u^k-u^{k-1}\| \le \sqrt{2s\|v^k-v^*(u^{k-1})\|^2 + 2t\|u^k-u^{k-1}\|^2} = \sqrt{2}\nu_k,$$

where the last equality is because of (4.19). Therefore, we can complete the proof by setting

$$c_2 = \max\{M, \sqrt{2}\mu\}. \quad \square$$

4.4. Convergence and Iteration Complexity for PPGDA and APGDA

Proving that Conditions 1 and 2 are satisfied allows us to utilize Lemma 3 and establish the following iteration complexity result.

Theorem 2 (Iteration Complexity). *Let $\{u^k, v^k\}_{k\geq 0}$ be the sequence generated by either PPGDA or APGDA, and assume that $\underline{\Theta} := \inf_{k\in\mathbb{N}} \Theta(u^k) > -\infty$ (e.g., when $\{u^k\}_{k\in\mathbb{N}}$ is bounded or when $\inf_{u\in\mathbb{R}^d} \Theta(u) > -\infty$). In addition, suppose that $\alpha < 1/L$, with*

$$L := L_{uv}^2 + L_\phi + \frac{2\Lambda L_{vu}^2}{\sigma^2} \tag{4.23}$$

and

$$\Lambda = \begin{cases} 2\kappa^2 + 3\kappa + 1 & \text{for (PPGDA)}, \\ 2\kappa^2 + \kappa & \text{for (APGDA)}. \end{cases} \tag{4.24}$$

Then, for every $\epsilon > 0$, there exists $k = O(\epsilon^{-2})$ such that

$$\|u^{k+1} - u^k\| \leq \epsilon, \quad \operatorname{dist}(0, \partial\Theta(u^k)) \leq \epsilon, \quad \text{and} \quad \|v^k - v^*(u^k)\| \leq \epsilon. \tag{4.25}$$

Proof. Consider the sequence $\{(u^k, v_k)\}_{k\in\mathbb{N}}$, with v_k as defined by (4.15) and (4.19) for PPGDA and APGDA, respectively. This sequence satisfies Conditions 1 and 2, as can be confirmed by Lemmas 9 and 10 for PPGDA and Lemmas 11 and 12 for APGDA. Together with the assumption that $\inf_{k\in\mathbb{N}} \Theta(u^k) < \infty$, this allows us to apply Lemma 3 with $\Psi \equiv \Theta$, $\underline{\Psi} \equiv \underline{\Theta}$, and $x^k \equiv u^k$ for every $k \geq 0$ and correspondingly, with the finite constants K_1 and K_2 as defined in Lemma 3, (ii) and (iii), respectively. Then, by (3.6) and (3.8), for any $N \in \mathbb{N}$, there exists $1 \leq k \leq N$ such that

$$\|u^{k+1} - u^k\| \leq \sqrt{K_1/N}, \quad \|v_k\| \leq \sqrt{K_1/N}, \quad \text{and} \quad \operatorname{dist}(0, \partial\Psi(x^k)) \leq \sqrt{K_2/N}. \tag{4.26}$$

In addition, in the following we prove that there exists $K_3 \geq 0$ such that

$$\|v^k - v^*(u^k)\| \leq \sqrt{K_3/N}.$$

Thus, given $\epsilon > 0$, we set $N = \lceil \max\{K_1, K_2, K_3\}/\epsilon^2 \rceil$, and it follows that there exists $k \leq N$ (i.e., $k = O(\epsilon^{-2})$) such that

$$\|u^{k+1} - u^k\| \leq \sqrt{K_1/N} \leq \epsilon, \quad \|v^k - v^*(u^k)\| \leq \sqrt{K_3/N} \leq \epsilon, \quad \text{and} \quad \operatorname{dist}(0, \partial\Theta(x^k)) \leq \sqrt{K_2/N} \leq \epsilon.$$

The value of the constant K_3 is determined separately for each of the algorithms. For PPGDA, the definition of v_k (cf. (4.15)) allows us to set

$$K_3 = \frac{K_1}{s},$$

with s as defined in the proof of Lemma 9. For APGDA, v_k is given by (4.19), with s and t as set in the proof of Lemma 11, and so, we have

$$\begin{aligned} \|v^k - v^*(u^k)\| &\leq \|v^k - v^*(u^{k-1})\| + \|v^*(u^k) - v^*(u^{k-1})\| \\ &\leq \|v^k - v^*(u^{k-1})\| + \frac{L_{vu}}{\sigma}\|u^k - u^{k-1}\| \\ &\leq \left(\frac{1}{\sqrt{s}} + \frac{L_{vu}}{\sigma\sqrt{t}}\right) v_k \\ &\leq \left(\frac{1}{\sqrt{s}} + \frac{L_{vu}}{\sigma\sqrt{t}}\right) \sqrt{\frac{K_1}{N}}, \end{aligned}$$

where the first inequality is because of the triangle inequality, the second is a result of the Lipschitz continuity of $v^*(\cdot)$ (cf. Lemma 1), the third is because of the definition of v_k (cf. (4.19)), and the last is a result of Lemma 3. Thus, in this case, we set

$$K_3 = K_1 \left(\frac{1}{\sqrt{s}} + \frac{L_{vu}}{\sigma\sqrt{t}}\right)^2. \quad \square$$

Remark 3 (Simplifying the Lipschitz Constants). Obviously, Inequalities (2.1)–(2.4) stay valid when we replace the Lipschitz constants with $\overline{L} = \max\{L_{uu}, L_{uv}, L_{vu}, L_{vv}\}$. In this case, the constant L, as defined in Theorem 2 (cf. (4.23)), can be expressed as

$$L = \overline{L}^2 + \overline{L} + 2\Lambda\kappa^2,$$

with $\kappa = \overline{L}/\sigma$ and Λ as defined in (4.24).

