Noname manuscript No.
(will be inserted by the editor)

A Unified Single-loop Alternating Gradient Projection Algorithm for Nonconvex-Concave and Convex-Nonconcave Minimax Problems

Zi Xu · Huiling Zhang · Yang Xu · Guanghui Lan

arXiv:2006.02032v5 [math.OC] 15 Jan 2023

Received: date / Accepted: date

Abstract Much recent research effort has been directed to the development of efficient algorithms for solving minimax problems with theoretical convergence guarantees due to the relevance of these problems to a few emergent applications. In this paper, we propose a unified single-loop alternating gradient projection (AGP) algorithm for solving smooth nonconvex-(strongly) concave and (strongly) convex-nonconcave minimax problems. AGP employs simple gradient projection steps for updating the primal and dual variables alternatively at each iteration. We show that it can find an ε-stationary point of the objective function in $\mathcal{O}\left(\varepsilon^{-2}\right)$ (resp. $\mathcal{O}\left(\varepsilon^{-4}\right)$) iterations under nonconvex-strongly concave (resp. nonconvex-concave) setting. Moreover, its gradient complexity to obtain an ε-stationary point of the objective function is bounded by $\mathcal{O}\left(\varepsilon^{-2}\right)$ (resp., $\mathcal{O}\left(\varepsilon^{-4}\right)$) under the strongly convex-nonconcave (resp., convex-nonconcave) setting. To the best of our knowledge, this is the first time that a simple and unified single-loop algorithm is developed for solving both nonconvex-(strongly) concave and (strongly) convex-nonconcave minimax problems. Moreover, the complexity results for solving the latter (strongly) convex-nonconcave minimax problems have never been obtained before in the literature. Numerical results show the efficiency of the proposed AGP algorithm.

Furthermore, we extend the AGP algorithm by presenting a block alternating proximal gradient (BAPG) algorithm for solving more general multi-block nonsmooth nonconvex-(strongly) concave and (strongly) convex-nonconcave minimax problems. We can similarly establish the gradient complexity of the proposed algorithm under these four different settings.

Keywords minimax optimization problem · alternating gradient projection algorithm · iteration complexity · single-loop algorithm · machine learning

Mathematics Subject Classification (2020) MSC 90C47 · MSC 90C26 · 90C30

Z. Xu was supported by National Natural Science Foundation of China under the grant 12071279 and by General Project of Shanghai Natural Science Foundation (No. 20ZR1420600). G. Lan's research was partly supported by National Science Foundation (NSF) grant CCF 1909298.

Zi Xu
Department of Mathematics, College of Sciences, Shanghai University, Shanghai 200444, P.R.China.
E-mail: xuzi@shu.edu.cn

Huiling Zhang
Department of Mathematics, College of Sciences, Shanghai University, Shanghai 200444, P.R.China.
E-mail: zhl18720009@i.shu.edu.cn

Yang Xu
Department of Mathematics, College of Sciences, Shanghai University, Shanghai 200444, P.R.China.
E-mail: fgjfhg3013@i.shu.edu.cn

Guanghui Lan
Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA.
Corresponding author. E-mail: george.lan@isye.gatech.edu

1 Introduction

We consider the following minimax optimization problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y), \tag{P}$$

where $\mathcal{X}\subseteq\mathbb{R}^n$ and $\mathcal{Y}\subseteq\mathbb{R}^m$ are nonempty closed and bounded convex sets, and $f:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is a smooth function. This problem has attracted more attention due to its wide applications in machine learning, signal processing, and many other research fields in recent years. Many practical problems can be formulated as in (P), such as the power control and transceiver design problem in signal processing [33], distributed nonconvex optimization [34,29,16], robust learning over multiple domains [46], statistical learning [1,19] and many others.

Minimax optimization problems have been studied for many years, but most previous works focused on convex-concave minimax problems, i.e., $f(x,y)$ is convex with respect to x and concave with respect to y [38,7,9]. Under this setting, Nemirovski [39] proposed a mirror-prox algorithm which returns an ε-saddle point within the complexity of $\mathcal{O}(1/\varepsilon)$ when $\mathcal{X}$ and $\mathcal{Y}$ are bounded sets. Nesterov [41] developed a dual extrapolation algorithm which owns the same complexity bound as in [39]. Monteiro and Svaiter [36,37] extended the complexity result to unbounded sets and composite objectives by using the hybrid proximal extragradient algorithm with a different termination criterion. Tseng [50] proved the same result using a refined convergence analysis. Abernethy et al. [2] presented a Hamiltonian gradient descent algorithm with last-iterate convergence under a "sufficient bilinear" condition. A few other papers have studied special cases in the convex-concave setting, for more details, we refer to [9,10,13,20,27,32,44,45] and the references therein.

However, recent applications of problem (P) in machine learning and signal processing urge the necessity of moving beyond this classical setting. For example, in a typical generative adversary network (GANs) problem formulation [48], the objective function $f(x,y)$ is nonconvex with respect to (w.r.t.) x and concave w.r.t. y. As another example, in a robust support vector machine problem, we have $y\equiv(y_u,y_v)\in\mathbb{R}^m\times\mathbb{R}$, $x\equiv(x_w,x_b)\in\mathbb{R}^m\times\mathbb{R}$ and the objective function $f(x,y)=y_v\left[\langle x_w,y_u\rangle+x_b\right]$. This problem is convex w.r.t. x, but not necessarily concave w.r.t. y.

Most of recent studies focus on nonconvex-(strongly) concave minimax problems. For nonconvex-strongly concave minimax problem, several recent works [24,47,31,33] have studied various algorithms, and all of them can achieve the gradient complexity of $\tilde{\mathcal{O}}\left(\kappa_y^2\varepsilon^{-2}\right)$ in terms of stationary point of $\Phi(\cdot)=\max_{y\in\mathcal{Y}} f(\cdot,y)$ (when $\mathcal{X}=\mathbb{R}^n$), or stationary point of f, where κ_y is the condition number for $f(x,\cdot)$. Lin et al.[32] propose an accelerated algorithm which can improve the gradient complexity bound to $\tilde{\mathcal{O}}\left(\sqrt{\kappa_y}\varepsilon^{-2}\right)$.

For general nonconvex-concave minimax problem, there are two types of algorithms, i.e., nested-loop algorithms and single-loop algorithms. One intensively studied type is nested-loop algorithms. Rafique et al. [47] propose a proximally guided stochastic mirror descent method (PG-SMD/PGSVRG), which updates x and y simultaneously, and provably converges to an approximate stationary point of $\Phi(\cdot)=\max_{y\in\mathcal{Y}} f(\cdot,y)$. However, only partial convergence results were established for nonconvex-linear minimax problem. Nouriehed et al. [43] propose an alternative multi-step framework that finds an ε-first order Nash equilibrium of f with $\tilde{\mathcal{O}}\left(\varepsilon^{-3.5}\right)$ gradient evaluations. Very recently, Thekumparampil et al. [49] propose a proximal dual implicit accelerated gradient algorithm and proved that the algorithm finds an approximate stationary point of $\Phi(\cdot)=\max_{y\in\mathcal{Y}} f(\cdot,y)$ with the rate of $\tilde{\mathcal{O}}\left(\varepsilon^{-3}\right)$. Under an equivalent notion of stationary point, Kong and Monteiro [25] propose an accelerated inexact proximal point smoothing method to achieve the same rate, however, at each outer iteration of their algorithm, a perturbed smooth approximation of the inner maximization subproblem needs to be solved, and the complexity of solving the inner problem has not been considered in [25]. Lin et al. [32] propose a class of accelerated algorithms for smooth nonconvex-concave minimax problems, which achieves a gradient complexity bound of $\tilde{\mathcal{O}}\left(\varepsilon^{-2.5}\right)$ in terms of stationary point of f. All these nested-loop algorithms either employ multiple gradient ascent steps for y's update to solve the inner subproblem exactly or inexactly, or further do similar acceleration for x's update by adding regularization terms to the inner objective function, and thus are relatively complicated to be implemented.

On the other hand, fewer studies have been directed to single-loop algorithms for nonconvex-concave minimax problems, even though these methods are more popular in practice due to their simplicity. One such method is the gradient descent-ascent (GDA) method, which performs a gradient descent step on x and a gradient ascent step on y simultaneously at each iteration. However, this algorithm fails to converge even for simple bilinear zero-sum games [28]. Many improved GDA algorithms have been proposed in [8,11,12,17,18,22]. However, the theoretical understanding of GDA is fairly limited. Very recently, by setting the stepsize of updating x in the order of ε^4, Lin et al. [31] proved that the iteration complexity of GDA to find an ε-stationary point of $\Phi(\cdot)=\max_{y\in\mathcal{Y}} f(\cdot,y)$ is bounded by $\tilde{\mathcal{O}}(\varepsilon^{-6})$ for nonconvex-concave minimax problems when $\mathcal{X}=\mathbb{R}^n$ and $\mathcal{Y}$ is a

convex compact set. If both x and y are constrained, the complexity of GDA still remains unknown. Jin et al. [24] propose a GDmax algorithm with iteration complexity $\tilde{\mathcal{O}}\left(\varepsilon^{-6}\right)$, which corresponds to the number of times the inner maximization problem is solved. Moreover, Lu et al. [33] proposed another single-loop algorithm for nonconvex minimax problems, namely the Hybrid Block Successive Approximation (HiBSA) algorithm, which can obtain an ε-stationary point of $f(x,y)$ in $\tilde{\mathcal{O}}\left(\varepsilon^{-4}\right)$ iterations when $\mathcal{X}$ and $\mathcal{Y}$ are convex compact sets. At each iteration, for updating x, one has to solve the subproblem of minimizing a strongly convex function majorizing the original function f for a fixed y, while for updating y one has to solve the subproblem of maximizing the original function $f(x,y)$ plus some regularization terms for a fixed x. Under the nonconvex-concave (not strongly concave) setting, both complexity results in [33] and [24] count the number of times the inner maximization problem is solved, without taking into account the complexity of solving the inner maximization problem. Moreover, different stopping rules have been adopted in these existing works, e.g., [32,33,43].

As mentioned earlier, another interesting class of minimax problems is the (strongly) convex-nonconcave setting of (P), i.e., $f(x,y)$ is (strongly) convex w.r.t. x and nonconcave w.r.t. y. However, for any given x, to solve the inner maximization subproblem, i.e., $\max_{y\in\mathcal{Y}} f(x,y)$, is already NP-hard. Due to this reason, almost all the existing nested-loop algorithms will lose their theoretical guarantees since they need to solve the inner subproblem exactly, or approximately with an error proportional to the accuracy ε. Most existing single-loop algorithms, e.g., HiBSA or GDmax, will also get stuck, since they require the solution of the inner maximization problem. One possible alternative approach would be to switch the order of the "min" and "max" operators. However, in general, $\min\limits_{x\in\mathcal{X}}\max\limits_{y\in\mathcal{Y}} f(x,y) \neq \max\limits_{y\in\mathcal{Y}}\min\limits_{x\in\mathcal{X}} f(x,y)$ when $f(x,y)$ is nonconvex w.r.t. x or nonconcave w.r.t. y. The set of stationary points for these problems obtained by switching the order of "min" and "max" could also be different under some criterions, e.g., $\|\nabla\Phi(\cdot)\| \leq \varepsilon$ with $\Phi(\cdot) = \max_{y\in\mathcal{Y}} f(\cdot,y)$ which is meaningful when $f(\cdot,y)$ is strongly concave with respect to y as defined in [32]. Whereas the set of stationary points for the above two problems might the same (e.g., in terms of the stationarity of f), the algorithms applied to these problems may have drastically different trajectories and would converge to quite different solutions.

It is worth mentioning the most general nonconvex-nonconcave minimax problems, i.e., $f(x,y)$ is nonconvex w.r.t. x and nonconcave w.r.t y. It should be noted that it is unclear whether a stationary point exists or not for solving these general minimax problems. Most recent works aimed at defining a notion of goodness or developing new practical algorithms for reducing oscillations and speeding up the convergence of gradient dynamics [3,12, 21,23,35]. The convergence results for all these algorithms hold either in local region or asymptotically and hence can not imply the global convergence rate. Recently, Flokas et al. [15] analyze the GDA algorithm for nonconvex-nonconcave Zero-Sum Games, i.e., $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$. Lin et al. [30] propose a proximal algorithm to solve a special case of the nonconvex-nonconcave minimax problem, where the function $f(x,y)$ satisfies a generalized monotone variational inequality condition (see [14]), and show its convergence to stationary points. Nouriehed et al. [43] study a special class of nonconvex-nonconcave minimax problems, in which $f(x,\cdot)$ satisfies the Polyak-Łojasiewic(PL) condition. More recently, Yang et al. [51] study the two-sided PL minimax problems and proposed a variance reduced strategy for solving them.

Contributions. In this paper, we focus on single-loop algorithms for solving nonconvex-concave and convex-nonconcave minimax problems. Our main contributions are as follows.

We propose a simple and unified single-loop Alternating Gradient Projection (AGP) algorithm for solving both nonconvex-(strongly) concave and (strongly) convex-nonconcave smooth minimax problems. At each iteration, only simple gradient projection steps are employed for updating x and y alternatively.

We analyze the gradient complexity of the proposed unified AGP algorithm under four different settings. For the nonconvex-concave setting, we show that an ε-stationary point of f can be obtained in $\mathcal{O}\left(\varepsilon^{-2}\right)$ (resp. $\mathcal{O}\left(\varepsilon^{-4}\right)$) iterations for nonconvex-strongly concave (resp. nonconvex concave) minimax problems. To the best of our knowledge, these represent the state-of-the-art single loop algorithms under nonconvex-concave setting. Secondly, we show that the gradient complexity to obtain an ε-stationary point of f is $\mathcal{O}\left(\varepsilon^{-2}\right)$ (resp., $\mathcal{O}\left(\varepsilon^{-4}\right)$) under the strongly convex-nonconcave (resp., convex-nonconcave) setting. To the best of our knowledge, these are the first two theoretically guaranteed convergence results reported in the literature under these settings. Existing single-loop algorithms under both nonconvex-concave and convex-nonconcave settings are summarized in Table 1.

As shown in Table 1, AGP matches the best-known $O(\epsilon^{-2})$ complexity for the nonconvex-strongly concave setting, whereas for the general nonconvex-concave setting, it improves the best-known complexity for single-loop algorithms from $O(\epsilon^{-6})$ to $O(\epsilon^{-4})$. A key step in our development is to construct a suitable potential function which involves function value plus some distance between two adjacent iterates for the nonconvex-strongly concave and strongly convex-nonconcave case, whereas an additional quadratic regularization term of x or y is needed for the general nonconvex-concave and convex-nonconcave case (see Lemma 3.4 and Lemma

4.4). To the best of our knowledge, this is the first time that such potential functions have been constructed for solving minimax problems. Moreover, the functional descent results in Lemmas 3.1, 3.3, 4.1, and 4.3 have not been reported before and appear to be novel from our point of view. It should be noted that the complexity of AGP does not match the best-known $O(\epsilon^{-2.5})$ complexity possessed by nested-loop algorithms for the general nonconvex-concave case. One possible reason is that only one gradient projection step is employed for solving the inner problem in single loop algorithms, and hence the error associated with the gradient for the outer problem will be larger than that for nested loop algorithms and such errors will accumulate as the algorithms proceeds. This also explains why their iteration complexity analysis might be more difficult than that of nested-looped algorithms. It is not evident to us whether or not this complexity bound obtained for single-loop algorithms can be further improved. Nevertheless, due to its simplicity and the fact that it does not require the input and fine-tuning of too many algorithmic parameters, the proposed AGP algorithm can numerically outperform the state-of-the-art nested loop algorithms as shown in Section 6.

Furthermore, AGP provides a flexible algorithmic framework that can be easily generalized to solve more complicated minimax problems. Based on the basic idea of the AGP algorithm, we propose a block alternating proximal gradient (BAPG) algorithm for solving more general nonsmooth multi-block nonconvex-(strongly) concave and (strongly) convex-nonconcave minimax problems. Each BAPG iteration requires only simple proximal gradient steps to update each block of the multi-block variables alternatively. We prove the gradient complexity of the proposed BAPG algorithm under these four different settings. To the best of our knowledge, existing algorithms (especially the ones with nested loop) can hardly be extended to these more complicated multi-block settings and these complexity results have not been obtained before.

Table 1 Single-loop Algorithms for solving (P) in one block smooth nonconvex-concave and convex-nonconcave settings

Algorithms	Nonconvex-concave setting		Convex-Nonconcave setting	
	strongly concave	general concave	strongly convex	general convex
GDA[31]	$\mathcal{O}\left(\varepsilon^{-2}\right)$	$\tilde{\mathcal{O}}\left(\varepsilon^{-6}\right)^{1}$	Unknown	Unknown
GDmax[24]	$\mathcal{O}\left(\varepsilon^{-2}\right)$	$\tilde{\mathcal{O}}\left(\varepsilon^{-6}\right)^{1,3}$	Unknown	Unknown
HiBSA[33]	$\mathcal{O}\left(\varepsilon^{-2}\right)$	$\tilde{\mathcal{O}}\left(\varepsilon^{-4}\right)^{2,3}$	Unknown	Unknown
AGP	$\mathcal{O}\left(\varepsilon^{-2}\right)$	$\mathcal{O}\left(\varepsilon^{-4}\right)^{2}$	$\mathcal{O}\left(\varepsilon^{-2}\right)$	$\mathcal{O}\left(\varepsilon^{-4}\right)$

[1] This complexity is to obtain an ε-stationary point of $\Phi(\cdot) = \max_{y\in\mathcal{Y}} f(\cdot, y)$, i.e., $\|\nabla\Phi(\cdot)\| \le \varepsilon$ when $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y}$ is a convex compact set.
[2] This complexity is to obtain an ε-stationary point of f which is defined as in Section 2, when $\mathcal{X}$ and $\mathcal{Y}$ are both convex compact sets.
[3] This complexity corresponds to the number of times the inner maximization problem is solved. Thus it does not consider the complexity of solving the inner problem.

The rest of this paper is organized as follows. In Section 2, we propose a unified alternating gradient projection (AGP) algorithm for nonconvex-(strongly) concave and (strongly) convex-concave minimax problems, and we then analyze the corresponding gradient complexity for four different settings in Section 3 and Section 4. We propose a block alternating proximal gradient (BAPG) algorithm for solving more general nonsmooth multi-block nonconvex-(strongly) concave and (strongly) convex-nonconcave minimax problems, and also establish the corresponding gradient complexity for four different settings in Section 5. We report some numerical results in Section 6 and make some concluding remarks in the last section.

Notation. For vectors, we use $\|\cdot\|$ to denote the l_2-norm. For a function $f(x, y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, we use $\nabla_x f(x, y)$ (or $\nabla_y f(x, y)$) to denote the partial gradient of f with respect to the first variable (or the second variable) at point (x, y). Let $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$ denote projections onto the sets $\mathcal{X}$ and $\mathcal{Y}$. Finally, we use the notation $\mathcal{O}(\cdot)$ to hide only absolute constants which do not depend on any problem parameter, and $\tilde{\mathcal{O}}(\cdot)$ notation to hide only absolute constants and log factors.