In order to prove subsequence and global convergence of the sequence $\{u^k\}_{k\geq 0}$, generated by either algorithm, to a critical point of Θ, it is necessary to complete our argument that the sequence $\{u^k, v_k\}_{k\geq 0}$ is a perturbed gradient-like descent sequence with regard to Θ, with v_k defined, respectively, via (4.15) for PPGDA and (4.19) for APGDA. Thus, we should establish that Condition 3 is satisfied for both algorithms, assuming that the generated sequence $\{u^k, v^k\}_{k\geq 0}$ is bounded.

Lemma 13 (Continuity Condition for $\Theta(u^k)$ (Condition 3)). *Assume that the sequence $\{(u^k, v^k)\}_{k\geq 0}$ generated by either PPGDA or APGDA is bounded, and let $\{v_k\}_{k\in\mathbb{N}}$ be as defined in (4.15) or (4.19), respectively. In addition, suppose that $\{u^k, v_k\}_{k\in\mathbb{N}}$ satisfies Condition 1 with respect to Θ. Let $\{u^k\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$ be a subsequence that converges to a point $\overline{u}$. Then,*

$$\limsup_{\substack{k\to\infty \\ k\in\mathcal{K}}} \Theta(u^k) \leq \Theta(\overline{u}). \tag{4.27}$$

Proof. Recall that $\Theta(u) = f(u) + \phi(u)$, where ϕ is continuous, as proved in Theorem 1. Therefore, we have

$$\limsup_{\substack{k\to\infty \\ k\in\mathcal{K}}} \Theta(u^k) = \limsup_{\substack{k\to\infty \\ k\in\mathcal{K}}} f(u^k) + \lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} \phi(u^k) = \limsup_{\substack{k\to\infty \\ k\in\mathcal{K}}} f(u^k) + \phi(\overline{u}),$$

and it remains to prove the following inequality:

$$\limsup_{\substack{k\to\infty \\ k\in\mathcal{K}}} f(u^k) \leq f(\overline{u}). \tag{4.28}$$

Let $k \in \mathcal{K}$ and $l = k-1$. The u-step can be stated as

$$\begin{aligned} u^k = u^{l+1} &\in \arg\min_{u\in\mathbb{R}^n}\left\{ f(u) + \langle \nabla_u c(u^l, w^l), u - u^l\rangle + \frac{1}{2\alpha}\|u - u^l\|^2 \right\} \\ &= \arg\min_{u\in\mathbb{R}^n}\left\{ f(u) + \langle \nabla_u c(u^{k-1}, w^{k-1}), u - u^{k-1}\rangle + \frac{1}{2\alpha}\|u - u^{k-1}\|^2 \right\}, \end{aligned}$$

and so, we obtain that

$$\begin{aligned} &f(u^k) + \langle \nabla_u c(u^{k-1}, w^{k-1}), u^k - u^{k-1}\rangle \\ &\leq f(u^k) + \langle \nabla_u c(u^{k-1}, w^{k-1}), u^k - u^{k-1}\rangle + \frac{1}{2\alpha}\|u^k - u^{k-1}\|^2 \\ &\leq f(\overline{u}) + \langle \nabla_u c(u^{k-1}, w^{k-1}), \overline{u} - u^{k-1}\rangle + \frac{1}{2\alpha}\|\overline{u} - u^{k-1}\|^2. \end{aligned}$$

Accordingly, we have

$$\begin{aligned} f(u^k) &\leq f(\overline{u}) + \langle \nabla_u c(u^{k-1}, w^{k-1}), \overline{u} - u^k\rangle + \frac{1}{\alpha}\|\overline{u} - u^{k-1}\|^2 \\ &\leq f(\overline{u}) + \|\nabla_u c(u^{k-1}, w^{k-1})\|\,\|\overline{u} - u^k\| + \frac{1}{2\alpha}(\|\overline{u} - u^k\| + \|u^k - u^{k-1}\|)^2, \end{aligned}$$

where the second inequality is because of the Cauchy–Schwartz and triangle inequalities.

To complete the proof and confirm that

$$\limsup_{\substack{k\to\infty \\ k\in\mathcal{K}}} f(u^k) \leq f(\overline{u}),$$

we show that

$$\sup_{k\in\mathcal{K}} \|\nabla_u c(u^{k-1}, w^{k-1})\| < \infty; \tag{4.29}$$

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} \|\overline{u} - u^k\| = 0; \tag{4.30}$$

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} \|u^k - u^{k-1}\| = 0. \tag{4.31}$$

As w^{k-1} equals either v^{k-1} or v^k, then (4.29) is a result of the continuity of $\nabla_u c(\cdot,\cdot)$ and the boundness of $\{u^k, v^k\}_{k\in\mathbb{N}}$. Next, the limit (4.30) follows immediately from the convergence of the subsequence $\{u^k\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$ to $\overline{u}$. Finally, we note that the sequence $\{u^k\}_{k\geq 0}$ is bounded and that $\{u^k, v_k\}_{k\geq 0}$ satisfies Condition 1 with respect to Θ. Thus, we can apply Lemma 3(i), which validates (4.31). □

So far, we have prepared the ground for proving subsequence and global convergence of $\{u_k\}_{k\in\mathbb{N}}$ to a critical point $\overline{u}$ of Θ. In the following proposition, we determine that in such cases, the appropriate subsequence of $\{v_k\}_{k\in\mathbb{N}}$ converges to the solution of the inner maximization problem of (M) (cf. (2.13)), with $u \equiv \overline{u}$.

Proposition 3. *Let $\{u^k, v^k\}_{k\geq 0}$ be the sequence generated by either PPGDA or APGDA, and assume that $\{u^k, v^k\}_{k\geq 0}$ is bounded. Furthermore, assume that the subsequence $\{u_k\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$ converges to $\overline{u} \in \mathbb{R}^n$. Then, $\{v^k\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$ converges to $v^*(\overline{u})$, with $v^*(\cdot)$ as defined in (2.13).*

Proof. First, we consider PPGDA. Lemma 9 establishes that the sequence $\{(u^k, v_k)\}_{k\in\mathbb{N}}$ satisfies Condition 1, with $v_k = \sqrt{s}\|v^k - v^*(u^k)\|$ and $s > 0$. Combined with the boundness of $\{u^k\}_{k\geq 0}$, this fact allows us to apply Lemma 3 and obtain that v_k converges to zero (cf. (3.5)) (i.e., $\|v^k - v^*(u^k)\|$ converges to zero). As $v^*(\cdot)$ is continuous (cf. Lemma 1) and $\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} u^k = \overline{u}$, it follows that

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} v^*(u^k) = v^*(\overline{u}),$$

and so,

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} v^k = v^*(\overline{u}).$$

With APGDA, Lemma 11 allows us to apply Lemma 3. In this case,

$$v_k = \sqrt{t\|u^k - u^{k-1}\|^2 + s\|v^k - v^*(u^{k-1})\|^2},$$

with $s > 0$ and $t > 0$. As $\lim_{k\to\infty} v_k = 0$, it follows that $\|v^k - v^*(u^{k-1})\|$ and $\|u^k - u^{k-1}\|$ converge to zero. As such and as $\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} u^k = \overline{u}$, it follows that $\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} u^{k-1} = \overline{u}$. Together with the continuity of v^* (cf. Lemma 1), if follows that $\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} v^*(u^{k-1}) = v^*(\overline{u})$, and so,

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} v^k = v^*(\overline{u}). \quad \square$$

Summarizing our results regarding both PPGDA and APGDA, we conclude that under the assumption that the sequence $\{u^k, v^k\}_{k\geq 0}$ is bounded, the sequence $\{u^k, v_k\}_{k\geq 0}$ is a perturbed gradient-like descent sequence with respect to Θ; see Definition 2 and Lemmas 9–13. Thus, we can apply Lemma 4. The stability of semialgebraic sets and functions under summation and partial maximization (see Bolte et al. [10], Bolte et al. [11]) allows us to establish that the primal objective function Θ is semialgebraic when the problem data are semialgebraic. Thus, we can apply Theorem 1, and together with Proposition 3, we obtain the following convergence result.

Theorem 3 (Subsequence and Global Convergence). *Suppose that the sequence $\{u^k, v^k\}_{k\geq 0}$, generated by either PPGDA or APGDA, is bounded. In addition, assume that $\alpha < 1/L$, with L as defined in Theorem 2. Then, the following implications hold.*

i. *Let Ω be the set of cluster points of the sequence $\{u^k\}_{k\geq 0}$. Then, Ω is a nonempty and compact set; $\omega \subseteq \operatorname{crit}\Theta$; $\lim_{k\to\infty}\operatorname{dist}(u^k, \Omega) = 0$; and Θ is finite and constant on Ω. Furthermore, let $\{u\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$ be a subsequence converging to point $\overline{u} \in \Omega$. Then, the subsequence $\{v\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$ converges to $v^*(\overline{u})$.*

ii. *Assume, in addition, that the functions $f(\cdot)$, $g(\cdot)$ and $c(\cdot,\cdot)$ are semialgebraic. Then, $\{u^k\}_{k\geq 0}$ has a finite length (i.e., $\sum_{k=1}^{\infty}\|u^{k+1} - u^k\| < \infty$), and it converges to a critical point $\overline{u} \in \operatorname{crit}\Theta$. Furthermore, $\{v^k\}_{k\geq 0}$ converges to the point $v^*(\overline{u})$.*

Remark 4. As pointed out in the literature (see, e.g., Bolte et al. [11, remark 6], Bolte et al. [12, remark 5]), when global convergence has been established when the data functions are semialgebraic, then the desingularizing function for semialgebraic problems can be chosen as

$$\varphi(t) = \frac{c}{\theta} t^{\theta},\ c > 0,\ \text{and}\ \theta \in (0,1]. \tag{4.32}$$

In this case, asymptotic rate of convergence results for the iterates are well known and common, and they can be established by standard arguments using the technique of Attouch and Bolte [1]. The work of Chen et al. [16] derived such asymptotic rate results, but it also requires to verify the additional assumption that the function H (cf. (1.2)) needs to satisfy the KL property with φ of the form (4.32).

5. Non-Euclidean Extension

We recall our interpretation of the suggested algorithms as an approximation for the PGA. Given the composite Problem (P),

$$\min_{u\in\mathbb{R}^n}\{\Theta(u) := f(u) + \phi(u)\},$$

the proximal gradient step is given by (2.22),

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^n}\left\{ f(u) + \langle \nabla\phi(u^k), u\rangle + \frac{1}{2\alpha}\|u - u^k\|^2 \right\}. \tag{5.1}$$

PGA and the suggested approximations, PPGDA and APGDA, require that

i. the function f is prox friendly (i.e., the proximal step's minimization problem is tractable) and
ii. the function $\phi(\cdot)$ is C^1 with a Lipschitz continuous gradient.

The framework and the Bregman proximal algorithm (BPG), suggested in Bauschke et al. [5] for the convex setting and then extended in Bolte et al. [13] for the nonconvex one, allow us to extend the applicability of PGA to cases where the requirements are not satisfied. We adopt here a simplified presentation, which is sufficient for our purposes, and we refer the reader to Bolte et al. [13] for more details and general results allowing extended valued functions h.

Definition 4 (L-Smooth Adaptability (Bolte et al. [13, Definition 2.2])). Let $\phi, h : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable functions, and assume that h is convex. Then, we say that the pair (ϕ, h) is L_ϕ-smooth adaptable if there exists some $L > 0$ such that $Lh - \phi$ is convex.