A continuously differentiable function $f(\cdot)$ is called θ-strongly convex if there exists a constant $\theta > 0$ such that for any $x, y \in \mathcal{X}$,

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\theta}{2}\|y - x\|^2. \tag{1.1}$$

If $-f$ satisfies (1.1), $f(\cdot)$ is called θ-strongly concave. A pair (x^*, y^*) is a Nash equilibrium (or equivalently a saddle point) of function f, if $\forall x \in \mathcal{X}$, $\forall y \in \mathcal{Y}$,

$$f(x^*, y) \le f(x^*, y^*) \le f(x, y^*). \tag{1.2}$$

A pair (x^*, y^*) is a local Nash equilibrium (or equivalently a local saddle point) of f, if there exists $\delta > 0$ such that for any (x, y) in $\mathcal{X} \times \mathcal{Y}$ and $\|x - x^*\| \le \delta, \|y - y^*\| \le \delta$, (1.2) is satisfied. A pair (x^*, y^*) is an ε-first-order

Nash equilibrium of function f, if $\mathcal{X}(x^*, y^*) \le \varepsilon$ and $\mathcal{Y}(x^*, y^*) \le \varepsilon$, where

$$\mathcal{X}(x^*, y^*) := -\min_x \langle \nabla_x f(x^*, y^*), x - x^* \rangle \quad s.t.\ x \in \mathcal{X}, \|x - x^*\| \le 1,$$
$$\mathcal{Y}(x^*, y^*) := \max_y \langle \nabla_y f(x^*, y^*), y - y^* \rangle \quad s.t.\ y \in \mathcal{Y}, \|y - y^*\| \le 1.$$

2 An Alternating Gradient Projection Algorithm for (P)

In this section, we propose a unified alternating gradient projection (AGP) algorithm that will be used later for solving a few different classes of (P). Each iteration of the proposed AGP algorithm consists of two gradient projection steps for updating both x and y. Instead of the original function $f(x, y)$, at the k-th iteration, AGP uses the gradient of a regularized version of the original function, i.e.,

$$f_k(x, y) = f(x, y) + \tfrac{b_k}{2}\|x\|^2 - \tfrac{c_k}{2}\|y\|^2, \tag{2.1}$$

where $b_k \ge 0$ and $c_k \ge 0$ are two regularization parameters. More specifically, for a given pair $(x_k, y_k) \in \mathcal{X} \times \mathcal{Y}$, AGP minimizes a linearized approximation of $f_k(x, y_k)$ plus some regularized terms to update x_k as follows:

$$\begin{aligned} x_{k+1} &= \arg\min_{x \in \mathcal{X}} \langle \nabla_x f_k(x_k, y_k), x - x_k \rangle + \tfrac{\beta_k}{2}\|x - x_k\|^2 \\ &= \mathcal{P}_{\mathcal{X}}\left(x_k - \tfrac{1}{\beta_k}\nabla_x f(x_k, y_k) - \tfrac{1}{\beta_k} b_k x_k\right), \end{aligned} \tag{2.2}$$

where $\mathcal{P}_{\mathcal{X}}$ is the projection operator onto $\mathcal{X}$ and $\beta_k > 0$ denotes a stepsize parameter. Similarly, it updates y_k by maximizing a linearized approximation of $f_k(x_{k+1}, y)$ minus some regularized terms, i.e.,

$$\begin{aligned} y_{k+1} &= \arg\max_{y \in \mathcal{Y}} \langle \nabla_y f_k(x_{k+1}, y_k), y - y_k \rangle - \tfrac{\gamma_k}{2}\|y - y_k\|^2 \\ &= \mathcal{P}_{\mathcal{Y}}\left(y_k + \tfrac{1}{\gamma_k}\nabla_y f(x_{k+1}, y_k) - \tfrac{1}{\gamma_k} c_k y_k\right), \end{aligned} \tag{2.3}$$

where $\mathcal{P}_{\mathcal{Y}}$ is the projection operator onto $\mathcal{Y}$ and $\gamma_k > 0$ is another stepsize parameter. The proposed AGP method is formally stated in Algorithm 1, where sequences $\{\beta_k\}$, $\{b_k\}$, $\{\gamma_k\}$, $\{c_k\}$ and the stopping rule in Step 4 will be specified later in each of the different problem settings to be studied.

Algorithm 1 (An Alternating Gradient Projection (AGP) Algorithm)

Step 1:Input $x_1, y_1, \beta_1, \gamma_1, b_1, c_1$; Set $k = 1$.
Step 2:Calculate β_k and b_k, and perform the following update for x_k:

$$x_{k+1} = \mathcal{P}_{\mathcal{X}}\left(x_k - \tfrac{1}{\beta_k}\nabla_x f(x_k, y_k) - \tfrac{1}{\beta_k} b_k x_k\right). \tag{2.4}$$

Step 3:Calculate γ_k and c_k, and perform the following update for y_k:

$$y_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(y_k + \tfrac{1}{\gamma_k}\nabla_y f(x_{k+1}, y_k) - \tfrac{1}{\gamma_k} c_k y_k\right). \tag{2.5}$$

Step 4:If some stationary condition is satisfied, stop; otherwise, set $k = k + 1$, go to Step 2.

Observe that if we set $b_k = 0$ and $c_k = 0$, the AGP algorithm is exactly the alternating version of the GDA algorithm, which is rather natural and has been widely used by practitioners for solving minimax problems, e.g., in generative adversarial networks. However, to the best of our knowledge, even for the alternating GDA algorithm, the convergence for solving (P) has never been established before in the literature. Moreover, if b_k or c_k is not equal to 0, AGP algorithm is completely new. It turns out that b_k or c_k plays a very crucial role to guarantee the convergence of AGP when solving general nonconvex-concave or convex-nonconcave minimax problems.

Before we prove the iteration complexity of AGP algorithm for solving (P), we define the stationarity gap as the termination criterion as follows.

Definition 2.1 At each iteration of Algorithm 1, the stationarity gap for problem (P) w.r.t. $f(x,y)$ is defined as:

$$\nabla G\left(x_k, y_k\right) := \begin{bmatrix} \beta_k \left(x_k - \mathcal{P}_{\mathcal{X}}(x_k - \frac{1}{\beta_k}\nabla_x f(x_k, y_k))\right) \\ \gamma_k \left(y_k - \mathcal{P}_{\mathcal{Y}}(y_k + \frac{1}{\gamma_k}\nabla_y f(x_k, y_k))\right) \end{bmatrix}.$$

We denote $\nabla G_k := \nabla G(x_k, y_k)$, $(\nabla G_k)_x := \beta_k(x_k - \mathcal{P}_{\mathcal{X}}(x_k - \frac{1}{\beta_k}\nabla_x f(x_k, y_k)))$, and $(\nabla G_k)_y := \gamma_k(y_k - \mathcal{P}_{\mathcal{Y}}(y_k + \frac{1}{\gamma_k}\nabla_y f(x_k, y_k)))$.

Note that the widely used metrics for convex-concave minimax problems such as the min-max value or the distance to the set of optimal solutions of the min-max problem are not applicable in nonconvex cases. For the latter cases we call (x_k, y_k) an ε-stationary point of $f(x,y)$ if $\|\nabla G(x_k, y_k)\| \leq \varepsilon$ with $\nabla G(x_k, y_k)$ being defined as in Definition 2.1, which has been widely used as the optimality measure, e.g., [32,33]. In the absence of constraints, $\|\nabla G(x_k, y_k)\| \leq \varepsilon$ reduces to the standard condition $\|\nabla_x f(x_k, y_k)\| \leq \varepsilon$ and $\|\nabla_y f(x_k, y_k)\| \leq \varepsilon$ for unconstrained problems. The vector $\nabla G(x_k, y_k)$ also refers to gradient mapping at (x_k, y_k), see [42] for the details.

Definition 2.2 At each iteration of Algorithm 1, the stationarity gap for problem (P) w.r.t. $f_k(x,y)$ is defined as:

$$\nabla \tilde{G}\left(x_k, y_k\right) := \begin{bmatrix} \beta_k \left(x_k - \mathcal{P}_{\mathcal{X}}(x_k - \frac{1}{\beta_k}\nabla_x f_k(x_k, y_k))\right) \\ \gamma_k \left(y_k - \mathcal{P}_{\mathcal{Y}}(y_k + \frac{1}{\gamma_k}\nabla_y f_k(x_k, y_k))\right) \end{bmatrix}.$$

We denote $\nabla \tilde{G}_k := \nabla \tilde{G}(x_k, y_k)$, $(\nabla \tilde{G}_k)_x := \beta_k(x_k - \mathcal{P}_{\mathcal{X}}(x_k - \frac{1}{\beta_k}\nabla_x f_k\left(x_k, y_k\right)))$, and $(\nabla \tilde{G}_k)_y := \gamma_k(y_k - \mathcal{P}_{\mathcal{Y}}(y_k + \frac{1}{\gamma_k}\nabla_y f_k\left(x_k, y_k\right)))$.

We also need to make the following assumption about the smoothness of $f(x,y)$.

Assumption 2.1 $f\left(x,y\right)$ has Lipschitz continuous gradients, i.e., there exist positive scalars L_{22}, L_{12}, L_{11}, L_{21} such that for any $x, \bar{x} \in \mathcal{X}$, $y, \bar{y} \in \mathcal{Y}$,

$$\begin{aligned} \|\nabla_x f\left(x,y\right) - \nabla_x f\left(\bar{x},y\right)\| &\leq L_{11}\|x - \bar{x}\|, \\ \|\nabla_x f\left(x,y\right) - \nabla_x f\left(x,\bar{y}\right)\| &\leq L_{21}\|y - \bar{y}\|, \\ \|\nabla_y f\left(x,\bar{y}\right) - \nabla_y f\left(x,y\right)\| &\leq L_{22}\|\bar{y} - y\|, \\ \|\nabla_y f\left(x,y\right) - \nabla_y f\left(\bar{x},y\right)\| &\leq L_{12}\|x - \bar{x}\|. \end{aligned}$$

We denote $L := \max\{L_{11}, L_{22}, L_{12}, L_{21}\}$.

Under this assumption, we can first prove the following lemma for estimating bounds on changes in the function value when x_k or y_k is updated at each iteration in Algorithm 1.

Lemma 2.1 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 1. Then we have

$$f\left(x_{k+1}, y_k\right) - f\left(x_k, y_k\right) \leq -\left(\beta_k - \tfrac{L_{11}}{2}\right)\|x_{k+1} - x_k\|^2 - \frac{b_k}{2}\left[\|x_{k+1}\|^2 - \|x_k\|^2 - \|x_{k+1} - x_k\|^2\right], \tag{2.6}$$

$$f(x_{k+1}, y_{k+1}) - f(x_{k+1}, y_k) \geq \left(\gamma_k - \tfrac{L_{22}}{2}\right)\|y_{k+1} - y_k\|^2 + \frac{c_k}{2}\left[\|y_{k+1}\|^2 - \|y_k\|^2 - \|y_{k+1} - y_k\|^2\right]. \tag{2.7}$$

Proof Firstly, by the optimality condition for (2.4), we have

$$\langle \nabla_x f\left(x_k, y_k\right) + \beta_k\left(x_{k+1} - x_k\right) + b_k x_k, x_k - x_{k+1}\rangle \geq 0. \tag{2.8}$$

By Assumption 2.1, the gradient of f is Lipschitz continuous, implying that

$$f\left(x_{k+1}, y_k\right) - f\left(x_k, y_k\right) \leq \langle \nabla_x f\left(x_k, y_k\right), x_{k+1} - x_k\rangle + \tfrac{L_{11}}{2}\|x_{k+1} - x_k\|^2. \tag{2.9}$$

Adding (2.8) and (2.9), we can easily show that

$$\begin{aligned} f\left(x_{k+1}, y_k\right) - f\left(x_k, y_k\right) &\leq -\left(\beta_k - \tfrac{L_{11}}{2}\right)\|x_{k+1} - x_k\|^2 - b_k\langle x_k, x_{k+1} - x_k\rangle \\ &= -\left(\beta_k - \tfrac{L_{11}}{2}\right)\|x_{k+1} - x_k\|^2 - \frac{b_k}{2}\left[\|x_{k+1}\|^2 - \|x_k\|^2 - \|x_{k+1} - x_k\|^2\right]. \end{aligned}$$

This completes the proof of (2.6). By the optimality condition for (2.5), we have

$$\langle \nabla_y f(x_{k+1}, y_k) - \gamma_k(y_{k+1} - y_k) - c_k y_k, y_k - y_{k+1} \rangle \leq 0. \tag{2.10}$$

By Assumption 2.1, the gradient of f is Lipschitz continuous, implying that

$$f(x_{k+1}, y_{k+1}) - f(x_{k+1}, y_k) \geq \langle \nabla_y f(x_{k+1}, y_k), y_{k+1} - y_k \rangle - \frac{L_{22}}{2} \|y_{k+1} - y_k\|^2. \tag{2.11}$$

The result (2.7) then follows by adding (2.10) and (2.11).

■

In the following two sections, we will establish the iteration complexity of the AGP algorithm under four different problem settings. Although there are some different technical details under different problem settings, the main process used in these proofs is similar. Firstly, we estimate a bound on the change of function values between two adjacent iterates, shown as in Lemmas 3.1, 3.3, 4.1 and 4.3 respectively. Then, we estimate an upper bound on the weighted distance between two adjacent pairs of iterates by constructing a suitable potential function according to different properties of the objective function, shown as in Lemmas 3.2, 3.4, 4.2 and 4.4 respectively. Finally, by using those upper bounds, we prove the complexity of the algorithm through some careful parameter selection, shown as in Theorems 3.1, 3.2, 4.1 and 4.2.

3 Complexity Analysis for Nonconvex-Concave Minimax Problems

3.1 Nonconvex-Strongly Concave Setting

In this subsection, we analyze the iteration complexity of Algorithm 1 for solving nonconvex-strongly concave minimax optimization problems, i.e., $f(x, y)$ is nonconvex w.r.t. x for any fixed $y \in \mathcal{Y}$, and μ-strongly concave w.r.t. y for any given $x \in \mathcal{X}$. Under this setting, $\forall k \geq 1$, we set

$$\beta_k = \eta, \gamma_k = \tfrac{1}{\rho}, \text{ and } b_k = c_k = 0 \tag{3.1}$$

in Algorithm 1, and simplify the update for x_k and y_k as follows:

$$x_{k+1} = \mathcal{P}_{\mathcal{X}} \left(x_k - \tfrac{1}{\eta} \nabla_x f(x_k, y_k) \right), \tag{3.2}$$

$$y_{k+1} = \mathcal{P}_{\mathcal{Y}} \left(y_k + \rho \nabla_y f(x_{k+1}, y_k) \right), \tag{3.3}$$

which is exactly the alternating version of GDA algorithm. Our goal in the remaining part of this subsection is to establish the iteration complexity of Algorithm 1 under the nonconvex-strongly concave setting.

We now establish an important recursion for the AGP algorithm.

Lemma 3.1 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (3.1). If $\eta > L_{11}$, then we have

$$\begin{aligned} & f(x_{k+1}, y_{k+1}) - f(x_k, y_k) \\ \leq & -\left(\tfrac{\eta}{2} - \tfrac{L_{12}^2 \rho}{2}\right) \|x_{k+1} - x_k\|^2 - \left(\tfrac{\mu}{2} - \tfrac{1}{\rho}\right) \|y_{k+1} - y_k\|^2 \\ & - \left(\mu - \tfrac{1}{2\rho} - \tfrac{\rho L_{22}^2}{2}\right) \|y_k - y_{k-1}\|^2. \end{aligned} \tag{3.4}$$

Proof The optimality condition for y_k in (3.3) implies that $\forall y \in \mathcal{Y}$ and $\forall k \geq 1$,

$$\langle \nabla_y f(x_{k+1}, y_k) - \frac{1}{\rho}(y_{k+1} - y_k), y - y_{k+1} \rangle \leq 0. \tag{3.5}$$

By choosing $y = y_k$ in (3.5), we have

$$\langle \nabla_y f(x_{k+1}, y_k) - \frac{1}{\rho}(y_{k+1} - y_k), y_k - y_{k+1} \rangle \leq 0. \tag{3.6}$$

On the other hand, by replacing k with $k-1$ and choosing $y = y_{k+1}$ in (3.5), we obtain

$$\langle \nabla_y f(x_k, y_{k-1}) - \frac{1}{\rho}(y_k - y_{k-1}), y_{k+1} - y_k \rangle \leq 0, \tag{3.7}$$

which, in view of the fact that $f(x,y)$ is μ-strongly concave w.r.t. y for any given $x \in \mathcal{X}$, then implies that

$$\begin{aligned}
&f(x_{k+1}, y_{k+1}) - f(x_{k+1}, y_k) \\
\leq &\langle \nabla_y f(x_{k+1}, y_k), y_{k+1} - y_k \rangle - \frac{\mu}{2}\|y_{k+1} - y_k\|^2 \\
\leq &\langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_{k-1}), y_{k+1} - y_k \rangle \\
&+ \frac{1}{\rho}\langle y_k - y_{k-1}, y_{k+1} - y_k \rangle - \frac{\mu}{2}\|y_{k+1} - y_k\|^2.
\end{aligned} \tag{3.8}$$

Denoting $v_{k+1} := (y_{k+1} - y_k) - (y_k - y_{k-1})$, we can write the first inner product term in the r.h.s. of (3.8) as

$$\begin{aligned}
&\langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_{k-1}), y_{k+1} - y_k \rangle \\
= &\langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_k), y_{k+1} - y_k \rangle \\
&+ \langle \nabla_y f(x_k, y_k) - \nabla_y f(x_k, y_{k-1}), v_{k+1} \rangle \\
&+ \langle \nabla_y f(x_k, y_k) - \nabla_y f(x_k, y_{k-1}), y_k - y_{k-1} \rangle.
\end{aligned} \tag{3.9}$$

Next, we estimate the three terms in the right hand side of (3.9) respectively. By Assumption 2.1 and the Cauchy-Schwarz inequality, we can bound the first two terms according to

$$\begin{aligned}
&\langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_k), y_{k+1} - y_k \rangle \\
\leq &\frac{L_{12}^2 \rho}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{2\rho}\|y_{k+1} - y_k\|^2,
\end{aligned} \tag{3.10}$$

and

$$\langle \nabla_y f(x_k, y_k) - \nabla_y f(x_k, y_{k-1}), v_{k+1} \rangle \leq \frac{\rho L_{22}^2}{2}\|y_k - y_{k-1}\|^2 + \frac{1}{2\rho}\|v_{k+1}\|^2. \tag{3.11}$$

For the third term, by μ-strongly-concavity of f with respect to y,

$$\langle \nabla_y f(x_k, y_k) - \nabla_y f(x_k, y_{k-1}), y_k - y_{k-1} \rangle \leq -\mu\|y_k - y_{k-1}\|^2. \tag{3.12}$$

Moreover, it can be easily checked that

$$\langle y_k - y_{k-1}, y_{k+1} - y_k \rangle = \frac{1}{2}\|y_k - y_{k-1}\|^2 + \frac{1}{2}\|y_{k+1} - y_k\|^2 - \frac{1}{2}\|v_{k+1}\|^2. \tag{3.13}$$

Plugging (3.9)-(3.13) into (3.8) and rearranging the terms, we conclude that

$$\begin{aligned}
&f(x_{k+1}, y_{k+1}) - f(x_{k+1}, y_k) \\
\leq &\frac{L_{12}^2 \rho}{2}\|x_{k+1} - x_k\|^2 - (\mu - \frac{1}{2\rho} - \frac{\rho L_{22}^2}{2})\|y_k - y_{k-1}\|^2 \\
&- (\frac{\mu}{2} - \frac{1}{\rho})\|y_{k+1} - y_k\|^2.
\end{aligned} \tag{3.14}$$

By setting $\beta_k = \eta$, $b_k = 0$ in (2.6) of Lemma 2.1 and the assumption $\eta > L_{11}$, we have

$$f(x_{k+1}, y_k) - f(x_k, y_k) \leq -\frac{\eta}{2}\|x_{k+1} - x_k\|^2. \tag{3.15}$$

The proof is completed by combining (3.14) with (3.15). ∎

One may want to take the telescoping sum of (3.4) in order to provide a bound on $\sum_k(\|x_{k+1} - x_k\|^2 + \|y_{k+1} - y_k\|^2)$. However, since $(1/\rho + \rho L_{22}^2)/2 \geq L_{22} \geq \mu$, the coefficient of the third term in the r.h.s. of (3.4) is always positive. As a result, we need to further refine this relation as shown below.