Given the convex function $h : \mathbb{R}^n \to \mathbb{R}$, called the *kernel-generating distance* (see Bolte et al. [13, definition 2.1] for further details), the associated Bregman distance (Bregman [14]) is defined by

$$D_h(x,y) \equiv h(x) - [h(y) + \langle \nabla h(y), y - x\rangle]. \tag{5.2}$$

Thanks to the gradient inequality, $D_h(x,y) \geq 0$ holds for any $x, y \in \mathbb{R}^n$, and when h is strictly convex, one has $D_h(x,y) = 0$ if and only if $x = y$. In general, D_h is not a distance in the usual sense; it is not symmetric (unless h is the usual squared Euclidean norm) and does not satisfy the triangle inequality (see, e.g., Bauschke et al. [5] and references therein for more properties and examples).

The following lemma readily translates the L_ϕ-smooth adaptability of the pair (ϕ, h) in Definition 4 into a descent-type property for ϕ; the proof follows immediately from the convexity of the function $L_\phi h - \phi$.

Lemma 14 (Descent Lemma (Bolte et al. [13], Lemma 2.1])). *Let $\phi, h : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable functions, and assume that h is convex. Then, (ϕ, h) is L_ϕ-smooth adaptable for some $L_\phi > 0$ if and only if*

$$\phi(x) \leq \phi(y) + \langle \nabla\phi(y), x - y\rangle + L_\phi D_h(x,y), \qquad \forall x, y \in \mathbb{R}^n. \tag{5.3}$$

Inequality (5.3) naturally extends the usual descent lemma, whereby the classical squared Euclidean distance between (x, y) (recovered with $h(\cdot) = 1/2\|\cdot\|^2$) is replaced by the Bregman distance D_h.

5.1. The Non-Euclidean Model and the Algorithm

For an L_ϕ-smooth adaptable pair (ϕ, h), the main iteration of a *BPG* method for solving the nonconvex composite Model (P) is given by

$$x^+ \in T_\lambda(x) \equiv \arg\min\{f(\xi) + \langle \nabla\phi(x), \xi - x\rangle + \lambda^{-1} D_h(\xi, x) : \xi \in \mathbb{R}^n\}, \tag{5.4}$$

where $\lambda \in (0, L_\phi^{-1}]$ is a step size.

In order to simplify the current exposition, we state conditions that guarantee both the convergence rate and the subsequence and global convergence results of parallel Bregman proximal gradient descent-ascent (PBGDA) and alternating Bregman proximal gradient descent-ascent (ABGDA), the Bregman variants of the PPGDA and APGDA. We make the following standing Assumption 2, which uses and modifies Assumption 1.

Assumption 2 (Non-Euclidean Model (M)). *Let $h:\mathbb{R}^n\to\mathbb{R}$ be a 1-strongly convex and C^1 function, with ∇h being Lipschitz continuous on any bounded subset of $\mathbb{R}^n$. Suppose that Model (M) satisfies the conditions of Assumption 1 with the following addition and change.*

i. *For any $a\in\mathbb{R}^n$ and $\alpha>0$, the Bregman proximal map of $f+\langle a,\cdot\rangle$,*

$$\operatorname{prox}^h_{\alpha(f+\langle a,\cdot\rangle)}(u):=\arg\min_{x\in\mathbb{R}^n}\left\{f(u)+\langle a,x\rangle+\frac{1}{\alpha}D_h(u,x)\right\},$$

is nonempty for every $u\in\mathbb{R}^n$.

ii. *The requirement that $\nabla_u c(\cdot,v)$ is L_{uu}-Lipschitz continuous, for every $v\in\mathbb{R}^m$ (cf. (2.1)), is replaced by the assumption that there exists $M_{uu}>0$ such that $(c(,\cdot,v),h)$ is M_{uu}-smooth adaptable, for every $v\in\mathbb{R}^m$ (i.e., $M_{uu}h-c(\cdot,v)$ is convex). In addition, assume that for all bounded subsets $\mathcal{U}\in\mathbb{R}^n$ and $\mathcal{V}\in\mathbb{R}^m$, there exists $M>0$ such that $\nabla c(\cdot,v)$ is M-Lipschitz continuous over $\mathcal{U}$, for every $v\in\mathcal{V}$.*

Remark 5. When $h+\alpha f$ is supercoercive for every $\alpha>0$ (see Bolte et al. [13, assumption B]), the Bregman proximal map $\operatorname{prox}^h_{\alpha(f+\langle a,\cdot\rangle)}(u)$ is guaranteed to be nonempty.

We are now ready to state the extension of PGDA in the non-Euclidean setting, with the choice of the proximal parameter $\alpha>0$ deferred to the corresponding complexity result (cf. Theorem 4).

Algorithm 2 (BGDA) – A Unified Scheme for Parallel/Alternating Bregman Proximal Gradient Descent Ascent (PBGDA/ABGDA)

Input: $(u^0,v^0)\in\mathbb{R}^n\times\mathbb{R}^m$, $\alpha>0$, and $\beta=1/L_{vv}$.

For $k=0,1,2,\ldots$

$$v^{k+1}=\arg\max_{v\in\mathbb{R}^m}\left\{\langle\nabla_v c(u^k,v^k),v\rangle-g(v)-\frac{1}{2\beta}\|v-v^k\|^2\right\},\tag{5.5}$$

$$w^k=\begin{cases}v^k & \text{for PBGDA,}\\ v^{k+1} & \text{for ABGDA,}\end{cases}\tag{5.6}$$

$$u^{k+1}\in\arg\min_{u\in\mathbb{R}^n}\left\{f(u)+\langle\nabla_u c(u^k,w^k),u\rangle+\frac{1}{\alpha}D_h(u,u^k)\right\}.\tag{5.7}$$

5.2. Convergence and Iteration Complexity for PBGDA and ABGDA

As can be expected, the change in our assumptions with respect to Model (M) (cf. Assumptions 1 and 2) and the updated u-proximal step of the algorithms (cf. (2.11) and (5.7)) require that we adjust some of the results obtained in the previous sections. First, we note that the Lipschitz continuity of $\nabla\phi$ as stated by Lemma 2 does not hold anymore. Instead, we prove an L-smooth adaptability property for $\phi(\cdot)$ by identifying an appropriate strongly convex *kernel-generating distance*.

Lemma 15 (L-Smooth Adaptability Property for ϕ). *Let $\psi:\mathbb{R}^n\to\mathbb{R}$ be defined by*

$$\psi(u):=M_{uu}h(u)+(L_{uv}\cdot L_{vu}/\sigma)\|u\|^2,\quad\forall u\in\mathbb{R}^n.$$

Then, the pair (ϕ,ψ) is 1-smooth adaptable: that is,

$$\phi(\overline{u})-\phi(u)-\langle\nabla\phi(u),\overline{u}-u\rangle\le D_\psi(\overline{u},u).\tag{5.8}$$

Proof. We recall that (cf. (2.14))

$$\phi(u):=\max_{v\in\mathbb{R}^n}\{c(u,v)-g(v)\}=c(u,v^*(u))-g(v^*(u)),$$

and so, for every $\overline{u},u\in\mathbb{R}^n$, we have

$$c(u,v^*(u))-g(v^*(u))\ge c(u,v^*(\overline{u}))-g(v^*(\overline{u})).$$

Combining this with the fact that $\nabla\phi(u) = \nabla_u c(u, v^*(u))$ (cf. (2.18)), we obtain that

$$\begin{aligned} D_\phi(\overline{u}, u) &= \phi(\overline{u}) - \phi(u) - \langle \nabla_u \phi(u), \overline{u} - u\rangle \\ &= c(\overline{u}, v^*(\overline{u})) - g(v^*(\overline{u})) - (c(u, v^*(u)) - g(v^*(u))) - \langle \nabla_u c(u, v^*(u)), \overline{u} - u\rangle \\ &\le c(\overline{u}, v^*(\overline{u})) - g(v^*(\overline{u})) - (c(u, v^*(\overline{u})) - g(v^*(\overline{u}))) - \langle \nabla_u c(u, v^*(u)), \overline{u} - u\rangle \\ &= c(\overline{u}, v^*(\overline{u})) - c(u, v^*(\overline{u})) - \langle \nabla_u c(u, v^*(u)), \overline{u} - u\rangle. \end{aligned}$$

By further algebraic manipulation, we obtain from the last inequality that

$$\begin{aligned} D_\phi(\overline{u}, u) &\le c(\overline{u}, v^*(\overline{u})) - c(u, v^*(\overline{u})) - \langle \nabla_u c(u, v^*(\overline{u})), \overline{u} - u\rangle + \langle \nabla_u c(u, v^*(\overline{u})) - \nabla_u c(u, v^*(u)), \overline{u} - u\rangle \\ &\le M_{uu} D_h(\overline{u}, u) + \langle \nabla_u c(u, v^*(\overline{u})) - \nabla_u c(u, v^*(u)), \overline{u} - u\rangle \\ &\le M_{uu} D_h(\overline{u}, u) + \|\nabla_u c(u, v^*(\overline{u})) - \nabla_u c(u, v^*(u))\|\|\overline{u} - u\| \\ &\le M_{uu} D_h(\overline{u}, u) + L_{uv}\|v^*(\overline{u}) - v^*(u)\|\|\overline{u} - u\| \\ &\le M_{uu} D_h(\overline{u}, u) + \frac{L_{uv} \cdot L_{vu}}{\sigma}\|\overline{u} - u\|^2, \end{aligned}$$

where the second inequality is because of the assumption that the pair $(c(\cdot, v^*(\overline{u})), h)$ is M_{uu}-smooth adaptable (cf. Assumption 2(ii)), the third is because of the Cauchy–Schwartz inequality, the fourth is a result of the Lipschitz continuity of $\nabla_u c(u, \cdot)$ (cf. (2.2)), and the last is because of Lemma 1. □

We continue with adapting the convergence analysis to the change in the proximal term of the u-step (cf. (5.7)). A careful review reveals that the impact of the u-step on the analysis of PGDA is based on Lemmas 7 and 8. Thus, by proving that these lemmas still hold, although with different constants, we can establish similar complexity and convergence results (up to constant values) for BGDA.

Lemma 16 (Function Values Gap for the Bregman u-Step). *Let $\{u^k, w^k\}_{k\in\mathbb{N}}$ be the sequence generated by BGDA. Then, for every $k \ge 0$, we have*

$$\Theta(u^{k+1}) - \Theta(u^k) \le \frac{1}{2}\left(L_{uv}^2 + M_\phi - \frac{1}{\alpha}\right)\|u^{k+1} - u^k\|^2 + \frac{1}{2}\|w^k - v^*(u^k)\|^2, \tag{5.9}$$

with

$$M_\phi = M_{uu} + \frac{2L_{uv} \cdot L_{vu}}{\sigma}. \tag{5.10}$$

Proof. The u-step (5.7) can be stated as

$$u^{k+1} \in \arg\min_{u\in\mathbb{R}^n}\left\{f(u) + \langle \nabla_u c(u^k, w^k), u - u^k\rangle + \frac{1}{\alpha}D_h(u, u^k)\right\},$$

and so, we have that (recall that $D_h(u^k, u^k) \equiv 0$)

$$f(u^{k+1}) - f(u^k) \le -\frac{1}{\alpha}D_h(u^{k+1}, u^k) - \langle \nabla_u c(u^k, w^k), u^{k+1} - u^k\rangle. \tag{5.11}$$

Having $\nabla\phi(u) = \nabla_u c(u, v^*(u)),\ \forall u \in \mathbb{R}^n$ (cf. Proposition 1) together with the descent result of Lemma 15, we obtain that

$$\phi(u^{k+1}) - \phi(u^k) \le \langle \nabla_u c(u^k, v^*(u^k)), u^{k+1} - u^k\rangle + M_{uu} D_h(u^{k+1}, u^k) + \frac{L_{uv} \cdot L_{vu}}{\sigma}\|u^{k+1} - u^k\|^2. \tag{5.12}$$