Lemma 3.2 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (3.1). Also let us denote

$$f_{k+1} := f(x_{k+1}, y_{k+1}), \quad S_{k+1} := \frac{2}{\rho^2 \mu}\|y_{k+1} - y_k\|^2,$$

$$F_{k+1} := f_{k+1} + S_{k+1} - (\mu + \frac{7}{2\rho} - \frac{\rho L_{22}^2}{2} - \frac{2L_{22}^2}{\mu})\|y_{k+1} - y_k\|^2.$$

If $\eta > L_{11}$, then

$$F_{k+1} - F_k \leq -\left(\tfrac{\eta}{2} - \tfrac{\rho L_{12}^2}{2} - \tfrac{2L_{12}^2}{\rho\mu^2}\right)\|x_{k+1} - x_k\|^2 - \left(\tfrac{3\mu - \rho L_{22}^2}{2} + \tfrac{\mu - 4\rho L_{22}^2}{2\rho\mu}\right)\|y_{k+1} - y_k\|^2. \tag{3.16}$$

Proof First by (3.6) and (3.7), we have

$$\frac{1}{\rho}\langle v_{k+1}, y_{k+1} - y_k\rangle \leq \langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f\left(x_k, y_{k-1}\right), y_{k+1} - y_k\rangle, \tag{3.17}$$

which together with (3.9) then imply that

$$\begin{aligned}\frac{1}{\rho}\langle v_{k+1}, y_{k+1} - y_k\rangle \leq & \langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f\left(x_k, y_k\right), y_{k+1} - y_k\rangle \\ & + \langle \nabla_y f\left(x_k, y_k\right) - \nabla_y f\left(x_k, y_{k-1}\right), v_{k+1}\rangle \\ & + \langle \nabla_y f\left(x_k, y_k\right) - \nabla_y f\left(x_k, y_{k-1}\right), y_k - y_{k-1}\rangle.\end{aligned} \tag{3.18}$$

Similar to (3.10), we can easily see that

$$\langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f\left(x_k, y_k\right), y_{k+1} - y_k\rangle \leq \frac{L_{12}^2}{2\mu}\|x_{k+1} - x_k\|^2 + \frac{\mu}{2}\|y_{k+1} - y_k\|^2. \tag{3.19}$$

By plugging (3.11),(3.12),(3.19) into (3.18), and using the identity $\frac{1}{\rho}\langle v_{k+1}, y_{k+1} - y_k\rangle = \frac{1}{2\rho}\|y_{k+1} - y_k\|^2 + \frac{1}{2\rho}\|v_{k+1}\|^2 - \frac{1}{2\rho}\|y_k - y_{k-1}\|^2$, we conclude that

$$\begin{aligned}&\frac{1}{2\rho}\|y_{k+1} - y_k\|^2 + \frac{1}{2\rho}\|v_{k+1}\|^2 - \frac{1}{2\rho}\|y_k - y_{k-1}\|^2 \\ \leq & \frac{L_{12}^2}{2\mu}\|x_{k+1} - x_k\|^2 + \frac{\mu}{2}\|y_{k+1} - y_k\|^2 + \frac{\rho L_{22}^2}{2}\|y_k - y_{k-1}\|^2 \\ & + \frac{1}{2\rho}\|v_{k+1}\|^2 - \mu\|y_k - y_{k-1}\|^2.\end{aligned} \tag{3.20}$$

Rearranging the terms of (3.20), we have

$$\begin{aligned}&\frac{1}{2\rho}\|y_{k+1} - y_k\|^2 - \frac{1}{2\rho}\|y_k - y_{k-1}\|^2 \\ \leq & \frac{L_{12}^2}{2\mu}\|x_{k+1} - x_k\|^2 + \frac{\mu}{2}\|y_{k+1} - y_k\|^2 - \left(\mu - \frac{\rho L_{22}^2}{2}\right)\|y_k - y_{k-1}\|^2.\end{aligned} \tag{3.21}$$

Multiplying $\frac{4}{\rho\mu}$ on both sides of (3.21) and using the definition of S_{k+1}, we obtain

$$S_{k+1} - S_k \leq \frac{2L_{12}^2}{\mu^2\rho}\|x_{k+1} - x_k\|^2 + \frac{2}{\rho}\|y_{k+1} - y_k\|^2 - \left(\frac{4}{\rho} - \frac{2L_{22}^2}{\mu}\right)\|y_k - y_{k-1}\|^2.$$

It then follows from (3.4) in Lemma 3.1 and the definition of F_k that

$$F_{k+1} - F_k \leq -\left(\frac{\eta}{2} - \frac{\rho L_{12}^2}{2} - \frac{2L_{12}^2}{\rho\mu^2}\right)\|x_{k+1} - x_k\|^2 - \left(\frac{3\mu - \rho L_{22}^2}{2} + \frac{\mu - 4\rho L_{22}^2}{2\rho\mu}\right)\|y_{k+1} - y_k\|^2.$$

■

We are now ready to establish the iteration complexity for the AGP algorithm in the nonconvex-strongly concave setting. In particular, letting $\nabla G\left(x_k, y_k\right)$ be defined as in Definition 2.1 and $\varepsilon > 0$ be a given target accuracy, we provide a bound on $T(\varepsilon)$, the first iteration index to achieve an ε-stationary point, i.e., $\|\nabla G(x_k, y_k)\| \leq \varepsilon$, which is equivalent to

$$T(\varepsilon) := \min\{k \mid \|\nabla G(x_k, y_k)\| \leq \varepsilon\}. \tag{3.22}$$

Theorem 3.1 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (3.1). If the relations $\eta > L_{11}, \eta > L_{12}^2\rho + \frac{4L_{12}^2}{\rho\mu^2}$, and $\rho \leq \frac{\mu}{4L_{22}^2}$ are satisfied, then it holds that

$$T(\varepsilon) \leq \frac{F_1 - \underline{F}}{d_1\varepsilon^2},$$

where $d_1 := \min\left\{\frac{\eta}{2} - \frac{\rho L_{12}^2}{2} - \frac{2L_{12}^2}{\rho\mu^2}, \frac{3\mu - \rho L_{22}^2}{2} + \frac{\mu - 4\rho L_{22}^2}{2\rho\mu}\right\} / \max\left\{\eta^2 + 2L_{12}^2, \frac{2}{\rho^2}\right\}$ and $\underline{F} := \underline{f} - (\mu + \frac{7}{2\rho} - \frac{\rho L_{22}^2}{2} - \frac{2L_{22}^2}{\mu})\sigma_y^2$ with $\underline{f} := \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x,y)$ and $\sigma_y := \max\{\|y_1 - y_2\| \mid y_1, y_2 \in \mathcal{Y}\}$.

Proof It follows immediately from (3.1) and (3.2) that

$$\|(\nabla G_k)_x\| = \eta\|x_{k+1} - x_k\|. \tag{3.23}$$

On the other hand, by (3.3) and the triangle inequality, we obtain that

$$\begin{aligned}
&\|(\nabla G_k)_y\| \\
\leq &\frac{1}{\rho}\|\mathcal{P}_{\mathcal{Y}}(y_k + \rho\nabla_y f(x_{k+1}, y_k)) - \mathcal{P}_{\mathcal{Y}}(y_k + \rho\nabla_y f(x_k, y_k))\| + \frac{1}{\rho}\|y_{k+1} - y_k\| \\
\leq &L_{12}\|x_{k+1} - x_k\| + \frac{1}{\rho}\|y_{k+1} - y_k\|.
\end{aligned} \tag{3.24}$$

By combining (3.23) and (3.24), and using the Cauchy-Schwarz inequality, we obtain

$$\|\nabla G_k\|^2 \leq \left(\eta^2 + 2L_{12}^2\right)\|x_{k+1} - x_k\|^2 + \frac{2}{\rho^2}\|y_{k+1} - y_k\|^2. \tag{3.25}$$

Observing that $d_1 > 0$. Multiplying both sides of (3.25) by d_1, and using (3.16) in Lemma 3.2, we have

$$d_1\|\nabla G_k\|^2 \leq F_k - F_{k+1}. \tag{3.26}$$

Summing up the above inequalities from $k = 1$ to $k = T(\varepsilon)$, we obtain

$$\sum_{k=1}^{T(\varepsilon)} d_1\|\nabla G_k\|^2 \leq F_1 - F_{T(\varepsilon)+1}. \tag{3.27}$$

Note that by the definition of F_{k+1} in Lemma 3.2, we have

$$\begin{aligned}
F_{T(\varepsilon)+1} &= f_{T(\varepsilon)+1} + S_{T(\varepsilon)+1} - (\mu + \frac{7}{2\rho} - \frac{\rho L_{22}^2}{2} - \frac{2L_{22}^2}{\mu})\|y_{T(\varepsilon)+1} - y_{T(\varepsilon)}\|^2 \\
&\geq \underline{f} - (\mu + \frac{7}{2\rho} - \frac{\rho L_{22}^2}{2} - \frac{2L_{22}^2}{\mu})\sigma_y^2 = \underline{F},
\end{aligned}$$

where the inequality follows from the definitions of $\underline{f}$ and σ_y, and the facts that $S_k \geq 0$ $(\forall k \geq 1)$ and $\mu + \frac{7}{2\rho} - \frac{\rho L_{22}^2}{2} - \frac{2L_{22}^2}{\mu} \geq 0$ due to the selection of ρ. We then conclude from (3.27) that $\sum_{k=1}^{T(\varepsilon)} d_1\|\nabla G_k\|^2 \leq F_1 - F_{T(\varepsilon)+1} \leq F_1 - \underline{F}$ which, in view of the definition of $T(\varepsilon)$, implies that $\varepsilon^2 \leq (F_1 - \underline{F})/(T(\varepsilon)\cdot d_1)$ or equivalently, $T(\varepsilon) \leq (F_1 - \underline{F})/(d_1\varepsilon^2)$. ∎

A few remarks are in place for the results obtained in Theorem 3.1. First, in view of Theorem 3.1, the gradient complexity of Algorithm 1 to obtain a stationary point that satisfies $\|\nabla G(x_k, y_k)\| \leq \varepsilon$ is given by $\mathcal{O}(L^2\varepsilon^{-2})$ under the nonconvex-strongly concave setting. Second, under this setting, very few single-loop algorithms have been investigated although a few other existing algorithms can achieve similar complexity bound. In particular, it seems that even the complexity bound for the GDA algorithm remains unknown under this setting when $\mathcal{X}$ and $\mathcal{Y}$ are both convex compact sets. Compared to the state-of-the-art algorithm in [32], we improve the complexity for the nonconvex-strongly concave setting by a logarithmic factor, since we do not need to solve the inner problem at each iteration. Third, as mentioned earlier, Algorithm 1 is a single-loop alternating gradient projection method with constant stepsizes, which is very easy to implement in practice.

3.2 Complexity Analysis for General Nonconvex-Concave Setting

In this subsection, we analyze the iteration complexity of Algorithm 1 for solving (P) under the general nonconvex-concave setting. Under this setting, $\forall k \geq 1$, we set

$$b_k = 0, \quad \beta_k = \bar{\eta} + \bar{\beta}_k, \quad \gamma_k = \frac{1}{\bar{\rho}}, \tag{3.28}$$

where $\bar{\eta} > 0$, $\bar{\rho} > 0$ are two constants, and $\bar{\beta}_k$ are stepsize parameters to be defined later. We need to make the following assumption on the parameters c_k.

Assumption 3.1 $\{c_k\}$ is a nonnegative monotonically decreasing sequence.

By Assumption 2.1 and $\nabla_y f_{k-1}(x, y) = \nabla_y f(x, y) - c_{k-1} y$, we have

$$\begin{aligned} &\|\nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1})\| \\ =&\|\nabla_y f(x_k, y_k) - \nabla_y f(x_k, y_{k-1}) - c_{k-1}(y_k - y_{k-1})\| \\ \leq&(L_{22} + c_{k-1})\|y_k - y_{k-1}\|. \end{aligned} \tag{3.29}$$

Denoting $L_{22}^{'} = L_{22} + c_1$, by Assumption 3.1 and (3.29), we have

$$\|\nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1})\| \leq L_{22}^{'}\|y_k - y_{k-1}\|.$$

It then follows from the above inequality and the strong concavity of $f_{k-1}(x_k, y)$ w.r.t. y (Theorem 2.1.12 in [40]) that

$$\begin{aligned} &\langle \nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1}), y_k - y_{k-1}\rangle \\ \leq& -\frac{1}{L_{22}^{'} + c_{k-1}}\|\nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1})\|^2 - \frac{c_{k-1}L_{22}^{'}}{L_{22}^{'} + c_{k-1}}\|y_k - y_{k-1}\|^2. \end{aligned} \tag{3.30}$$

This is a key inequality that we will use to establish some important recursions for the AGP method under the nonconvex-concave setting in the following two results. This is also one of the key differences between the proof for the nonconvex-strongly concave setting and the one for the nonconvex-concave setting.

Lemma 3.3 Suppose that Assumptions 2.1 and 3.1 hold. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (3.28). If $\forall k, \bar{\beta}_k > L_{11}$ and $\bar{\rho} \leq \frac{2}{L_{22}^{'} + c_1}$, then

$$\begin{aligned} &f(x_{k+1}, y_{k+1}) - f(x_k, y_k) \\ \leq& -\left(\bar{\eta} + \frac{\bar{\beta}_k}{2} - \frac{\bar{\rho}L_{12}^2}{2}\right)\|x_{k+1} - x_k\|^2 + \frac{1}{\bar{\rho}}\|y_{k+1} - y_k\|^2 + \frac{1}{2\bar{\rho}}\|y_k - y_{k-1}\|^2 \\ &+ \frac{c_{k-1}}{2}(\|y_{k+1}\|^2 - \|y_k\|^2). \end{aligned} \tag{3.31}$$

Proof The optimality condition for y_k in (2.5) implies that, $\forall y \in \mathcal{Y}$, $\forall k \geq 1$,

$$\langle \nabla_y f_k(x_{k+1}, y_k) - \frac{1}{\bar{\rho}}(y_{k+1} - y_k), y - y_{k+1}\rangle \leq 0. \tag{3.32}$$

By choosing $y = y_k$ in (3.32), we have

$$\langle \nabla_y f_k(x_{k+1}, y_k) - \frac{1}{\bar{\rho}}(y_{k+1} - y_k), y_k - y_{k+1}\rangle \leq 0. \tag{3.33}$$

On the other hand, by replacing k with $k-1$ and choosing $y = y_{k+1}$ in (3.32), we obtain

$$\langle \nabla_y f_{k-1}(x_k, y_{k-1}) - \frac{1}{\bar{\rho}}(y_k - y_{k-1}), y_{k+1} - y_k\rangle \leq 0. \tag{3.34}$$

The concavity of $f_k(x_{k+1}, y)$ w.r.t. y together with (3.34) then imply that

$$\begin{aligned}
&f_k(x_{k+1}, y_{k+1}) - f_k(x_{k+1}, y_k)\\
\leq&\langle \nabla_y f_k(x_{k+1}, y_k) - \nabla_x f_{k-1}(x_k, y_{k-1}), y_{k+1} - y_k\rangle + \frac{1}{\bar{\rho}}\langle y_k - y_{k-1}, y_{k+1} - y_k\rangle\\
=&\langle \nabla_y f_k(x_{k+1}, y_k) - \nabla_y f_{k-1}(x_k, y_k), y_{k+1} - y_k\rangle + \langle \nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1}), v_{k+1}\rangle\\
&+ \langle \nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1}), y_k - y_{k-1}\rangle + \frac{1}{\bar{\rho}}\langle y_k - y_{k-1}, y_{k+1} - y_k\rangle,
\end{aligned} \tag{3.35}$$

where $v_{k+1} = y_{k+1} - y_k - (y_k - y_{k-1})$. We now provide bounds on the inner product terms of (3.35). Firstly, by the definition of $f_k(x_{k+1}, y_k)$ and $f_{k-1}(x_k, y_k)$, Assumptions 2.1 and 3.1, and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\langle \nabla_y f_k(x_{k+1}, y_k) - \nabla_y f_{k-1}(x_k, y_k), y_{k+1} - y_k\rangle\\
=&\langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_k), y_{k+1} - y_k\rangle - (c_k - c_{k-1})\langle y_k, y_{k+1} - y_k\rangle\\
\leq&\frac{\bar{\rho}L_{12}^2}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{2\bar{\rho}}\|y_{k+1} - y_k\|^2 - \frac{c_k - c_{k-1}}{2}(\|y_{k+1}\|^2 - \|y_k\|^2)\\
&+ \frac{c_k - c_{k-1}}{2}\|y_{k+1} - y_k\|^2\\
\leq&\frac{\bar{\rho}L_{12}^2}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{2\bar{\rho}}\|y_{k+1} - y_k\|^2 - \frac{c_k - c_{k-1}}{2}(\|y_{k+1}\|^2 - \|y_k\|^2).
\end{aligned} \tag{3.36}$$

Secondly, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\langle \nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1}), v_{k+1}\rangle \leq&\frac{\bar{\rho}}{2}\|\nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1})\|^2\\
&+ \frac{1}{2\bar{\rho}}\|v_{k+1}\|^2.
\end{aligned} \tag{3.37}$$

Thirdly, it follows from (3.30) that

$$\begin{aligned}
&\langle \nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1}), y_k - y_{k-1}\rangle\\
\leq&-\frac{1}{L_{22}^{'} + c_{k-1}}\|\nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1})\|^2 - \frac{c_{k-1}L_{22}^{'}}{L_{22}^{'} + c_{k-1}}\|y_k - y_{k-1}\|^2\\
\leq&-\frac{1}{L_{22}^{'} + c_{k-1}}\|\nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1})\|^2.
\end{aligned} \tag{3.38}$$

Also observe that

$$\frac{1}{\bar{\rho}}\langle y_{k+1} - y_k, y_k - y_{k-1}\rangle = \frac{1}{2\bar{\rho}}\|y_{k+1} - y_k\|^2 + \frac{1}{2\bar{\rho}}\|y_k - y_{k-1}\|^2 - \frac{1}{2\bar{\rho}}\|v_{k+1}\|^2. \tag{3.39}$$

Plugging (3.36)-(3.39) into (3.35), and using the definition of $f_k(x_{k+1}, y_{k+1})$ and $f_k(x_{k+1}, y_k)$ and the assumption $\frac{\bar{\rho}}{2} \leq \frac{1}{L_{22}^{'} + c_1}$, we obtain

$$\begin{aligned}
f(x_{k+1}, y_{k+1}) - f(x_{k+1}, y_k) \leq&\frac{\bar{\rho}L_{12}^2}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{\bar{\rho}}\|y_{k+1} - y_k\|^2\\
&+ \frac{1}{2\bar{\rho}}\|y_k - y_{k-1}\|^2 + \frac{c_{k-1}}{2}(\|y_{k+1}\|^2 - \|y_k\|^2).
\end{aligned} \tag{3.40}$$

By setting $\beta_k = \bar{\eta} + \bar{\beta}_k$, $b_k = 0$ in (2.6) of Lemma 2.1 and the assumption $\bar{\beta}_k > L_{11}$, we have

$$f(x_{k+1}, y_k) - f(x_k, y_k) \leq -\left(\bar{\eta} + \frac{\bar{\beta}_k}{2}\right)\|x_{k+1} - x_k\|^2. \tag{3.41}$$

The result in (3.31) then follows by adding (3.40) and (3.41). ∎

It turns out that from Lemma 3.3 we can not obtain an upper bound on the positively weighted sum of $\|x_{k+1} - x_k\|^2$ and $\|y_{k+1} - y_k\|^2$ to provide an upper bound for $\|\nabla G_k\|$. We need to further refine this result in (3.31) to overcome this difficulty. In particular, we obtain below a new inequality as in (3.46) to further investigate the relation between $\|x_{k+1} - x_k\|^2$ and $\|y_{k+1} - y_k\|^2$. Then by using this new inequality, we construct

a new potential function as shown in the following important result for Algorithm 1. Note that in Lemma 3.2 we have constructed a potential function which involves the function value plus some distance between two adjacent iterates for the nonconvex-strongly concave cases, whereas in the following lemma an additional quadratic regularization term, i.e., $\|y_{k+1}\|^2$, is added to construct another potential function for the general nonconvex-concave case.

Lemma 3.4 Suppose that Assumptions 2.1 and 3.1 hold. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (3.28). Also let us denote

$$\mathcal{S}_{k+1} := \frac{8}{\bar{\rho}^2 c_{k+1}}\|y_{k+1} - y_k\|^2 + \frac{8}{\bar{\rho}}\left(1 - \frac{c_k}{c_{k+1}}\right)\|y_{k+1}\|^2,$$
$$\mathcal{F}_{k+1} := f(x_{k+1}, y_{k+1}) + \mathcal{S}_{k+1} - \frac{15}{2\bar{\rho}}\|y_{k+1} - y_k\|^2 - \frac{c_k}{2}\|y_{k+1}\|^2.$$

If

$$\bar{\beta}_k > L_{11},\ \frac{1}{c_{k+1}} - \frac{1}{c_k} \le \frac{\bar{\rho}}{5},\ \bar{\rho} \le \frac{2}{L_{22}^{'} + c_1}, \tag{3.42}$$

then $\forall k \ge 2$,

$$\begin{aligned}\mathcal{F}_{k+1} - \mathcal{F}_k \le &-\left(\bar{\eta} + \frac{\bar{\beta}_k}{2} - \frac{\bar{\rho}L_{12}^2}{2} - \frac{16L_{12}^2}{\bar{\rho}c_k^2}\right)\|x_{k+1} - x_k\|^2 + \frac{c_{k-1} - c_k}{2}\|y_{k+1}\|^2 \\ &- \frac{9}{10\bar{\rho}}\|y_{k+1} - y_k\|^2 + \frac{8}{\bar{\rho}}\left(\frac{c_{k-1}}{c_k} - \frac{c_k}{c_{k+1}}\right)\|y_{k+1}\|^2.\end{aligned} \tag{3.43}$$

Proof By (3.33) and (3.34), we have

$$\frac{1}{\bar{\rho}}\langle v_{k+1}, y_{k+1} - y_k\rangle \le \langle \nabla_y f_k(x_{k+1}, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1}), y_{k+1} - y_k\rangle. \tag{3.44}$$

Similar to (3.35) in Lemma 3.3, (3.44) can be rewritten as

$$\begin{aligned}&\frac{1}{\bar{\rho}}\langle v_{k+1}, y_{k+1} - y_k\rangle \\ \le&\langle \nabla_y f_k(x_{k+1}, y_k) - \nabla_y f_{k-1}(x_k, y_k), y_{k+1} - y_k\rangle + \langle \nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1}), v_{k+1}\rangle \\ &+ \langle \nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1}), y_k - y_{k-1}\rangle.\end{aligned}$$

Using an argument similar to the proof of (3.36)-(3.39), the relation in (3.30), and the Cauchy-Schwarz inequality, we conclude from the above inequality that

$$\begin{aligned}&\frac{1}{2\bar{\rho}}\|y_{k+1} - y_k\|^2 + \frac{1}{2\bar{\rho}}\|v_{k+1}\|^2 - \frac{1}{2\bar{\rho}}\|y_k - y_{k-1}\|^2 \\ \le&\frac{L_{12}^2}{2a_k}\|x_{k+1} - x_k\|^2 + \frac{a_k}{2}\|y_{k+1} - y_k\|^2 - \frac{c_k - c_{k-1}}{2}(\|y_{k+1}\|^2 - \|y_k\|^2) \\ &+ \frac{\bar{\rho}}{2}\|\nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1})\|^2 - \frac{1}{L_{22}^{'} + c_{k-1}}\|\nabla_y f_{k-1}(x_k, y_k) - \nabla_y f_{k-1}(x_k, y_{k-1})\|^2 \\ &+ \frac{1}{2\bar{\rho}}\|v_{k+1}\|^2 - \frac{c_{k-1}L_{22}^{'}}{L_{22}^{'} + c_{k-1}}\|y_k - y_{k-1}\|^2 + \frac{c_k - c_{k-1}}{2}\|y_{k+1} - y_k\|^2,\end{aligned} \tag{3.45}$$

for any $a_k > 0$. Observing by $c_1 \le L_{22}^{'}$, and Assumption 3.1, we have

$$-\frac{c_{k-1}L_{22}^{'}}{c_{k-1} + L_{22}^{'}} \le -\frac{c_{k-1}L_{22}^{'}}{2L_{22}^{'}} = -\frac{c_{k-1}}{2} \le -\frac{c_k}{2}.$$

Combining $\bar{\rho} \le \frac{2}{L_{22}^{'} + c_1}$ and rearranging the terms in (3.45), we obtain

$$\begin{aligned}&\frac{1}{2\bar{\rho}}\|y_{k+1} - y_k\|^2 + \frac{c_k - c_{k-1}}{2}\|y_{k+1}\|^2 \\ \le&\frac{1}{2\bar{\rho}}\|y_k - y_{k-1}\|^2 + \frac{c_k - c_{k-1}}{2}\|y_k\|^2 + \frac{L_{12}^2}{2a_k}\|x_{k+1} - x_k\|^2 + \frac{a_k}{2}\|y_{k+1} - y_k\|^2 \\ &- \frac{c_k}{2}\|y_k - y_{k-1}\|^2.\end{aligned}$$

By multiplying $\frac{16}{\bar{\rho}c_k}$ on both sides of the above inequality, we then obtain

$$\begin{aligned}&\frac{8}{\bar{\rho}^2 c_k}\|y_{k+1}-y_k\|^2+\frac{8}{\bar{\rho}}\left(1-\frac{c_{k-1}}{c_k}\right)\|y_{k+1}\|^2\\ \le&\frac{8}{\bar{\rho}^2 c_k}\|y_k-y_{k-1}\|^2+\frac{8}{\bar{\rho}}(1-\frac{c_{k-1}}{c_k})\|y_k\|^2+\frac{8L_{12}^2}{\bar{\rho}c_k a_k}\|x_{k+1}-x_k\|^2\\ &+\frac{8a_k}{\bar{\rho}c_k}\|y_{k+1}-y_k\|^2-\frac{8}{\bar{\rho}}\|y_k-y_{k-1}\|^2. \end{aligned}\tag{3.46}$$

Setting $a_k=\frac{c_k}{2}$ in the above inequality, and using the definition of $\mathcal{S}_{k+1}$ and (3.42), we have

$$\begin{aligned}\mathcal{S}_{k+1}-\mathcal{S}_k\le&\frac{8}{\bar{\rho}}\left(\frac{c_{k-1}}{c_k}-\frac{c_k}{c_{k+1}}\right)\|y_{k+1}\|^2+\frac{16L_{12}^2}{\bar{\rho}c_k^2}\|x_{k+1}-x_k\|^2+\frac{28}{5\bar{\rho}}\|y_{k+1}-y_k\|^2\\ &-\frac{8}{\bar{\rho}}\|y_k-y_{k-1}\|^2. \end{aligned}\tag{3.47}$$

Combining (3.47) and (3.31) in Lemma 3.3, and using the definition of $\mathcal{F}_{k+1}$, we conclude

$$\begin{aligned}\mathcal{F}_{k+1}-\mathcal{F}_k\le&-\left(\bar{\eta}+\frac{\bar{\beta}_k}{2}-\frac{\bar{\rho}L_{12}^2}{2}-\frac{16L_{12}^2}{\bar{\rho}c_k^2}\right)\|x_{k+1}-x_k\|^2+\frac{c_{k-1}-c_k}{2}\|y_{k+1}\|^2\\ &-\frac{9}{10\bar{\rho}}\|y_{k+1}-y_k\|^2+\frac{8}{\bar{\rho}}\left(\frac{c_{k-1}}{c_k}-\frac{c_k}{c_{k+1}}\right)\|y_{k+1}\|^2. \end{aligned}$$

■

We are now ready to establish the iteration complexity for the AGP algorithm to achieve an ε-stationary point in the general nonconvex-concave setting.

Theorem 3.2 Suppose that Assumption 2.1 holds. Let $\{(x_k,y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (3.28). If $\bar{\rho}\le\frac{1}{10L_{22}}$, $\tau>\max\{\frac{19^2(L_{11}+2\bar{\eta}-\bar{\rho}L_{12}^2)}{20^2\cdot 16\bar{\rho}L_{12}^2},2\}$, $c_k=\frac{19}{20\bar{\rho}k^{1/4}}, k\ge 1$, then for any given $\varepsilon>0$,

$$T(\varepsilon)\le\max\left(\left(\frac{2\cdot 80^2\bar{\rho}(\tau-2)L_{12}^2d_2d_3}{19^2\varepsilon^2}+1\right)^2,\frac{19^4\hat{\sigma}_y^4}{10^4\bar{\rho}^4\varepsilon^4}\right),$$

where $d_2=\mathcal{F}_2-\underline{\mathcal{F}}+(8\cdot 2^{1/4}+\frac{19}{40})\frac{\hat{\sigma}_y^2}{\bar{\rho}}$, $\bar{d}_1=\frac{8\tau^2}{(\tau-2)^2}+\frac{19^4\left(2\left(\bar{\rho}L_{12}^2-\bar{\eta}\right)^2+2L_{12}^2\right)}{64\cdot 20^4\bar{\rho}^2(\tau-2)^2L_{12}^4}$, $d_3=\max\{\bar{d}_1,\frac{19^2}{1440\sqrt{2}(\tau-2)\bar{\rho}^2L_{12}^2}\}$, $\hat{\sigma}_y:=\max\{\|y\|\mid y\in\mathcal{Y}\}$, $\underline{\mathcal{F}}:=\underline{f}-\left(2^{13/4}+15+\frac{19}{40}\right)\frac{\hat{\sigma}_y^2}{\bar{\rho}}$ with $\underline{f}:=\min_{(x,y)\in\mathcal{X}\times\mathcal{Y}}f(x,y)$.

Proof By $\bar{\rho}\le\frac{1}{10L_{22}}$ and $c_k=\frac{19}{20\bar{\rho}k^{1/4}},\forall k\ge 1$, let us denote $\bar{\beta}_k=\bar{\rho}L_{12}^2+\frac{16\tau L_{12}^2}{\bar{\rho}c_k^2}-2\bar{\eta}$, $\alpha_k=\frac{8(\tau-2)L_{12}^2}{\bar{\rho}c_k^2}$, we can easily see that the relations in (3.42) are satisfied. It follows from the selection of $\bar{\beta}_k$ and α_k that

$$\bar{\eta}+\frac{\bar{\beta}_k}{2}-\frac{\bar{\rho}L_{12}^2}{2}-\frac{16L_{12}^2}{\bar{\rho}c_k^2}=\alpha_k.$$

This observation, in view of Lemma 3.4, then immediately implies that

$$\begin{aligned}\alpha_k\|x_{k+1}-x_k\|^2+\frac{9}{10\bar{\rho}}\|y_{k+1}-y_k\|^2\le&\mathcal{F}_k-\mathcal{F}_{k+1}+\frac{8}{\bar{\rho}}\left(\frac{c_{k-1}}{c_k}-\frac{c_k}{c_{k+1}}\right)\|y_{k+1}\|^2\\ &+\frac{c_{k-1}-c_k}{2}\|y_{k+1}\|^2. \end{aligned}\tag{3.48}$$

We can easily check from the definition of $f_k(x,y)$ that

$$\|\nabla G_k\|-\|\nabla\tilde{G}_k\|\le c_k\|y_k\|.$$

By replacing f with f_k, η with $\bar{\beta}_k+\bar{\eta}$, similar to (3.23) and (3.24), we immediately obtain that

$$\|(\nabla\tilde{G}_k)_x\|=\left(\bar{\beta}_k+\bar{\eta}\right)\|x_{k+1}-x_k\|,\tag{3.49}$$

and

$$\|(\nabla \tilde{G}_k)_y\| \leq \frac{1}{\bar{\rho}}\|y_{k+1} - y_k\| + L_{12}\|x_{k+1} - x_k\|. \tag{3.50}$$

Combining (3.49) and (3.50), and using the Cauchy-Schwarz inequality, we have

$$\|\nabla \tilde{G}_k\|^2 \leq \left(\left(\bar{\beta}_k + \bar{\eta}\right)^2 + 2L_{12}^2\right)\|x_{k+1} - x_k\|^2 + \frac{2}{\bar{\rho}^2}\|y_{k+1} - y_k\|^2. \tag{3.51}$$

Since both α_k and $\bar{\beta}_k$ are in the same order when k becomes large enough, it then follows from the definition of $\bar{d}_1$ that $\forall k \geq 1$,

$$\begin{aligned}
\frac{\left(\bar{\beta}_k + \bar{\eta}\right)^2 + 2L_{12}^2}{\alpha_k^2} &= \frac{\left(\bar{\rho}L_{12}^2 + \frac{16\tau L_{12}^2}{\bar{\rho}c_k^2} - \bar{\eta}\right)^2 + 2L_{12}^2}{\alpha_k^2} \\
&\leq \frac{2\left(\frac{16\tau L_{12}^2}{\bar{\rho}c_k^2}\right)^2 + 2\left(\bar{\rho}L_{12}^2 - \bar{\eta}\right)^2 + 2L_{12}^2}{\alpha_k^2} \\
&= \frac{8\tau^2}{(\tau-2)^2} + \frac{19^4\left(2\left(\bar{\rho}L_{12}^2 - \bar{\eta}\right)^2 + 2L_{12}^2\right)}{64 \cdot 20^4 \bar{\rho}^2(\tau-2)^2 L_{12}^4 k} \leq \bar{d}_1.
\end{aligned} \tag{3.52}$$

Combining the previous two inequalities in (3.52) and (3.51), we obtain

$$\|\nabla \tilde{G}_k\|^2 \leq \bar{d}_1(\alpha_k)^2\|x_{k+1} - x_k\|^2 + \frac{2}{\bar{\rho}^2}\|y_{k+1} - y_k\|^2. \tag{3.53}$$

Denote $d_k^{(2)} = \frac{1}{\max\left\{\bar{d}_1\alpha_k, \frac{20}{9\bar{\rho}}\right\}}$. By multiplying $d_k^{(2)}$ on the both sides of (3.53), and using (3.48), we have

$$d_k^{(2)}\|\nabla \tilde{G}_k\|^2 \leq \mathcal{F}_k - \mathcal{F}_{k+1} + \frac{8}{\bar{\rho}}\left(\frac{c_{k-1}}{c_k} - \frac{c_k}{c_{k+1}}\right)\|y_{k+1}\|^2 + \frac{c_{k-1} - c_k}{2}\|y_{k+1}\|^2, \tag{3.54}$$

where the last inequality follows since $d_k^{(2)}$ is a decreasing sequence. Denoting

$$\tilde{T}(\varepsilon) := \min\{k \mid \|\nabla \tilde{G}(x_k, y_k)\| \leq \frac{\varepsilon}{2}, k \geq 2\},$$

Summing both sides of (3.54) from $k = 2$ to $k = \tilde{T}(\varepsilon)$, we then obtain

$$\begin{aligned}
&\sum_{k=2}^{\tilde{T}(\varepsilon)} d_k^{(2)}\|\nabla \tilde{G}_k\|^2 \\
\leq &\mathcal{F}_2 - \mathcal{F}_{\tilde{T}(\varepsilon)} + \frac{8}{\bar{\rho}}\left(\frac{c_1}{c_2} - \frac{c_{\tilde{T}(\varepsilon)}}{c_{\tilde{T}(\varepsilon)+1}}\right)\hat{\sigma}_y^2 + \frac{c_1 - c_{\tilde{T}(\varepsilon)}}{2}\hat{\sigma}_y^2 \\
\leq &\mathcal{F}_2 - \mathcal{F}_{\tilde{T}(\varepsilon)} + \frac{8c_1}{\bar{\rho}c_2}\hat{\sigma}_y^2 + \frac{c_1}{2}\hat{\sigma}_y^2 \\
= &\mathcal{F}_2 - \mathcal{F}_{\tilde{T}(\varepsilon)} + \left(8 \cdot 2^{1/4} + \frac{19}{40}\right)\frac{\hat{\sigma}_y^2}{\bar{\rho}}.
\end{aligned} \tag{3.55}$$

Note that by the definition of $\mathcal{F}_{k+1}$ in Lemma 3.4, we have

$$\begin{aligned}
\mathcal{F}_{\tilde{T}(\varepsilon)} &\geq \underline{f} - \frac{8\hat{\sigma}_y^2}{\bar{\rho}}\left(1 + \frac{1}{\tilde{T}(\varepsilon)}\right)^{1/4} - \frac{15\hat{\sigma}_y^2}{\bar{\rho}} - \frac{c_1\hat{\sigma}_y^2}{2} \\
&= \underline{f} - \left(2^{13/4} + 15 + \frac{19}{40}\right)\frac{\hat{\sigma}_y^2}{\bar{\rho}} = \underline{\mathcal{F}},
\end{aligned}$$

where $\underline{f} := \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x, y)$. We then conclude from (3.55) that

$$\sum_{k=2}^{\tilde{T}(\varepsilon)} d_k^{(2)} \|\nabla \tilde{G}_k\|^2 \le \mathcal{F}_2 - \underline{\mathcal{F}} + \left(8 \cdot 2^{1/4} + \frac{19}{40}\right) \frac{\hat{\sigma}_y^2}{\bar{\rho}} = d_2. \tag{3.56}$$