Summing Inequalities (5.11) and (4.9), we have

$$\begin{aligned}\Theta(u^{k+1}) - \Theta(u^k) &= f(u^{k+1}) - f(u^k) + \phi(u^{k+1}) - \phi(u^k)\\ &\le \left(M_{uu} - \frac{1}{\alpha}\right) D_h(u^{k+1}, u^k) + \frac{L_{uv}\cdot L_{vu}}{\sigma}\|u^{k+1} - u^k\|^2 + \langle \nabla_u c(u^k, v^*(u^k)) - \nabla_u c(u^k, w^k), u^{k+1} - u^k\rangle\\ &\le \frac{1}{2}\left(M_\phi - \frac{1}{\alpha}\right)\|u^{k+1} - u^k\|^2 + \langle \nabla_u c(u^k, v^*(u^k)) - \nabla_u c(u^k, w^k), u^{k+1} - u^k\rangle, \end{aligned} \tag{5.13}$$

where the second inequality is a result of having $M_{uu} - 1/\alpha < 0$ and the assumption that h is 1-strongly convex (cf. Assumption 2) (i.e., that $D_h(u^{k+1}, u^k) \ge \frac{1}{2}\|u^{k+1} - u^k\|^2, \quad \forall k \in \mathbb{N}$). Finally, as in Lemma 7, we derive the desired result using the Cauchy–Schwartz inequality, the L_{uv}-Lipschitz continuity of $\nabla_u c(u^k, \cdot)$ (cf. (2.2)), and Proposition 2(i) applied with $\gamma \equiv L_{uv}$:

$$\begin{aligned}\Theta(u^{k+1}) - \Theta(u^k) &\le \frac{1}{2}\left(M_\phi - \frac{1}{\alpha}\right)\|u^{k+1} - u^k\|^2 + \|\nabla_u c(u^k, v^*(u^k)) - \nabla_u c(u^k, w^k)\|\,\|u^{k+1} - u^k\|\\ &\le \frac{1}{2}\left(M_\phi - \frac{1}{\alpha}\right)\|u^{k+1} - u^k\|^2 + L_{uv}\|v^*(u^k) - w^k)\|\|u^{k+1} - u^k\|\\ &\le \frac{1}{2}\left(M_\phi - \frac{1}{\alpha}\right)\|u^{k+1} - u^k\|^2 + L_{uv}\left(\frac{L_{uv}}{2}\|u^{k+1} - u^k\|^2 + \frac{1}{2L_{uv}}\|v^*(u^k) - w^k\|^2\right)\\ &= \frac{1}{2}\left(L_{uv}^2 + M_\phi - \frac{1}{\alpha}\right)\|u^{k+1} - u^k\|^2 + \frac{1}{2}\|v^*(u^k) - w^k)\|^2. \quad \square\end{aligned}$$

Lemma 17 (Subgradient Bound for the Bregman u-Step). *Assume that the sequence $\{u^k, v^k\}_{k\ge 0}$ generated by either PBGDA or ABGDA is bounded, and let w^k be as defined in (5.6), for all $k \ge 0$. Then, there exists $M > 0$ such that for every $k \ge 0$, there exists $\xi^{k+1} \in \partial\Theta(u^{k+1})$, which satisfies the following inequality*:

$$\|\xi^{k+1}\| \le M(\|u^{k+1} - u^k\| + \|w^k - v^*(u^k)\|). \tag{5.14}$$

Proof. The u-step (5.7) can be stated as

$$u^{k+1} \in \arg\min_{u\in\mathbb{R}^n}\left\{f(u) + \langle \nabla_u c(u^k, w^k), u\rangle + \frac{1}{\alpha}D_h(u, u^k)\right\}.$$

Writing the optimality condition, thanks to (4.11), we have

$$0 \in \partial f(u^{k+1}) + \frac{1}{\alpha}(\nabla h(u^{k+1}) - \nabla h(u^k)) + \nabla_u c(u^k, w^k). \tag{5.15}$$

On the other hand (and again thanks to (4.11)), we have

$$\partial\Theta(u^{k+1}) = \partial f(u^{k+1}) + \nabla\phi(u^{k+1}) = \partial f(u^{k+1}) + \nabla_u c(u^{k+1}, v^*(u^{k+1})). \tag{5.16}$$

Therefore, setting

$$\xi^{k+1} = -\frac{1}{\alpha}(\nabla h(u^{k+1}) - \nabla h(u^k)) - \nabla_u c(u^k, w^k) + \nabla_u c(u^{k+1}, v^*(u^{k+1})),$$

it follows from (5.15) and (5.16) that $\xi^{k+1} \in \partial\Theta(u^{k+1})$, and

$$\begin{aligned}\|\xi^{k+1}\| &= \left\|-\frac{1}{\alpha}(\nabla h(u^{k+1}) - \nabla h(u^k)) + \nabla_u c(u^{k+1}, v^*(u^{k+1})) - \nabla_u c(u^k, w^k)\right\|\\ &= \left\|-\frac{1}{\alpha}(\nabla h(u^{k+1}) - \nabla h(u^k)) + \nabla_u c(u^{k+1}, v^*(u^{k+1})) - \nabla_u c(u^{k+1}, w^k) + \nabla_u c(u^{k+1}, w^k) - \nabla_u c(u^k, w^k)\right\|\\ &\le \frac{1}{\alpha}\|\nabla h(u^{k+1}) - \nabla h(u^k)\| + \|\nabla_u c(u^{k+1}, v^*(u^{k+1})) - \nabla_u c(u^{k+1}, w^k)\| + \|\nabla_u c(u^{k+1}, w^k) - \nabla_u c(u^k, w^k)\|,\end{aligned}$$

where the inequality is because of the triangle inequality.