We can see from the selection of d_3 that $d_3 = \max\{\bar{d}_1, \frac{20}{9\bar{\rho}\alpha_2}\}$. Observe that α_k is an increasing sequence, when $k \ge 2$, we have that $d_3 \ge \max\{\bar{d}_1, \frac{20}{9\bar{\rho}\alpha_k}\}$, which implies that $d_k^{(2)} \ge \frac{1}{d_3 \alpha_k}$, by multiplying d_3 on the both sides of (3.56), and combining the definition of d_2, we have $\sum_{k=2}^{\tilde{T}(\varepsilon)} \frac{1}{\alpha_k} \|\nabla \tilde{G}_k\|^2 \le d_2 d_3$, which, by the definition of $\tilde{T}(\varepsilon)$, implies that

$$\frac{\varepsilon^2}{4} \le \frac{d_2 d_3}{\sum_{k=2}^{\tilde{T}(\varepsilon)} \frac{1}{\alpha_k}}. \tag{3.57}$$

Note that when $c_k = \frac{19}{20\bar{\rho}k^{1/4}}$, $\alpha_k = \frac{2 \cdot 40^2 \bar{\rho}(\tau-2)L_{12}^2\sqrt{k}}{19^2}$. By using the fact $\sum_{k=2}^{\tilde{T}(\varepsilon)} 1/\sqrt{k} \ge \sqrt{\tilde{T}(\varepsilon)} - 1$ and (3.57), we conclude $\frac{\varepsilon^2}{4} \le \frac{2 \cdot 40^2 \bar{\rho}(\tau-2)L_{12}^2 d_2 d_3}{19^2\left(\sqrt{\tilde{T}(\varepsilon)}-1\right)}$ or equivalently,

$$\tilde{T}(\varepsilon) \le \left(\frac{2 \cdot 80^2 \bar{\rho}(\tau-2)L_{12}^2 d_2 d_3}{19^2 \varepsilon^2} + 1\right)^2.$$

On the other hand, if $k \ge \frac{19^4 \hat{\sigma}_y^4}{10^4 \bar{\rho}^4 \varepsilon^4}$, then $c_k = \frac{19}{20\bar{\rho}k^{1/4}} \le \frac{\varepsilon}{2\hat{\sigma}_y}$. This inequality together with the definition of $\hat{\sigma}_y$ then imply that $c_k \|y_k\| \le \frac{\varepsilon}{2}$. Therefore, there exists a

$$\begin{aligned} T(\varepsilon) &\le \max(\tilde{T}(\varepsilon), \frac{19^4 \hat{\sigma}_y^4}{10^4 \bar{\rho}^4 \varepsilon^4}) \\ &\le \max\left(\left(\frac{2 \cdot 80^2 \bar{\rho}(\tau-2)L_{12}^2 d_2 d_3}{19^2 \varepsilon^2} + 1\right)^2, \frac{19^4 \hat{\sigma}_y^4}{10^4 \bar{\rho}^4 \varepsilon^4}\right), \end{aligned}$$

such that $\|\nabla G_k\| \le \|\nabla \tilde{G}_k\| + c_k \|y_k\| \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$. ∎

According to Theorem 3.2, we can show that, by specifying γ_k and c_k as in the order of $k^{1/2}$ and $1/k^{1/4}$, the gradient complexity of Algorithm 1 to obtain an ε-stationarity point of f in the nonconvex-concave setting can be bounded by $\mathcal{O}(L^4\varepsilon^{-4})$. In this setting the stepsize for updating x_k is in the order of $k^{-1/2}$, while the one for updating y_k is a constant at iteration k.

It is worth mentioning two closely related works to ours. One is the HiBSA algorithm proposed in [33] for nonconvex-concave problems. It seems that our iteration complexity is slightly better than that of the HiBSA algorithm by a logarithmic factor. More importantly, the two subproblems for updating x_k and y_k in AGP are much easier than those in the HiBSA algorithm, as the latter method needs to solve $\max_{y \in \mathcal{Y}} f(x, y)$ at each iteration to update y_k under the general nonconvex-concave (not necessarily strongly concave) setting. The other related method is the GDA algorithm. By setting the stepsize to update x_k in the order of ε^4, Lin et al. [31] proved that the iteration complexity of GDA to return an ε-stationary point of $\Phi(\cdot) = \max_{y \in \mathcal{Y}} f(\cdot, y)$ is bounded by $\tilde{\mathcal{O}}(\varepsilon^{-6})$ for nonconvex-concave minimax problems when $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y}$ is a convex compact set. In this paper, both $\mathcal{X}$ and $\mathcal{Y}$ are convex compact sets, hence the outer problem is a constrained minimization problem instead of an unconstrained one. Under this setting, the stopping rule, i.e., $\|\nabla \Phi(\cdot)\| \le \varepsilon$, used in [31] for unconstrained minimization setting needs to be reconsidered. Moreover, the step sizes to update x_k in AGP is a decreasing sequence in the order of $k^{-1/2}$, which is significantly larger than $\mathcal{O}(\varepsilon^4)$ stepsize allowed in GDA, and thus may result in better practical performance.

Note that for solving unconstrained bilinear minimax problem, the alternating GDA algorithm with any fixed step size will cause recurrence [6]. In the classic convex optimization literature, such a divergence issue was usually handled by incorporating averaging, smoothing, or direct acceleration techniques (see, e.g., Sections 3.5-3.8 and Sections 4.3-4.5 of [26]). However, the proposed AGP algorithm for nonconvex minimax problems is not equivalent to the alternating GDA algorithm, since a regularized version of the original function is incorporated, and a variable stepsize policy has been used for the nonconvex-concave setting. Hence, the results of our paper do not contradict with existing ones.

4 Complexity Analysis for Convex-Nonconcave Minimax Problems

In this section, we establish the convergence of AGP algorithm for the cases where f is convex w.r.t. x, but possibly nonconcave w.r.t. y. These are important minimax problems but the studies on their solution methods are still quite limited. Although there exists some symmetry between nonconvex-concave and convex-nonconcave minimax problems, the complexity analysis of the same AGP algorithm for nonconvex-concave setting cannot be trivially extended to that for convex-nonconcave setting.

4.1 Strongly Convex-Nonconcave Setting

In this subsection, we analyze the iteration complexity of Algorithm 1 for solving strongly convex-nonconcave minimax optimization problems (P), i.e., $f(x, y)$ is θ-strongly convex w.r.t. x for any fixed $y \in \mathcal{Y}$, and nonconcave w.r.t. y for any given $x \in \mathcal{X}$. Under this setting, $\forall k \geq 1$, we set

$$\beta_k = \tfrac{1}{\zeta}, \gamma_k = \nu, b_k = c_k = 0, \tag{4.1}$$

in Algorithm 1, and simplify the update for x_k and y_k as follows:

$$x_{k+1} = \mathcal{P}_{\mathcal{X}}\left(x_k - \zeta \nabla_x f(x_k, y_k)\right), \tag{4.2}$$

$$y_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(y_k + \tfrac{1}{\nu}\nabla_y f(x_{k+1}, y_k)\right). \tag{4.3}$$

Our goal in the remaining part of this subsection is to establish the iteration complexity of Algorithm 1 under the strongly convex-nonconcave setting. The convergence analysis for this setting is different from that of the nonconvex-strongly concave setting in Section 3.

Lemma 4.1 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (4.1). If $\nu > L_{22}$, then we have

$$\begin{aligned} & f(x_{k+1}, y_{k+1}) - f(x_k, y_k) \\ \geq & \frac{\nu}{2}\|y_{k+1} - y_k\|^2 - \frac{L_{21}^2\zeta}{2}\|y_k - y_{k-1}\|^2 + (\theta - \frac{1}{2\zeta} - \frac{\zeta L_{11}^2}{2})\|x_k - x_{k-1}\|^2 \\ & + (\frac{\theta}{2} - \frac{1}{\zeta})\|x_{k+1} - x_k\|^2. \end{aligned} \tag{4.4}$$

Proof Similar to (3.5)-(3.7) in the proof of Lemma 3.1, by the optimality condition for x_k in (4.2) implies that $\forall x \in \mathcal{X}$ and $\forall k \geq 1$,

$$\langle \nabla_x f(x_k, y_k) + \frac{1}{\zeta}(x_{k+1} - x_k), x - x_{k+1}\rangle \geq 0, \tag{4.5}$$

$$\langle \nabla_x f(x_k, y_k) + \frac{1}{\zeta}(x_{k+1} - x_k), x_k - x_{k+1}\rangle \geq 0, \tag{4.6}$$

$$\langle \nabla_x f(x_{k-1}, y_{k-1}) + \frac{1}{\zeta}(x_k - x_{k-1}), x_{k+1} - x_k\rangle \geq 0, \tag{4.7}$$

which, in view of the fact that $f(x, y)$ is θ-strongly convex w.r.t. x for any given $y \in \mathcal{Y}$, then implies that

$$\begin{aligned} & f(x_{k+1}, y_k) - f(x_k, y_k) \\ \geq & \langle \nabla_x f(x_k, y_k), x_{k+1} - x_k\rangle + \frac{\theta}{2}\|x_{k+1} - x_k\|^2 \\ \geq & \langle \nabla_x f(x_k, y_k) - \nabla_x f(x_{k-1}, y_{k-1}), x_{k+1} - x_k\rangle \\ & - \frac{1}{\zeta}\langle x_k - x_{k-1}, x_{k+1} - x_k\rangle + \frac{\theta}{2}\|x_{k+1} - x_k\|^2. \end{aligned} \tag{4.8}$$

Denoting $m_{k+1} := (x_{k+1} - x_k) - (x_k - x_{k-1})$, by Assumption 2.1, the Cauchy-Schwarz inequality and the θ-strongly convexity of f w.r.t. x, we can estimate the first inner product term in the r.h.s. of (4.8) as

$$\begin{aligned}
&\langle \nabla_x f(x_k, y_k) - \nabla_x f(x_{k-1}, y_{k-1}), x_{k+1} - x_k \rangle \\
=&\langle \nabla_x f(x_k, y_k) - \nabla_x f\left(x_k, y_{k-1}\right), x_{k+1} - x_k \rangle \\
&+ \langle \nabla_x f\left(x_k, y_{k-1}\right) - \nabla_x f(x_{k-1}, y_{k-1}), m_{k+1} \rangle \\
&+ \langle \nabla_x f\left(x_k, y_{k-1}\right) - \nabla_x f(x_{k-1}, y_{k-1}), x_k - x_{k-1} \rangle \qquad (4.9)\\
\geq& - \frac{L_{21}^2 \zeta}{2} \|y_k - y_{k-1}\|^2 - \frac{1}{2\zeta} \|x_{k+1} - x_k\|^2 - \frac{\zeta L_{11}^2}{2} \|x_k - x_{k-1}\|^2 \\
&- \frac{1}{2\zeta} \|m_{k+1}\|^2 + \theta \|x_k - x_{k-1}\|^2. \qquad (4.10)
\end{aligned}$$

Moreover, it can be easily checked that

$$\langle x_k - x_{k-1}, x_{k+1} - x_k \rangle = \frac{1}{2} \|x_k - x_{k-1}\|^2 + \frac{1}{2} \|x_{k+1} - x_k\|^2 - \frac{1}{2} \|m_{k+1}\|^2. \qquad (4.11)$$

Plugging (4.10) and (4.11) into (4.8) and rearranging the terms, we conclude that

$$\begin{aligned}
f(x_{k+1}, y_k) - f(x_k, y_k) \geq& - \frac{L_{21}^2 \zeta}{2} \|y_k - y_{k-1}\|^2 + (\theta - \frac{1}{2\zeta} - \frac{\zeta L_{11}^2}{2}) \|x_k - x_{k-1}\|^2 \\
&+ (\frac{\theta}{2} - \frac{1}{\zeta}) \|x_{k+1} - x_k\|^2. \qquad (4.12)
\end{aligned}$$

By setting $\gamma_k = \nu$, $c_k = 0$ in (2.7) of Lemma 2.1 and the assumption $\nu > L_{22}$, we have

$$f(x_{k+1}, y_{k+1}) - f(x_{k+1}, y_k) \geq \frac{\nu}{2} \|y_{k+1} - y_k\|^2. \qquad (4.13)$$

The proof is completed by combining (4.12) with (4.13). ■

Next, we further refine this relation in (4.4) as shown below. Note that in Lemma 3.2 we have constructed a potential function for the nonconvex-strongly concave case, whereas in the following lemma, an additional term involving the distance between two adjacent iterates, i.e., $\|x_{k+1} - x_k\|^2$, is added to construct another potential function for the strongly convex-nonconcave case.

Lemma 4.2 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (4.1). Denote

$$\hat{f}_{k+1} := f\left(x_{k+1}, y_{k+1}\right), \ \hat{S}_{k+1} := -\frac{2}{\zeta^2 \theta} \|x_{k+1} - x_k\|^2,$$

$$\hat{F}_{k+1} := \hat{f}_{k+1} + \hat{S}_{k+1} + (\theta + \frac{7}{2\zeta} - \frac{\zeta L_{11}^2}{2} - \frac{2L_{11}^2}{\theta}) \|x_{k+1} - x_k\|^2 - (\frac{L_{21}^2 \zeta}{2} + \frac{2L_{21}^2}{\theta^2 \zeta}) \|y_{k+1} - y_k\|^2.$$

If $\nu > L_{22}$, then $\forall k \geq 1$,

$$\begin{aligned}
\hat{F}_{k+1} - \hat{F}_k \geq& \left(\frac{\nu}{2} - \frac{\zeta L_{21}^2}{2} - \frac{2L_{21}^2}{\zeta \theta^2} \right) \|y_{k+1} - y_k\|^2 \\
&+ \left(\frac{3\theta - \zeta L_{11}^2}{2} + \frac{\theta - 4\zeta L_{11}^2}{2\zeta\theta} \right) \|x_{k+1} - x_k\|^2. \qquad (4.14)
\end{aligned}$$

Proof First by (4.6) and (4.7), we have

$$\frac{1}{\zeta} \langle m_{k+1}, x_k - x_{k+1} \rangle \geq \langle \nabla_x f(x_k, y_k) - \nabla_x f\left(x_{k-1}, y_{k-1}\right), x_{k+1} - x_k \rangle, \qquad (4.15)$$

which together with (4.9) then imply that

$$\begin{aligned}\frac{1}{\zeta}\langle m_{k+1}, x_k - x_{k+1}\rangle \geq & \langle \nabla_x f(x_k, y_k) - \nabla_x f\left(x_k, y_{k-1}\right), x_{k+1} - x_k\rangle \\ & + \langle \nabla_x f\left(x_k, y_{k-1}\right) - \nabla_x f\left(x_{k-1}, y_{k-1}\right), m_{k+1}\rangle \\ & + \langle \nabla_x f\left(x_k, y_{k-1}\right) - \nabla_x f\left(x_{k-1}, y_{k-1}\right), x_k - x_{k-1}\rangle \qquad (4.16) \\ \geq & -\frac{L_{21}^2}{2\theta}\|y_k - y_{k-1}\|^2 - \frac{\theta}{2}\|x_{k+1} - x_k\|^2 - \frac{\zeta L_{11}^2}{2}\|x_k - x_{k-1}\|^2 \\ & - \frac{1}{2\zeta}\|m_{k+1}\|^2 + \theta\|x_k - x_{k-1}\|^2, \qquad (4.17)\end{aligned}$$

where the second inequality is similar to the proof of (4.10) except that for the first term in the r.h.s., we use $\langle \nabla_x f(x_k, y_k) - \nabla_x f\left(x_k, y_{k-1}\right), x_{k+1} - x_k\rangle \geq -\frac{L_{21}^2}{2\theta}\|y_k - y_{k-1}\|^2 - \frac{\theta}{2}\|x_{k+1} - x_k\|^2$. By using the identity $\frac{1}{\zeta}\langle m_{k+1}, x_k - x_{k+1}\rangle = \frac{1}{2\zeta}\|x_k - x_{k-1}\|^2 - \frac{1}{2\zeta}\|x_{k+1} - x_k\|^2 - \frac{1}{2\zeta}\|m_{k+1}\|^2$, we conclude from (4.17) that

$$\begin{aligned}& \frac{1}{2\zeta}\|x_k - x_{k-1}\|^2 - \frac{1}{2\zeta}\|x_{k+1} - x_k\|^2 - \frac{1}{2\zeta}\|m_{k+1}\|^2 \\ \geq & -\frac{L_{21}^2}{2\theta}\|y_k - y_{k-1}\|^2 - \frac{\theta}{2}\|x_{k+1} - x_k\|^2 - \frac{\zeta L_{11}^2}{2}\|x_k - x_{k-1}\|^2 \\ & - \frac{1}{2\zeta}\|m_{k+1}\|^2 + \theta\|x_k - x_{k-1}\|^2. \qquad (4.18)\end{aligned}$$

Rearranging the terms of (4.18), we have

$$\begin{aligned}& \frac{1}{2\zeta}\|x_k - x_{k-1}\|^2 - \frac{1}{2\zeta}\|x_{k+1} - x_k\|^2 \\ \geq & -\frac{L_{21}^2}{2\theta}\|y_k - y_{k-1}\|^2 - \frac{\theta}{2}\|x_{k+1} - x_k\|^2 + \left(\theta - \frac{\zeta L_{11}^2}{2}\right)\|x_k - x_{k-1}\|^2. \qquad (4.19)\end{aligned}$$

Multiplying $\frac{4}{\zeta\theta}$ on both sides of (4.19) and using the definition of $\hat{S}_{k+1}$, we obtain

$$\hat{S}_{k+1} - \hat{S}_k \geq -\frac{2L_{21}^2}{\theta^2\zeta}\|y_k - y_{k-1}\|^2 - \frac{2}{\zeta}\|x_{k+1} - x_k\|^2 + \left(\frac{4}{\zeta} - \frac{2L_{11}^2}{\theta}\right)\|x_k - x_{k-1}\|^2.$$

The proof is completed by (4.4) in Lemma 4.1 and the definition of $\hat{F}_k$. ∎

We are now ready to establish the iteration complexity for the AGP algorithm in the strongly convex-nonconcave setting.

Theorem 4.1 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (4.1). If the relations $\nu > L_{22}, \nu > L_{21}^2\zeta + \frac{4L_{21}^2}{\zeta\theta^2}$, $\zeta \leq \frac{\theta}{4L_{11}^2}$ are satisfied, then $\forall \varepsilon > 0$, it holds that

$$T(\varepsilon) \leq \frac{\overline{F} - \hat{F}_1}{\hat{d}_1\varepsilon^2},$$

where $\hat{d}_1 := \frac{\min\left\{\frac{\nu}{2} - \frac{\zeta L_{21}^2}{2} - \frac{2L_{21}^2}{\zeta\theta^2}, \frac{3\theta - \zeta L_{11}^2}{2} + \frac{\theta - 4\zeta L_{11}^2}{2\zeta\theta}\right\}}{\max\left\{\frac{1}{\zeta^2} + 2L_{12}^2, 2\nu^2\right\}}$ and $\overline{F} = \bar{f} + (\theta + \frac{7}{2\zeta} - \frac{\zeta L_{11}^2}{2} - \frac{2L_{11}^2}{\theta})\sigma_x^2$ with $\overline{f} := \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x, y)$ and $\sigma_x = \max\{\|x_1 - x_2\| \mid \forall x_1, x_2 \in \mathcal{X}\}$.

Proof By (4.2), we immediately obtain

$$\|(\nabla G_k)_x\| \leq \frac{1}{\zeta}\|x_{k+1} - x_k\|. \qquad (4.20)$$

On the other hand, similar to the proof of (3.24), by (4.3) and the triangle inequality, and the nonexpansiveness of the projection operator $\mathcal{P}_{\mathcal{Y}}$, we conclude that

$$\|(\nabla G_k)_y\| \leq \nu\|y_{k+1} - y_k\| + L_{12}\|x_{k+1} - x_k\|. \qquad (4.21)$$

By combining (4.20) and (4.21), and using the Cauchy-Schwarz inequality, we obtain

$$\|\nabla G_k\|^2 \leq 2\nu^2\|y_{k+1}-y_k\|^2 + \left(\frac{1}{\zeta^2}+2L_{12}^2\right)\|x_{k+1}-x_k\|^2. \tag{4.22}$$

Observing that $\hat{d}_1 > 0$. Multiplying both sides of (4.22) by $\hat{d}_1$, and using (4.14) in Lemma 4.2, we have

$$\hat{d}_1\|\nabla G_k\|^2 \leq \hat{F}_{k+1} - \hat{F}_k. \tag{4.23}$$

Summing up the above inequalities from $k=1$ to $k=T(\varepsilon)$, we obtain

$$\sum_{k=1}^{T(\varepsilon)} \hat{d}_1\|\nabla G_k\|^2 \leq \hat{F}_{T(\varepsilon)+1} - \hat{F}_1. \tag{4.24}$$

Note that by the definition of $\hat{F}_{k+1}$ in Lemma 4.2, we have

$$\begin{aligned}\hat{F}_{T(\varepsilon)+1} &\leq \hat{f}_{T(\varepsilon)+1} + (\theta + \frac{7}{2\zeta} - \frac{\zeta L_{11}^2}{2} - \frac{2L_{11}^2}{\theta})\|x_{T(\varepsilon)+1} - x_{T(\varepsilon)}\|^2 \\ &\leq \bar{f} + (\theta + \frac{7}{2\zeta} - \frac{\zeta L_{11}^2}{2} - \frac{2L_{11}^2}{\theta})\sigma_x^2 = \bar{F},\end{aligned}$$

where the last inequality follows from the definitions of $\bar{f}$ and σ_x, and $\theta + \frac{7}{2\zeta} - \frac{\zeta L_{11}^2}{2} - \frac{2L_{11}^2}{\theta} \geq 0$ due to the selection of ζ. We then conclude from (4.24) that $\sum_{k=1}^{T(\varepsilon)} \hat{d}_1\|\nabla G_k\|^2 \leq \bar{F} - \hat{F}_1$ which, in view of the definition of $T(\varepsilon)$, implies that $\varepsilon^2 \leq (\bar{F} - \hat{F}_1)/(T(\varepsilon)\cdot\hat{d}_1)$ or equivalently, $T(\varepsilon) \leq (\bar{F} - \hat{F}_1)/(\hat{d}_1\varepsilon^2)$. ∎

Theorem 4.1 shows that the number of gradient evaluations performed by Algorithm 1 to obtain an ε-stationary point of f is bounded by $\mathcal{O}\left(L^2\varepsilon^{-2}\right)$ under the strongly convex-nonconcave setting. To the best of our knowledge, this is the first theoretical guarantee that has been obtained in the literature for solving this class of minimax problems.

4.2 Complexity Analysis for Convex-Nonconcave Setting

In this subsection, we analyze the iteration complexity of Algorithm 1 applied to the general convex-nonconcave setting, for which $f(x,y)$ is convex w.r.t. x for any fixed $y \in \mathcal{Y}$, and nonconcave w.r.t. y for any given $x \in \mathcal{X}$. Under this setting, $\forall k \geq 1$, we set

$$\beta_k = \tfrac{1}{\zeta}, \gamma_k = \bar{\nu} + \bar{\gamma}_k, b_k = q_k, c_k = 0, \tag{4.25}$$

where $\bar{\gamma}_k$ and q_k are stepsize parameters to be defined later. We need to make the following assumption on the parameters q_k.

Assumption 4.1 $\{q_k\}$ is a nonnegative monotonically decreasing sequence.

By Assumption 2.1 and $\nabla_x f_{k-1}(x,y) = \nabla_x f(x,y) + q_{k-1}x$, we have

$$\|\nabla_x f_{k-1}(x_k,y_{k-1}) - \nabla_x f_{k-1}(x_{k-1},y_{k-1})\| \leq (L_{11}+q_{k-1})\|x_k - x_{k-1}\|. \tag{4.26}$$

Denoting $L_{11}^{'} = L_{11} + q_1$, by Assumption 4.1 and (4.26), we have

$$\|\nabla_x f_{k-1}(x_k,y_{k-1}) - \nabla_x f_{k-1}(x_{k-1},y_{k-1})\| \leq L_{11}^{'}\|x_k - x_{k-1}\}.$$

It then follows from the above inequality and the strong convexity of $f_{k-1}(x,y_{k-1})$ w.r.t. x (Theorem 2.1.12 in [40]) that

$$\begin{aligned}&\langle \nabla_x f_{k-1}(x_k,y_{k-1}) - \nabla_x f_{k-1}(x_{k-1},y_{k-1}), x_k - x_{k-1}\rangle \\ &\geq \frac{1}{L_{11}^{'}+q_{k-1}}\|\nabla_x f_{k-1}(x_k,y_{k-1}) - \nabla_x f_{k-1}(x_{k-1},y_{k-1})\|^2 + \frac{q_{k-1}L_{11}^{'}}{L_{11}^{'}+q_{k-1}}\|x_k - x_{k-1}\|^2.\end{aligned} \tag{4.27}$$

By using the strong convexity of $f_{k-1}(x,y_{k-1})$ w.r.t. x instead of the strong concavity of $f_{k-1}(x_k,y)$ w.r.t. y, this inequality provides a lower bound for the inner product instead of an upper bound shown as in (3.30). This is a key inequality that we will use to establish some important recursions for the AGP algorithm under the convex-nonconcave setting in the following two results.

Similar to Lemma 4.1, we first provide an estimate on the increase of the function values from $f(x_k,y_k)$ to $f(x_{k+1},y_{k+1})$.

Lemma 4.3 Suppose that Assumption 2.1 and 4.1 hold. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (4.25). If $\forall k, \bar{\gamma}_k > L_{22}$ and $\bar{\zeta} \leq \frac{2}{L_{11}^{'}+q_1}$, then

$$\begin{aligned}
&f(x_{k+1}, y_{k+1}) - f(x_k, y_k) \\
\geq &\left(\bar{\nu} + \frac{\bar{\gamma}_k}{2}\right) \|y_{k+1} - y_k\|^2 - \frac{L_{21}^2 \bar{\zeta}}{2} \|y_k - y_{k-1}\|^2 - \frac{1}{\bar{\zeta}} \|x_{k+1} - x_k\|^2 - \frac{1}{2\bar{\zeta}} \|x_k - x_{k-1}\|^2 \\
&- \frac{q_{k-1}}{2} (\|x_{k+1}\|^2 - \|x_k\|^2).
\end{aligned} \tag{4.28}$$

Proof Similar to (4.5)-(4.9) in the proof of Lemma 4.1, by replacing f with f_k, ζ with $\bar{\zeta}$ respectively, and setting $\theta = 0$, we obtain that

$$\begin{aligned}
&f_k(x_{k+1}, y_k) - f_k(x_k, y_k) \\
\geq &\langle \nabla_x f_k(x_k, y_k) - \nabla_x f_{k-1}(x_k, y_{k-1}), x_{k+1} - x_k \rangle + \langle \nabla_x f_{k-1}(x_k, y_{k-1}) - \nabla_x f_{k-1}(x_{k-1}, y_{k-1}), m_{k+1} \rangle \\
&+ \langle \nabla_x f_{k-1}(x_k, y_{k-1}) - \nabla_x f_{k-1}(x_{k-1}, y_{k-1}), x_k - x_{k-1} \rangle - \frac{1}{\bar{\zeta}} \langle x_k - x_{k-1}, x_{k+1} - x_k \rangle.
\end{aligned} \tag{4.29}$$

Next, we prove lower bound for the four terms in the r.h.s. of (4.29) which are different from that in Lemma 4.1. By using the convexity of $f_{k-1}(x, y_{k-1})$ w.r.t x instead of the concavity of $f_{k-1}(x_k, y)$ w.r.t y, the opposite side Cauchy-Schwarz inequality, and replacing $\bar{\rho}, c_k, c_{k-1}, L_{12}, L_{22}^{'}$ by $\bar{\zeta}, q_k, q_{k-1}, L_{21}, L_{11}^{'}$ respectively, Assumptions 2.1 and 4.1, similar to the proof of (3.36)-(3.38) we conclude that

$$\begin{aligned}
&\langle \nabla_x f_k(x_k, y_k) - \nabla_x f_{k-1}(x_k, y_{k-1}), x_{k+1} - x_k \rangle \\
\geq &- \frac{L_{21}^2 \bar{\zeta}}{2} \|y_k - y_{k-1}\|^2 - \frac{1}{2\bar{\zeta}} \|x_{k+1} - x_k\|^2 + \frac{q_k - q_{k-1}}{2} (\|x_{k+1}\|^2 - \|x_k\|^2),
\end{aligned} \tag{4.30}$$

and

$$\begin{aligned}
\langle \nabla_x f_{k-1}(x_k, y_{k-1}) - \nabla_x f_{k-1}(x_{k-1}, y_{k-1}), m_{k+1} \rangle \geq &- \frac{\bar{\zeta}}{2} \|\nabla_x f_{k-1}(x_k, y_{k-1}) - \nabla_x f_{k-1}(x_{k-1}, y_{k-1})\|^2 \\
&- \frac{1}{2\bar{\zeta}} \|m_{k+1}\|^2,
\end{aligned} \tag{4.31}$$

and

$$\begin{aligned}
&\langle \nabla_x f_{k-1}(x_k, y_{k-1}) - \nabla_x f_{k-1}(x_{k-1}, y_{k-1}), x_k - x_{k-1} \rangle \\
\geq &\frac{1}{L_{11}^{'} + q_{k-1}} \|\nabla_x f_{k-1}(x_k, y_{k-1}) - \nabla_x f_{k-1}(x_{k-1}, y_{k-1})\|^2.
\end{aligned} \tag{4.32}$$

Moreover, it can be easily checked that

$$\frac{1}{\bar{\zeta}} \langle x_{k+1} - x_k, x_k - x_{k-1} \rangle = \frac{1}{2\bar{\zeta}} \|x_{k+1} - x_k\|^2 + \frac{1}{2\bar{\zeta}} \|x_k - x_{k-1}\|^2 - \frac{1}{2\bar{\zeta}} \|m_{k+1}\|^2. \tag{4.33}$$

Plugging (4.30)-(4.33) into (4.29), and by the definition of $f_k(x_{k+1}, y_k)$ and $f_k(x_k, y_k)$ and $\frac{\bar{\zeta}}{2} \leq \frac{1}{L_{11}^{'}+q_1}$, we conclude that

$$f(x_{k+1}, y_k) - f(x_k, y_k) \geq - \frac{L_{21}^2 \bar{\zeta}}{2} \|y_k - y_{k-1}\|^2 - \frac{1}{\bar{\zeta}} \|x_{k+1} - x_k\|^2 - \frac{1}{2\bar{\zeta}} \|x_k - x_{k-1}\|^2 - \frac{q_{k-1}}{2} (\|x_{k+1}\|^2 - \|x_k\|^2). \tag{4.34}$$

By setting $\gamma_k = \bar{\nu} + \bar{\gamma}_k$, $c_k = 0$ in (2.7) of Lemma 2.1 and the assumption $\bar{\gamma}_k > L_{22}$, we have

$$f(x_{k+1}, y_{k+1}) - f(x_{k+1}, y_k) \geq \left(\bar{\nu} + \frac{\bar{\gamma}_k}{2}\right) \|y_{k+1} - y_k\|^2. \tag{4.35}$$

The proof is completed by combing (4.34) and (4.35). ∎

We need to further refine the relation in (4.28) in order to establish the convergence of the AGP algorithm as shown below. Note that in Lemma 4.2 we have constructed a potential function which involves the function value plus some distance between two adjacent iterates for the strongly convex-nonconcave case, whereas in the following lemma an additional quadratic regularization term, i.e., $\|x_{k+1}\|^2$, is added to construct another potential function for the general convex-nonconcave case.

Lemma 4.4 Suppose that Assumptions 2.1 and 4.1 hold. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (4.25). Denote

$$\begin{aligned}\hat{\mathcal{S}}_{k+1} &:= -\frac{8}{\bar{\zeta}^2 q_{k+1}}\|x_{k+1}-x_k\|^2 - \frac{8}{\bar{\zeta}}\left(1-\frac{q_k}{q_{k+1}}\right)\|x_{k+1}\|^2,\\ \hat{\mathcal{F}}_{k+1} &:= f(x_{k+1}, y_{k+1}) + \hat{\mathcal{S}}_{k+1} + \frac{15}{2\bar{\zeta}}\|x_{k+1}-x_k\|^2 + \frac{q_k}{2}\|x_{k+1}\|^2\\ &\quad - \left(\frac{\bar{\zeta}L_{21}^2}{2} + \frac{16L_{21}^2}{\bar{\zeta}(q_{k+1})^2}\right)\|y_{k+1}-y_k\|^2.\end{aligned}$$

If

$$\bar{\gamma}_k > L_{22},\ \frac{1}{q_{k+1}} - \frac{1}{q_k} \le \frac{\bar{\zeta}}{5},\ \bar{\zeta} \le \frac{2}{L_{11}^{'} + q_1}, \tag{4.36}$$

then $\forall k \ge 1$,

$$\begin{aligned}&\hat{\mathcal{F}}_{k+1} - \hat{\mathcal{F}}_k\\ \ge &\left(\bar{\nu} + \frac{\bar{\gamma}_k}{2} - \frac{\bar{\zeta}L_{21}^2}{2} - \frac{16L_{21}^2}{\bar{\zeta}(q_{k+1})^2}\right)\|y_{k+1}-y_k\|^2 + \frac{q_k - q_{k-1}}{2}\|x_{k+1}\|^2\\ &+ \frac{9}{10\bar{\zeta}}\|x_{k+1}-x_k\|^2 + \frac{8}{\bar{\zeta}}\left(\frac{q_k}{q_{k+1}} - \frac{q_{k-1}}{q_k}\right)\|x_{k+1}\|^2.\end{aligned}$$

Proof Similar to the proof of (4.16), by replacing f with f_k, we have

$$\begin{aligned}\frac{1}{\bar{\zeta}}\langle m_{k+1}, x_k - x_{k+1}\rangle \ge &\langle \nabla_x f_k(x_k, y_k) - \nabla_x f_{k-1}(x_k, y_{k-1}), x_{k+1} - x_k\rangle\\ &+ \langle \nabla_x f_{k-1}(x_k, y_{k-1}) - \nabla_x f_{k-1}(x_{k-1}, y_{k-1}), m_{k+1}\rangle\\ &+ \langle \nabla_x f_{k-1}(x_k, y_{k-1}) - \nabla_x f_{k-1}(x_{k-1}, y_{k-1}), x_k - x_{k-1}\rangle.\end{aligned}$$

Using an argument similar to the proof of (4.30)-(4.33), and the Cauchy-Schwarz inequality, we conclude from the above inequality that

$$\begin{aligned}&\frac{1}{2\bar{\zeta}}\|x_k - x_{k-1}\|^2 - \frac{1}{2\bar{\zeta}}\|x_{k+1} - x_k\|^2 - \frac{1}{2\bar{\zeta}}\|m_{k+1}\|^2\\ \ge &-\frac{L_{21}^2}{2\bar{a}_k}\|y_k - y_{k-1}\|^2 - \frac{\bar{a}_k}{2}\|x_{k+1} - x_k\|^2 + \frac{q_k - q_{k-1}}{2}(\|x_{k+1}\|^2 - \|x_k\|^2)\\ &- \frac{q_k - q_{k-1}}{2}\|x_{k+1} - x_k\|^2 - \frac{\bar{\zeta}}{2}\|\nabla_x f_{k-1}(x_k, y_{k-1}) - \nabla_x f_{k-1}(x_{k-1}, y_{k-1})\|^2\\ &+ \frac{1}{L_{11}^{'} + q_{k-1}}\|\nabla_x f_{k-1}(x_k, y_{k-1}) - \nabla_x f_{k-1}(x_{k-1}, y_{k-1})\|^2 - \frac{1}{2\bar{\zeta}}\|m_{k+1}\|^2\\ &+ \frac{q_{k-1}L_{11}^{'}}{L_{11}^{'} + q_{k-1}}\|x_k - x_{k-1}\|^2,\end{aligned} \tag{4.37}$$

where $\bar{a}_k > 0$. Observing that $q_1 \le L_{11}^{'}$ and by Assumption 4.1, we have

$$\frac{q_{k-1}L_{11}^{'}}{q_{k-1} + L_{11}^{'}} \ge \frac{q_{k-1}L_{11}^{'}}{2L_{11}^{'}} = \frac{q_{k-1}}{2} \ge \frac{q_k}{2}.$$

By $\bar{\zeta} \le \frac{2}{L_{11}^{'} + q_1}$ and Assumption 4.1, rearranging the terms in (4.37), we obtain

$$\begin{aligned}&-\frac{1}{2\bar{\zeta}}\|x_{k+1} - x_k\|^2 - \frac{q_k - q_{k-1}}{2}\|x_{k+1}\|^2\\ \ge &-\frac{1}{2\bar{\zeta}}\|x_k - x_{k-1}\|^2 - \frac{q_k - q_{k-1}}{2}\|x_k\|^2 - \frac{L_{21}^2}{2\bar{a}_k}\|y_k - y_{k-1}\|^2\\ &- \frac{\bar{a}_k}{2}\|x_{k+1} - x_k\|^2 + \frac{q_k}{2}\|x_k - x_{k-1}\|^2.\end{aligned}$$

By multiplying $\frac{16}{\zeta q_k}$ on both sides of the above inequality, we then obtain

$$\begin{aligned}
&-\frac{8}{\bar{\zeta}^2 q_k}\|x_{k+1}-x_k\|^2-\frac{8}{\bar{\zeta}}\left(1-\frac{q_{k-1}}{q_k}\right)\|x_{k+1}\|^2\\
\geq&-\frac{8}{\bar{\zeta}^2 q_k}\|x_k-x_{k-1}\|^2-\frac{8}{\bar{\zeta}}(1-\frac{q_{k-1}}{q_k})\|x_k\|^2-\frac{8L_{21}^2}{\bar{\zeta}q_k\bar{a}_k}\|y_k-y_{k-1}\|^2\\
&-\frac{8\bar{a}_k}{\bar{\zeta}q_k}\|x_{k+1}-x_k\|^2+\frac{8}{\bar{\zeta}}\|x_k-x_{k-1}\|^2. \qquad (4.38)
\end{aligned}$$

Setting $\bar{a}_k=\frac{q_k}{2}$ in the above inequality, and using the definition of $\hat{\mathcal{S}}_{k+1}$ and (4.36), we have

$$\begin{aligned}
\hat{\mathcal{S}}_{k+1}-\hat{\mathcal{S}}_k\geq&\frac{8}{\bar{\zeta}}\left(\frac{q_k}{q_{k+1}}-\frac{q_{k-1}}{q_k}\right)\|x_{k+1}\|^2-\frac{16L_{21}^2}{\bar{\zeta}(q_k)^2}\|y_k-y_{k-1}\|^2\\
&-\frac{28}{5\bar{\zeta}}\|x_{k+1}-x_k\|^2+\frac{8}{\bar{\zeta}}\|x_k-x_{k-1}\|^2. \qquad (4.39)
\end{aligned}$$

The proof is completed by combining (4.39) and (4.28) in Lemma 4.3, and using the definition of $\hat{\mathcal{F}}_{k+1}$. ■

Note that the proof of Lemma 4.4 is different from that of Lemma 3.4, since different potential functions, i.e., $\hat{\mathcal{F}}_{k+1}$ and $\mathcal{F}_{k+1}$ respectively, are used to establish the convergence of the proposed algorithms. The construction of these potential functions is the key step for our convergence analysis of AGP algorithms. We are now ready to establish the iteration complexity for the AGP algorithm to achieve an ε-stationary point for solving (P) under general convex-nonconcave setting.