We assume that the sequence $\{u^k, v^k\}_{k\in\mathbb{N}}$ is bounded and accordingly, that $\{w^k\}_{k\in\mathbb{N}}$ is bounded as well. Thus, because of the Lipschitz continuity over bounded sets of ∇h and $\nabla_u c(\cdot, w^k)$ as portrayed by Assumption 2, there exist $M_h > 0$ and $M_c > 0$ such that for every $k \in \mathbb{N}$,

$$\|\nabla h(u^{k+1}) - \nabla h(u^k)\| \le M_h\|u^{k+1} - u^k\| \quad \text{and} \quad \|\nabla_u c(u^{k+1}, w^k) - \nabla_u c(u^k, w^k)\| \le M_c\|u^{k+1} - u^k\|.$$

Together with the L_{uv}-Lipschitz continuity of $\nabla_u c(u^{k+1}, \cdot)$ (cf. (2.2)), we obtain that

$$\begin{aligned}\|\xi^{k+1}\| &\le \left(\frac{M_h}{\alpha} + M_c\right)\|u^{k+1} - u^k\| + L_{uv}\|v^*(u^{k+1}) - w^k\| \\ &= \left(\frac{M_h}{\alpha} + M_c\right)\|u^{k+1} - u^k\| + L_{uv}\|v^*(u^{k+1}) - v^*(u^k) + v^*(u^k) - w^k\| \\ &\le \left(\frac{M_h}{\alpha} + M_c\right)\|u^{k+1} - u^k\| + L_{uv}\|v^*(u^{k+1}) - v^*(u^k)\| + L_{uv}\|v^*(u^k) - w^k\|,\end{aligned}$$

where the last inequality is because of the triangle inequality.

Finally, applying Lemma 1, we have

$$\begin{aligned}\|\xi^{k+1}\| &\le \left(\frac{M_h}{\alpha} + M_c + \frac{L_{uv}\cdot L_{vu}}{\sigma}\right)\|u^{k+1} - u^k\| + L_{uv}\|v^*(u^k) - w^k\| \\ &\le M\Big(\|u^{k+1} - u^k\| + \|v^*(u^k) - w^k\|\Big),\end{aligned}$$

with

$$M = \max\left\{\frac{M_h}{\alpha} + M_c + \frac{L_{uv}\cdot L_{vu}}{\sigma}, L_{uv}\right\}. \quad \square$$

The analysis allows us to state the following convergence and convergence rate results for the non-Euclidean setting characterized by Assumption 2.

Theorem 4 (Iteration Complexity for PBGDA/ABGDA). *Suppose that the sequence $\{u^k, v^k\}_{k\ge0}$, generated by either PBGDA or ABGDA, is bounded. In addition, assume that $\alpha < 1/L$, with*

$$L := L_{uv}^2 + M_\phi + \frac{2\Lambda L_{vu}^2}{\sigma^2}, \tag{5.17}$$

where M_ϕ is as defined in (5.10) and

$$\Lambda = \begin{cases} 2\kappa^2 + 3\kappa + 1 & \text{for PBGDA,} \\ 2\kappa^2 + \kappa & \text{for ABGDA.}\end{cases} \tag{5.18}$$

Then, for every $\epsilon > 0$, there exists $k = O(\epsilon^{-2})$ such that

$$\|u^{k+1} - u^k\| \le \epsilon, \quad \operatorname{dist}(0, \partial\Theta(u^k)) \le \epsilon, \quad \text{and} \quad \|v^k - v^*(u^k)\| \le \epsilon. \tag{5.19}$$

Theorem 5 (Subsequence and Global Convergence of PBGDA/ABGDA). *Suppose that the sequence $\{u^k, v^k\}_{k\ge0}$, generated by either PBGDA or ABGDA, is bounded. In addition, assume that $\alpha < 1/L$, with L as defined in Theorem 4. Then, the following implications hold.*

i. *Let Ω be the set of cluster points of the sequence $\{u^k\}_{k\ge0}$. Then, Ω is a nonempty and compact set; $\omega \subseteq \operatorname{crit}\Theta$; $\lim_{k\to\infty}\operatorname{dist}(u^k, \Omega) = 0$; and Θ is finite and constant on Ω. Furthermore, let $\{u\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$ be a subsequence converging to point $\overline{u} \in \Omega$. Then, the subsequence $\{v\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$ converges to $v^*(\overline{u})$.*

ii. *Assume, in addition, that the functions $f(\cdot)$, $g(\cdot)$ and $c(\cdot,\cdot)$ are semialgebraic. Then, $\{u^k\}_{k\ge0}$ has a finite length (i.e., $\sum_{k=1}^{\infty}\|u^{k+1} - u^k\| < \infty$), and it converges to a critical point $\overline{u} \in \operatorname{crit}\Theta$. Furthermore, $\{v^k\}_{k\ge0}$ converges to the point $v^*(\overline{u})$.*

Endnote

[1] We recall that a function $\psi(\cdot)$ is l-weakly convex if $\psi(\cdot) + (l/2)\|\cdot\|^2$ is convex.