Theorem 4.2 Suppose that Assumptions 2.1 holds. Let $\{(x_k,y_k)\}$ be a sequence generated by Algorithm 1 with parameter settings in (4.25). If $\bar{\zeta}\leq\frac{1}{10L_{11}}$, $\tau>\max\{\frac{19^2(L_{22}+2\bar{\nu}-\bar{\zeta}L_{21}^2)}{20^2\cdot 16\bar{\zeta}L_{21}^2},2\}$, $q_k=\frac{19}{20\bar{\zeta}k^{1/4}}, k\geq 1$, then for any given $\varepsilon>0$,

$$T(\varepsilon)\leq\max\left(\left(\frac{2\cdot 80^2\bar{\zeta}(\tau-2)L_{21}^2\hat{d}_3\hat{d}_4}{19^2\varepsilon^2}+2\right)^2,\frac{19^4\hat{\sigma}_x^4}{10^4\bar{\zeta}^4\varepsilon^4}\right),$$

where $\hat{\sigma}_x:=\max\{\|x\|\mid x\in\mathcal{X}\}$,

$$\hat{d}_3:=\hat{\mathcal{F}}_0-\hat{\mathcal{F}}_2+\frac{8\cdot 2^{1/4}}{\bar{\zeta}}\hat{\sigma}_x^2+\frac{19}{40\bar{\zeta}}\hat{\sigma}_x^2,$$

$\hat{d}_4:=\max\{\hat{D}_1,\frac{10+20\bar{\zeta}^2L_{21}^2}{9\bar{\zeta}p_2}\}$ with $\hat{\mathcal{D}}_1:=\frac{16\tau^2}{(\tau-2)^2}+\frac{19^4(\bar{\zeta}L_{21}^2-\bar{\nu})^2}{16\cdot 20^4(\tau-2)^2L_{21}^4\bar{\zeta}^2}$, $\hat{\mathcal{F}}_0:=\bar{f}+\left(2^{13/4}+15+\frac{19}{40}\right)\frac{\hat{\sigma}_x^2}{\bar{\zeta}}$ with $\bar{f}:=\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}}f(x,y)$.

Proof By $\bar{\zeta}\leq\frac{1}{10L_{11}}$,$q_k=\frac{19}{20\bar{\zeta}k^{1/4}},\forall k\geq 1$, let us denote $\bar{\gamma}_k=\bar{\zeta}L_{21}^2+\frac{16\tau L_{21}^2}{\bar{\zeta}(q_{k+1})^2}-2\bar{\nu}$, $p_k=\frac{8(\tau-2)L_{21}^2}{\bar{\zeta}(q_{k+1})^2}$, it can be easily checked that the relations in (4.36) are satisfied. It follows from the selection of $\bar{\gamma}_k$ and p_k that

$$\bar{\nu}+\frac{\bar{\gamma}_k}{2}-\frac{\bar{\zeta}L_{21}^2}{2}-\frac{16L_{21}^2}{\bar{\zeta}(q_{k+1})^2}=p_k.$$

This observation, in view of Lemma 4.4, then immediately implies that

$$\begin{aligned}
p_k\|y_{k+1}-y_k\|^2+\frac{9}{10\bar{\zeta}}\|x_{k+1}-x_k\|^2\leq&\hat{\mathcal{F}}_{k+1}-\hat{\mathcal{F}}_k+\frac{8}{\bar{\zeta}}\left(\frac{q_{k-1}}{q_k}-\frac{q_k}{q_{k+1}}\right)\|x_{k+1}\|^2\\
&+\frac{q_{k-1}-q_k}{2}\|x_{k+1}\|^2. \qquad (4.40)
\end{aligned}$$

We can easily check from the definition of $f_k(x_k,y_k)$ that

$$\|\nabla G_k\|-\|\nabla\tilde{G}_k\|\leq q_k\|x_k\|.$$

Similar to (4.22) in the proof of Theorem 4.1, by replacing f with f_k, ∇G_k with $\nabla\tilde{G}_k$, ζ with $\bar{\zeta}$, ν with $\bar{\gamma}_k+\bar{\nu}$ respectively, we conclude that

$$\|\nabla\tilde{G}_k\|^2\leq 2\left(\bar{\gamma}_k+\bar{\nu}\right)^2\|y_{k+1}-y_k\|^2+\left(\frac{1}{\bar{\zeta}^2}+2L_{12}^2\right)\|x_{k+1}-x_k\|^2. \qquad (4.41)$$

Since both p_k and $\bar{\gamma}_k$ are in the same order when k becomes large enough, it then follows from the definition of $\hat{D}_1$ that $\forall k \geq 1$,

$$\hat{D}_1 \geq \frac{2\left(\bar{\gamma}_k + \bar{\nu}\right)^2}{(p_k)^2}. \tag{4.42}$$

Combining the previous two inequalities in (4.42) and (4.41), we obtain

$$\|\nabla \tilde{G}_k\|^2 \leq \hat{D}_1 (p_k)^2 \|y_{k+1} - y_k\|^2 + \left(\frac{1}{\bar{\zeta}^2} + 2L_{12}^2\right) \|x_{k+1} - x_k\|^2. \tag{4.43}$$

Denote $\hat{d}_k^{(2)} = \frac{1}{\max\left\{\hat{D}_1 p_k, \frac{10+20\bar{\zeta}^2 L_{12}^2}{9\bar{\zeta}}\right\}}$. By multiplying $\hat{d}_k^{(2)}$ on the both sides of (4.43), and using (4.40), we get

$$\hat{d}_k^{(2)} \|\nabla \tilde{G}_k\|^2 \leq \hat{\mathcal{F}}_{k+1} - \hat{\mathcal{F}}_k + \frac{8}{\bar{\zeta}} \left(\frac{q_{k-1}}{q_k} - \frac{q_k}{q_{k+1}}\right) \|x_{k+1}\|^2 + \frac{q_{k-1} - q_k}{2} \|x_{k+1}\|^2. \tag{4.44}$$

Denote $\hat{T}(\varepsilon) := \min\{k \mid \|\nabla \tilde{G}(x_k, y_k)\| \leq \frac{\varepsilon}{2}, k \geq 2\}$. By summing both sides of (4.44) from $k = 2$ to $k = \hat{T}(\varepsilon)$, we then obtain

$$\begin{aligned}
&\sum_{k=2}^{\hat{T}(\varepsilon)} \hat{d}_k^{(2)} \|\nabla \tilde{G}_k\|^2 \\
\leq& \hat{\mathcal{F}}_{\hat{T}(\varepsilon)+1} - \hat{\mathcal{F}}_2 + \frac{8}{\bar{\zeta}} \left(\frac{q_1}{q_2} - \frac{q_{\hat{T}(\varepsilon)}}{q_{\hat{T}(\varepsilon)+1}}\right) \hat{\sigma}_x^2 + \frac{q_1 - q_{\hat{T}(\varepsilon)}}{2} \hat{\sigma}_x^2 \\
\leq& \hat{\mathcal{F}}_{\hat{T}(\varepsilon)+1} - \hat{\mathcal{F}}_2 + \frac{8q_1}{\bar{\zeta} q_2} \hat{\sigma}_x^2 + \frac{q_1}{2} \hat{\sigma}_x^2 \\
=& \hat{\mathcal{F}}_{\hat{T}(\varepsilon)+1} - \hat{\mathcal{F}}_2 + \frac{8 \cdot 2^{1/4}}{\bar{\zeta}} \hat{\sigma}_x^2 + \frac{19}{40\bar{\zeta}} \hat{\sigma}_x^2.
\end{aligned} \tag{4.45}$$

Note that by the definition of $\hat{\mathcal{F}}_{k+1}$ in Lemma 4.4, we have

$$\begin{aligned}
\hat{\mathcal{F}}_{\hat{T}(\varepsilon)+1} &\leq \bar{f} + \frac{8\hat{\sigma}_x^2}{\bar{\zeta}} \left(1 + \frac{1}{\hat{T}(\varepsilon)}\right)^{1/4} + \frac{15\hat{\sigma}_x^2}{\bar{\zeta}} + \frac{q_1 \hat{\sigma}_x^2}{2} \\
&= \bar{f} + \left(2^{13/4} + 15 + \frac{19}{40}\right) \frac{\hat{\sigma}_x^2}{\bar{\zeta}} = \hat{\mathcal{F}}_0,
\end{aligned}$$

where $\bar{f} := \max_{(x,y) \in \mathcal{X} \times \mathcal{Y}} f(x, y)$. We then conclude from (4.45) that

$$\sum_{k=2}^{\hat{T}(\varepsilon)} \hat{d}_k^{(2)} \|\nabla \tilde{G}_k\|^2 \leq \hat{\mathcal{F}}_0 - \hat{\mathcal{F}}_2 + \frac{8 \cdot 2^{1/4}}{\bar{\zeta}} \hat{\sigma}_x^2 + \frac{19}{40\bar{\zeta}} \hat{\sigma}_x^2 = \hat{d}_3. \tag{4.46}$$

Note that $\hat{d}_4 = \max\{\hat{D}_1, \frac{10+20\bar{\zeta}^2 L_{12}^2}{9\bar{\zeta} p_2}\}$. Since p_k is an increasing sequence when $k \geq 2$, we have that $\hat{d}_4 \geq \max\{\hat{D}_1, \frac{10+20\bar{\zeta}^2 L_{12}^2}{9\bar{\zeta} p_k}\}$, which implies that $\hat{d}_k^{(2)} \geq \frac{1}{\hat{d}_4 p_k}$. By multiplying $\hat{d}_4$ on the both sides of (4.46), and using the definition of $\hat{d}_3$, we have $\sum_{k=2}^{\hat{T}(\varepsilon)} \frac{1}{p_k} \|\nabla \tilde{G}_k\|^2 \leq \hat{d}_3 \hat{d}_4$, which, by the definition of $\hat{T}(\varepsilon)$, implies that

$$\frac{\varepsilon^2}{4} \leq \frac{\hat{d}_3 \hat{d}_4}{\sum_{k=2}^{\hat{T}(\varepsilon)} \frac{1}{p_k}}. \tag{4.47}$$

Using the assumptions $q_k = \frac{19}{20\zeta k^{1/4}}$ and $p_k = \frac{2 \cdot 40^2 \bar{\zeta}(\tau-2) L_{21}^2 \sqrt{k+1}}{19^2}$, (4.47) and the fact $\sum_{k=2}^{\hat{T}(\varepsilon)} 1/\sqrt{k+1} \geq \sqrt{\hat{T}(\varepsilon)} - 2$, we conclude that $\frac{\varepsilon^2}{4} \leq \frac{2 \cdot 40^2 \bar{\zeta}(\tau-2) L_{21}^2 \hat{d}_3 \hat{d}_4}{19^2 \left(\sqrt{\hat{T}(\varepsilon)} - 2\right)}$ or equivalently,

$$\hat{T}(\varepsilon) \leq \left(\frac{2 \cdot 80^2 \bar{\zeta}(\tau-2) L_{21}^2 \hat{d}_3 \hat{d}_4}{19^2 \varepsilon^2} + 2\right)^2.$$

On the other hand, if $k > \frac{19^4\hat{\sigma}_x^4}{10^4\bar{\zeta}^4\varepsilon^4}$, then $q_k = \frac{19}{20\zeta k^{1/4}} \leq \frac{\varepsilon}{2\hat{\sigma}_x}$, this inequality together with the definition of $\hat{\sigma}_x$ then imply that $q_k\|x_k\| \leq \frac{\varepsilon}{2}$. Therefore, there exists a

$$T(\varepsilon) \leq \max(\hat{T}(\varepsilon), \frac{19^4\hat{\sigma}_x^4}{10^4\bar{\zeta}^4\varepsilon^4}) \leq \max\left(\left(\frac{2\cdot 80^2\bar{\zeta}(\tau-2)L_{21}^2\hat{d}_3\hat{d}_4}{19^2\varepsilon^2}+2\right)^2, \frac{19^4\hat{\sigma}_x^4}{10^4\bar{\zeta}^4\varepsilon^4}\right),$$

such that $\|\nabla G_k\| \leq \|\nabla\tilde{G}_k\| + q_k\|x_k\| \leq \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon$. ∎

Theorem 4.2 shows that the number of gradient evaluations performed by Algorithm 1 to obtain an ε-stationary point of f which satisfies (4.36) is bounded by $\mathcal{O}\left(L^4\varepsilon^{-4}\right)$ under the general convex-nonconcave setting. To the best of our knowledge, this is the first theoretical guarantee that has been obtained in the literature for solving this class of minimax problems. As mentioned in Section 1, it is difficult to extend existing algorithms, especially those nested-loop methods, for minimax optimization to the general convex-nonconcave setting. One possible approach would be to switch the order of the "min" and "max" operators. However, in general, $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y) \neq \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} f(x,y)$ when $f(x,y)$ is nonconvex w.r.t. x or nonconcave w.r.t. y. Even if the set of stationary points for the above two problems are the same, e.g., by using the same stationarity criterion as in this paper, the algorithm applied to these problems will converge to different solutions. This can be seen by applying the unified AGP algorithm to solve $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y)$ and $\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} f(x,y)$. The AGP algorithm will likely converge to different stationary points because different potential functions (i.e., $\mathcal{F}_{k+1}$ and $\hat{\mathcal{F}}_{k+1}$ in Lemma 3.6 and Lemma 4.6) have to be applied for analyzing the convergence of AGP for solving these two problems.

For nonconvex-strongly concave or strongly convex-nonconcave setting, the AGP algorithm with properly chosen parameters is actually equivalent to the alternating GDA algorithm. By constructing suitable potential function, i.e., F_{k+1}, we were able to show the $O(\epsilon^{-2})$ iteration complexity of this method. Whereas for nonconvex-concave or convex-nonconcave setting, where b_k or c_k is not equal to 0, AGP algorithm is completely new and the selection of these parameters b_k or c_k plays a very crucial role to guarantee the convergence of the APG method. In these cases, the key step in our proof is also to construct a suitable potential function, i.e., $\mathcal{F}_{k+1}$ and $\hat{\mathcal{F}}_{k+1}$ with b_k or c_k playing a very crucial role (see Lemma 3.6 and Lemma 4.6 respectively). To our best knowledge, this is the first time that such potential functions have been constructed for solving minimax problems.

5 Block-wise Nonsmooth Nonconvex Minimax Problems

In this section, we consider a more general block-wise nonsmooth minimax problem as follows.

$$\begin{aligned}\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\ & l(x,y) := f(x^{(1)},x^{(2)},\cdots,x^{(K_1)},y^{(1)},y^{(2)},\cdots,y^{(K_2)}) + \sum_{i=1}^{K_1} h_i(x^{(i)}) - \sum_{j=1}^{K_2} g_j(y^{(j)}), \\ \text{s.t.}\quad & x^{(i)} \in \mathcal{X}_i, \quad i=1,\cdots,K_1, \\ & y^{(j)} \in \mathcal{Y}_j, \quad j=1,\cdots,K_2,\end{aligned} \tag{BP}$$

where $f:\mathbb{R}^{d_xK_1+d_yK_2}\to\mathbb{R}$ is a continuously differentiable function, $h_i:\mathbb{R}^{d_x}\to\mathbb{R}$ and $g_j:\mathbb{R}^{d_y}\to\mathbb{R}$ are some convex continuous possibly nonsmooth functions, $x = \left[x^{(1)},x^{(2)},\cdots,x^{(K_1)}\right] \in \mathcal{X} = \mathcal{X}_1\times\mathcal{X}_2\times\cdots\times\mathcal{X}_{K_1} \subset \mathbb{R}^{d_xK_1}$ and $y = \left[y^{(1)},y^{(2)},\cdots,y^{(K_2)}\right] \in \mathcal{Y} = \mathcal{Y}_1\times\mathcal{Y}_2\times\cdots\times\mathcal{Y}_{K_2} \subset \mathbb{R}^{d_yK_2}$ are the block variables, $\mathcal{X}_i (i=1,\cdots,K_1)$ and $\mathcal{Y}_j$ $(j=1,\cdots,K_2)$ are nonempty compact convex sets. For notational simplicity, we denote $f\left(x^{(1)},x^{(2)},\cdots,x^{(K_1)},y^{(1)},y^{(2)},\cdots,y^{(K_2)}\right)$ by $f(x,y)$.

Note that (BP) reduces to (P) if we set $K_1 = K_2 = 1$, $h_i(\cdot) = 0$ and $g_j(\cdot) = 0$, which means that (P) is a special case of (BP). There exist very few known existing nested-loop algorithms that are proposed to solve (BP) with multi-block structure. However, the minimax problems with block structure are important in machine learning and signal processing, e.g., distributed training [33].

The difficulty to solve (BP) comes from two aspects. One is that for any given j, $f(x,y^{(1)},y^{(2)},\cdots,y^{(K_2)})$ is nonconcave w.r.t. $y^{(j)}$ for any given x and other blocks of y, i.e., $y^{(1)},y^{(2)},\cdots,y^{(j-1)},y^{(j+1)},\cdots,y^{(K_2-1)}$ and $y^{(K_2)}$. For any given x, to solve the inner maximization subproblems with respect to $y^{(j)}$ is already NP-hard. Due to this reason, almost all the existing nested-loop algorithms will lose their theoretic guarantees since they need to solve the inner subproblem exactly, or approximately with an error proportional to the accuracy ε, unless we

can exchange the order of "min" and "max" operators. However, in general, $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} l(x,y) \neq \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} l(x,y)$ when $f(x,y)$ is nonconvex w.r.t. x or nonconcave w.r.t. y. Although $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} l(x,y)$ and $\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} l(x,y)$ share the same stationary point set, when we use the same algorithm to solve them, it may converge to different stationary points. Another difficulty comes from the multi-block structure. To the best of our knowledge, there are very few nested-loop algorithms (and complexity results) for solving the aforementioned multi-block structure nonconvex minimax problems before our work. On the other hand, single loop algorithms do not need to solve the inner subproblem, and can be easily generalized to handle the multi-block structure.