References

[1] Attouch H, Bolte J (2009) On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Programming* 116(1–2):5–16.
[2] Auslender A, Teboulle M (2005) Interior projection-like methods for monotone variational inequalities. *Math. Programming* 104(1):39–68.
[3] Barazandeh B, Razaviyayn M (2020) Solving non-convex non-differentiable min-max games using proximal gradient method. *ICASSP 2020–2020 IEEE Internat. Conf. Acoustics Speech Signal Processing (ICASSP)* (IEEE, Piscataway, NJ), 3162–3166.
[4] Barazandeh B, Razaviyayn M, Sanjabi M (2019) Training generative networks using random discriminators. *2019 IEEE Data Sci. Workshop (DSW)* (IEEE, Piscataway, NJ), 327–332.
[5] Bauschke HH, Bolte J, Teboulle M (2016) A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.* 42(2):330–348.
[6] Beck A, Teboulle M (2009) Gradient-based algorithms with applications to signal recovery. Palomar DP, Eldar YC, eds. *Convex Optimization in Signal Processing and Communications* (Cambridge University Press, Cambridge, UK), 42–88.
[7] Ben-Tal A, Nemirovski A (1999) Robust solutions of uncertain linear programs. *Oper. Res. Lett.* 25(1):1–13.
[8] Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization*, vol. 28 (Princeton University Press, Princeton, NJ).
[9] Bertsekas D, Nedic A, Ozdaglar A (2003) *Convex Analysis and Optimization*, vol. 1 (Athena Scientific, Nashua, NH).
[10] Bolte J, Daniilidis A, Lewis A (2007) The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.* 17(4):1205–1223.
[11] Bolte J, Sabach S, Teboulle M (2014) Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Programming* 146(1–2):459–494.
[12] Bolte J, Sabach S, Teboulle M (2018) Nonconvex lagrangian-based optimization: Monitoring schemes and global convergence. *Math. Oper. Res.* 43(4):1210–1232.
[13] Bolte J, Sabach S, Teboulle M, Vaisbourd Y (2018) First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM J. Optim.* 28(3):2131–2151.
[14] Bregman LM (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* 7(3):200–217.
[15] Chambolle A, Pock T (2016) On the ergodic convergence rates of a first-order primal–dual algorithm. *Math. Programming* 159(1):253–287.
[16] Chen Z, Zhou Y, Xu T, Liang Y (2021) Proximal gradient descent-ascent: Variable convergence under KŁ geometry. *Internat. Conf. Learn. Representations*.
[17] Cohen E, Hallak N, Teboulle M (2022) A dynamic alternating direction of multipliers for nonconvex minimization with nonlinear functional equality constraints. *J. Optim. Theory Appl.* 193(1):324–353.
[18] Cohen E, Sabach S, Teboulle M (2021) Non-Euclidean proximal methods for convex-concave saddle-point problems. *J. Appl. Numerical Optim.* 3(1):43–60.
[19] Danskin JM (1966) The theory of max-min, with applications. *SIAM J. Appl. Math.* 14(4):641–664.
[20] Daskalakis C, Skoulakis S, Zampetakis M (2021) The complexity of constrained min-max optimization. *Proc. 53rd Annual ACM SIGACT Sympos. Theory Comput.* (ACM, New York), 1466–1478.
[21] Davis D, Drusvyatskiy D (2019) Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.* 29(1):207–239.
[22] Diakonikolas J, Daskalakis C, Jordan MI (2021) Efficient methods for structured nonconvex-nonconcave min-max optimization. *Internat. Conf. Artificial Intelligence Statist.* (PMLR, New York), 2746–2754.
[23] Drori Y, Sabach S, Teboulle M (2015) A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. *Oper. Res. Lett.* 43(2):209–214.
[24] Fiez T, Ratliff L, Mazumdar E, Faulkner E, Narang A (2021) Global convergence to local minmax equilibrium in classes of nonconvex zero-sum games. *Adv. Neural Inform. Processing Systems* 34:29049–29063.
[25] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. *Adv. Neural Inform. Processing Systems* 27:139–144.
[26] Korpelevich GM (1976) The extragradient method for finding saddle points and other problems. *Matecon* 12:747–756.
[27] Kurdyka K (1998) On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier (Grenoble)* 48(3):769–783.
[28] Lin T, Jin C, Jordan M (2020) On gradient descent ascent for nonconvex-concave minimax problems. *Internat. Conf. Machine Learn.* (PMLR, New York), 6083–6093.
[29] Liu M, Rafique H, Lin Q, Yang T (2021) First-order convergence theory for weakly-convex-weakly-concave min-max problems. *J. Machine Learn. Res.* 22(1):169.
[30] Lojasiewicz S (1963) Une propriété topologique des sous-ensembles analytiques réels. *Colloques internationaux du C.N.R.S 117. Les Équations aux Dérivées Partielles* (Gauthier-Villars, Paris), 87–89.
[31] Lu S, Tsaknakis I, Hong M, Chen Y (2020) Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *IEEE Trans. Signal Processing* 68:3676–3691.
[32] Mokhtari A, Ozdaglar A, Pattathil S (2020) A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *Internat. Conf. Artificial Intelligence Statist.* (PMLR, New York), 1497–1507.
[33] Moreau JJ (1965) Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* 93:273–299.
[34] Nemirovski A (2004) Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.* 15(1):229–251.
[35] Nouiehed M, Sanjabi M, Huang T, Lee JD, Razaviyayn M (2019) Solving a class of non-convex min-max games using iterative first order methods. *Adv. Neural Inform. Processing Systems* 32:14934–14942.
[36] Ostrovskii DM, Lowy A, Razaviyayn M (2021) Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *SIAM J. Optim.* 31(4):2508–2538.
[37] Pock T, Cremers D, Bischof H, Chambolle A (2009) An algorithm for minimizing the Mumford–Shah functional. *2009 IEEE 12th Internat. Conf. Comput. Vision* (IEEE, Piscataway, NJ), 1133–1140.

[38] Rafique H, Liu M, Lin Q, Yang T (2021) Weakly-convex–concave min–max optimization: Provable algorithms and applications in machine learning. *Optim. Methods Software* 37(3):1087–1121.
[39] Razaviyayn M, Huang T, Lu S, Nouiehed M, Sanjabi M, Hong M (2020) Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine* 37(5):55–66.
[40] Rockafellar R (1970) *Convex Analysis* (Princeton University Press, Princeton, NJ).
[41] Rockafellar R, Wets J (2004) *Variational Analysis* (Springer, Berlin).
[42] Teboulle M (2018) A simplified view of first order methods for optimization. *Math. Programming* 170(1):67–96.
[43] Thekumparampil KK, Jain P, Netrapalli P, Oh S (2019) Efficient algorithms for smooth minimax optimization. *Adv. Neural Inform. Processing Systems* 32:12680–12691.
[44] von Neumann J (1928) Zur theorie der gesellschaftsspiele. *Math. Ann.* 100(1):295–320.