Similar to the idea of AGP algorithm, we propose a block alternating proximal gradient algorithm (BAPG) to solve (BP). Instead of the original function $l(x,y)$, the BAPG algorithm uses the gradient of a regularized version of the original function, i.e.,

$$\bar{l}_k(x,y) := \bar{f}_k(x,y) + \sum_{i=1}^{K_1} h_i(x^{(i)}) - \sum_{j=1}^{K_2} g_j(y^{(j)}), \tag{5.1}$$

where $\bar{f}_k(x,y) = f(x,y) + \frac{\hat{b}_k}{2}\sum_{i=1}^{K_1}\|x^{(i)}\|^2 - \frac{\hat{c}_k}{2}\sum_{j=1}^{K_2}\|y^{(j)}\|^2$ with regularization parameters $\hat{b}_k \geq 0$ and $\hat{c}_k \geq 0$ at the kth iteration. Before presenting the detailed algorithm, we give some notations as follows. We denote two proximity operators for $x^{(i)}$ and $y^{(j)}$ as follows,

$$\mathrm{Prox}_{h_i,\mathcal{X}_i}^{\beta^{(i)}}(\upsilon^{(i)}) := \arg\min_{x^{(i)}\in\mathcal{X}_i} h_i(x^{(i)}) + \frac{\beta^{(i)}}{2}\|x^{(i)} - \upsilon^{(i)}\|^2, \tag{5.2}$$

$$\mathrm{Prox}_{g_j,\mathcal{Y}_j}^{\xi^{(j)}}(\omega^{(j)}) := \arg\max_{y^{(j)}\in\mathcal{Y}_j} -g_j(y^{(j)}) - \frac{\xi^{(j)}}{2}\|y^{(j)} - \omega^{(j)}\|^2. \tag{5.3}$$

Let k be the number of iteration. Denote $x_k = \left[x_k^{(1)}, x_k^{(2)}, \cdots, x_k^{(K_1)}\right]$, $y_k = \left[y_k^{(1)}, y_k^{(2)}, \cdots, y_k^{(K_2)}\right]$ and define

$$\begin{aligned}
v_{k+1}^{(i)} &:= \left[x_{k+1}^{(1)}, x_{k+1}^{(2)}, \cdots, x_{k+1}^{(i-1)}, x_k^{(i)}, \cdots x_k^{(K_1)}\right],\\
v_{k+1}^{(-i)} &:= \left[x_{k+1}^{(1)}, x_{k+1}^{(2)}, \cdots, x_{k+1}^{(i-1)}, x_k^{(i+1)}, \cdots x_k^{(K_1)}\right],\\
w_{k+1}^{(j)} &:= \left[y_{k+1}^{(1)}, y_{k+1}^{(2)}, \cdots, y_{k+1}^{(j-1)}, y_k^{(j)}, \cdots y_k^{(K_2)}\right],\\
w_{k+1}^{(-j)} &:= \left[y_{k+1}^{(1)}, y_{k+1}^{(2)}, \cdots, y_{k+1}^{(j-1)}, y_k^{(j+1)}, \cdots y_k^{(K_2)}\right].
\end{aligned}$$

Each iteration of the proposed BAPG algorithm conducts two proximal gradient steps for updating both x and y. More specifically, at the kth iteration, it updates x_k by minimizing a linearized approximation of $\bar{l}_k(x,y_k)$ with the gradient at point $\left(v_{k+1}^{(i)}, y_k\right)$, i.e., for each $i = 1, \cdots, K_1$,

$$\begin{aligned}
x_{k+1}^{(i)} &= \arg\min_{x^{(i)}\in\mathcal{X}_i} \left\langle \nabla_{x^{(i)}}\bar{f}_k(v_{k+1}^{(i)}, y_k), x^{(i)} - x_k^{(i)}\right\rangle + h_i(x^{(i)}) + \frac{\beta_k^{(i)}}{2}\|x^{(i)} - x_k^{(i)}\|^2\\
&= \mathrm{Prox}_{h_i,\mathcal{X}_i}^{\beta_k^{(i)}}\left(x_k^{(i)} - \frac{1}{\beta_k^{(i)}}\nabla_{x^{(i)}} f(v_{k+1}^{(i)}, y_k) - \frac{1}{\beta_k^{(i)}}\hat{b}_k x_k^{(i)}\right),
\end{aligned}$$

where $\mathrm{Prox}_{h_i,\mathcal{X}_i}^{\beta_k^{(i)}}$ is the proximal operator which is defined in (5.2) and $\beta_k^{(i)} > 0$. Similarly, it updates y_k by maximizing a linearized approximation of $\bar{l}_k(x_{k+1},y)$ minus some regularized terms, i.e.,

$$\begin{aligned}
y_{k+1}^{(j)} &= \arg\max_{y^{(j)}\in\mathcal{Y}_j} \langle \nabla_{y^{(j)}}\bar{f}_k(x_{k+1}, w_{k+1}^{(j)}), y^{(j)} - y_k^{(j)}\rangle - g_j(y^{(j)}) - \frac{\xi_k^{(j)}}{2}\|y^{(j)} - y_k^{(j)}\|^2\\
&= \mathrm{Prox}_{g_j,\mathcal{Y}_j}^{\xi_k^{(j)}}\left(y_k^{(j)} + \frac{1}{\xi_k^{(j)}}\nabla_{y^{(j)}} f(x_{k+1}, w_{k+1}^{(j)}) - \frac{1}{\xi_k^{(j)}}\hat{c}_k y_k^{(j)}\right),
\end{aligned}$$

where $\mathrm{Prox}_{g_j,\mathcal{Y}_j}^{\xi_k^{(j)}}$ is the proximal operator which is defined in (5.3) and $\xi_k^{(j)} > 0$. The proposed BAPG algorithm is formally stated in Algorithm 2, where sequences $\beta_k^{(i)}$, $\hat{b}_k$, $\xi_k^{(j)}$, $\hat{c}_k$ and the stopping rule in Step 4 will be

Algorithm 2 (BAPG Algorithm)

Step 1: Input: x_1, y_1, $\hat{b}_1$, $\hat{c}_1$, $\beta_1^{(i)}$ $(i = 1, \cdots, K_1)$, $\xi_1^{(j)}$ $(j = 1, \cdots, K_2)$; Set $k = 1$.

Step 2: Calculate $\hat{b}_k$, $\beta_k^{(i)}$, and for $i = 1, \ldots, K_1$, perform the following update for $x_k^{(i)}$:

$$x_{k+1}^{(i)} = \mathrm{Prox}_{h_i,\mathcal{X}_i}^{\beta_k^{(i)}}\left(x_k^{(i)} - \frac{1}{\beta_k^{(i)}}\nabla_{x^{(i)}} f(v_{k+1}^{(i)}, y_k) - \frac{1}{\beta_k^{(i)}}\hat{b}_k x_k^{(i)}\right). \tag{5.4}$$

Step 3: Calculate $\hat{c}_k$, $\xi_k^{(j)}$, and for $j = 1, \ldots, K_2$ and perform the following update for $y_k^{(j)}$:

$$y_{k+1} = \mathrm{Prox}_{g_j,\mathcal{Y}_j}^{\xi_k^{(j)}}\left(y_k^{(j)} + \frac{1}{\xi_k^{(j)}}\nabla_{y^{(j)}} f(x_{k+1}, w_{k+1}^{(j)}) - \frac{1}{\xi_k^{(j)}}\hat{c}_k y_k^{(j)}\right). \tag{5.5}$$

Step 4: If some stationary condition is satisfied, stop; otherwise, set $k = k + 1$, go to Step 2.

specified later in each of the different problem settings to be studied. Note that it reduces to the AGP algorithm when $K_1 = K_2 = 1$, $h_i(\cdot) = 0$ and $g_j(\cdot) = 0$.

Before analyzing the convergence of Algorithm 2, we define the stationarity gap as the termination criterion as follows.

Definition 5.1 At each iteration of Algorithm 2, the stationarity gap for problem (BP) w.r.t. $l(x, y)$ is defined as:

$$\nabla\mathcal{G}\left(x_k, y_k\right) := \begin{bmatrix} \beta_k^{(1)}\left(x_k^{(1)} - \mathrm{Prox}_{h_1,\mathcal{X}_1}^{\beta_k^{(1)}}(x_k^{(1)} - \frac{1}{\beta_k^{(1)}}\nabla_{x^{(1)}} f(x_k, y_k))\right) \\ \vdots \\ \beta_k^{(K_1)}\left(x_k^{(K_1)} - \mathrm{Prox}_{h_{K_1},\mathcal{X}_{K_1}}^{\beta_k^{(K_1)}}(x_k^{(K_1)} - \frac{1}{\beta_k^{(K_1)}}\nabla_{x^{(K_1)}} f(x_k, y_k))\right) \\ \xi_k^{(1)}\left(y_k^{(1)} - \mathrm{Prox}_{g_1,\mathcal{Y}_1}^{\xi_k^{(1)}}\left(y_k^{(1)} + \frac{1}{\xi_k^{(1)}}\nabla_{y^{(1)}} f(x_k, y_k)\right)\right) \\ \vdots \\ \xi_k^{(K_2)}\left(y_k^{(K_2)} - \mathrm{Prox}_{g_{K_2},\mathcal{Y}_{K_2}}^{\xi_k^{(K_2)}}\left(y_k^{(K_2)} + \frac{1}{\xi_k^{(K_2)}}\nabla_{y^{(K_2)}} f(x_k, y_k)\right)\right) \end{bmatrix}.$$

For simplicity, we denote $\nabla\mathcal{G}_k := \nabla\mathcal{G}(x_k, y_k)$, $(\nabla\mathcal{G}_k)_{x^{(i)}} := \beta_k^{(i)}(x_k^{(i)} - \mathrm{Prox}_{h_i,\mathcal{X}_i}^{\beta_k^{(i)}}(x_k^{(i)} - \frac{1}{\beta_k^{(i)}}\nabla_{x^{(i)}} f(x_k, y_k)))$ and $(\nabla\mathcal{G}_k)_{y^{(j)}} := \xi_k^{(j)}\left(y_k^{(j)} - \mathrm{Prox}_{g_j,\mathcal{Y}_j}^{\xi_k^{(j)}}(y_k^{(j)} + \frac{1}{\xi_k^{(j)}}\nabla_{y^{(j)}} f\left(x_k, y_k\right))\right)$.

Definition 5.2 At each iteration of Algorithm 2, the stationarity gap for problem (BP) w.r.t. $\bar{l}_k(x, y)$, denoted by $\nabla\bar{\mathcal{G}}\left(x_k, y_k\right)$, is defined almost the same as in Definition 5.1 except by replacing f with $\bar{f}_k$. For simplicity, we denote $\nabla\bar{\mathcal{G}}_k := \nabla\bar{\mathcal{G}}(x_k, y_k)$, $(\nabla\bar{\mathcal{G}}_k)_{x^{(i)}} := \beta_k^{(i)}(x_k^{(i)} - \mathrm{Prox}_{h_i,\mathcal{X}_i}^{\beta_k^{(i)}}(x_k^{(i)} - \frac{1}{\beta_k^{(i)}}\nabla_{x^{(i)}}\bar{f}_k\left(x_k, y_k\right)))$ and $(\nabla\bar{\mathcal{G}}_k)_{y^{(j)}} := \xi_k^{(j)}\left(y_k^{(j)} - \mathrm{Prox}_{g_j,\mathcal{Y}_j}^{\xi_k^{(j)}}(y_k^{(j)} + \frac{1}{\xi_k^{(j)}}\nabla_{y^{(j)}}\bar{f}_k\left(x_k, y_k\right))\right)$.

5.1 Complexity Analysis for Nonconvex-(Strongly) Concave Setting

5.1.1 Nonconvex-Strongly Concave Setting

In this subsection, we analyze the iteration complexity of Algorithm 2 for solving the nonsmooth one-sided block-wise nonconvex-strongly concave minimax optimization problem (BP) with $K_2 = 1$, i.e., $f(x, y)$ is nonconvex w.r.t. x for any fixed $y \in \mathcal{Y}$, and $\hat{\mu}$-strongly concave w.r.t. y for any given $x \in \mathcal{X}$. Under this setting, $\forall k \geq 1$, we set

$$\beta_k^{(i)} = \hat{\eta}, \xi_k^{(j)} = \tfrac{1}{\hat{\rho}}, \text{ and } \hat{b}_k = \hat{c}_k = 0. \tag{5.6}$$

Lemma 5.1 below shows a descent result for the x_k update.

Lemma 5.1 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.6). If $\forall k, \hat{\eta} > L_{11}$, we have

$$l(x_{k+1}, y_k) - l(x_k, y_k) \leq -\frac{\hat{\eta}}{2}\|x_{k+1} - x_k\|^2. \tag{5.7}$$

Proof By the optimality condition for (5.4), $\forall i$ we have

$$\langle \nabla_{x^{(i)}} f(v_{k+1}^{(i)}, y_k) + \partial h_i(x_{k+1}^{(i)}) + \hat{\eta}(x_{k+1}^{(i)} - x_k^{(i)}), x_k^{(i)} - x_{k+1}^{(i)} \rangle \geq 0. \tag{5.8}$$

By Assumption 2.1 and the convexity of $h_i(x)$, we obtain

$$\begin{aligned} & l(x_{k+1}^{(i)}, v_{k+1}^{(-i)}, y_k) - l(x_k^{(i)}, v_{k+1}^{(-i)}, y_k) \\ \leq & \langle \nabla_{x^{(i)}} f(v_{k+1}^{(i)}, y_k) + \partial h_i(x_{k+1}^{(i)}), x_{k+1}^{(i)} - x_k^{(i)} \rangle + \frac{L_{11}}{2}\|x_{k+1}^{(i)} - x_k^{(i)}\|^2. \end{aligned} \tag{5.9}$$

Adding (5.8) and (5.9), and summing up the inequality from $i = 1$ to $i = K_1$, we have

$$l(x_{k+1}, y_k) - l\left(x_k, y_k\right) \leq -\sum_{i=1}^{K_1}\left(\hat{\eta} - \frac{L_{11}}{2}\right)\|x_{k+1}^{(i)} - x_k^{(i)}\|^2.$$

By using the assumption that $\hat{\eta} > L_{11}$, we complete the proof. ∎

The rest of the proof is almost the same as that of the AGP algorithm shown in Section 3 since the ascent step of y's update can be similarly estimated. By replacing $\nabla_y f(x_{k+1}, y_k)$ with $\nabla_y f(x_{k+1}, y_k) - \partial g_1(y_{k+1})$, and using $\langle \partial g_1(y_{k+1}) - \partial g_1(y_k), y_{k+1} - y_k \rangle \geq 0$, and the fact that $\|x_{k+1}^{(i)} - x_k^{(i)}\| \leq \|x_{k+1} - x_k\|$ and $\|v_{k+1}^{(i)} - x_k\| \leq \|x_{k+1} - x_k\|$, we can show similar results to those in Lemmas 3.1 and 3.2. Then, similar to the proof of Theorem 3.1, we obtain the following convergence result by some parameters replacement, e.g., η with $\hat{\eta} + L_{11}$ in (3.23) and η^2 with $K_1\left(\hat{\eta} + L_{11}\right)^2$ in (3.25). We omit the proof details here for simplicity.

In particular, letting $\nabla\mathcal{G}\left(x_k, y_k\right)$ be defined as in Definition 5.1 and $\varepsilon > 0$ be a given target accuracy, we provide a bound on $\mathcal{T}(\varepsilon)$, the first iteration index to achieve an ε-stationary point, i.e., $\|\nabla\mathcal{G}(x_k, y_k)\| \leq \varepsilon$, which is equivalent to

$$\mathcal{T}(\varepsilon) := \min\{k \mid \|\nabla\mathcal{G}(x_k, y_k)\| \leq \varepsilon\}. \tag{5.10}$$

Theorem 5.1 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.6). If

$$\hat{\eta} > L_{11}, \hat{\eta} > L_{12}^2\hat{\rho} + \frac{4L_{12}^2}{\hat{\rho}\hat{\mu}^2}, \text{ and } \hat{\rho} \leq \frac{\hat{\mu}}{4L_{22}^2},$$

then it holds that

$$\mathcal{T}\left(\varepsilon\right) \leq \frac{L_1 - \underline{L}}{r_1\varepsilon^2},$$

where $r_1 := \frac{\min\left\{\frac{\hat{\eta}}{2} - \frac{\hat{\rho}L_{12}^2}{2} - \frac{2L_{12}^2}{\hat{\rho}\hat{\mu}^2}, \frac{3\hat{\mu} - \hat{\rho}L_{22}^2}{2} + \frac{\hat{\mu} - 4\hat{\rho}L_{22}^2}{2\hat{\rho}\hat{\mu}}\right\}}{\max\left\{K_1(\hat{\eta} + L_{11})^2 + 2L_{12}^2, \frac{2}{\hat{\rho}^2}\right\}}$, $L_1 := l(x_1, y_1) + \frac{2\delta_y^2}{\hat{\rho}^2\hat{\mu}}$ and $\underline{L} := \underline{l} - (\hat{\mu} + \frac{7}{2\hat{\rho}} - \frac{\hat{\rho}L_{22}^2}{2} - \frac{2L_{22}^2}{\hat{\mu}})\delta_y^2$ with $\underline{l} := \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} l(x, y)$ and $\delta_y := \max\{\|y_1 - y_2\| \mid y_1, y_2 \in \mathcal{Y}\}$.

Theorem 5.1 implies that the iteration complexity of Algorithm 2 to obtain an ε-stationary point for solving general block-wise nonsmooth nonconvex-strongly concave minimax problems (BP) is bounded by $\mathcal{O}\left(L^2\varepsilon^{-2}\right)$.

5.1.2 Nonconvex-Concave Setting

We analyze the iteration complexity of Algorithm 2 for solving (BP) in the nonconvex-concave setting. Under this setting, let $K_2 = 1$, and $\forall k \geq 1$, we set

$$\beta_k^{(i)} = \iota + \vartheta_k, \quad \xi_k^{(j)} = \frac{1}{\xi}, \text{ and } \hat{b}_k = 0, \tag{5.11}$$

where ϑ_k is stepsize parameter to be defined later.

Assumption 5.1 $\{\hat{c}_k\}$ is a nonnegative monotonically decreasing sequence.

Lemma 5.2 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.11). If $\forall k, \vartheta_k > L_{11}$, we have

$$l(x_{k+1}, y_k) - l(x_k, y_k) \leq -\left(\iota + \frac{\vartheta_k}{2}\right) \|x_{k+1} - x_k\|^2. \tag{5.12}$$

Proof The proof is the same with that of Lemma 5.1 except replacing $\beta_k^{(i)}$ by $\iota + \vartheta_k$. We omit the details here. ■

The rest of the proof is almost the same as that of the AGP algorithm shown in Section 3 since the ascent step of y's update can be similarly estimated. By replacing $\nabla_y f_k(x_{k+1}, y_k)$ with $\nabla_y \bar{f}_k(x_{k+1}, y_k) - \partial g_1(y_{k+1})$, and using $\langle \partial g_1(y_{k+1}) - \partial g_1(y_k), y_{k+1} - y_k \rangle \geq 0$, and the fact that $\|x_{k+1}^{(i)} - x_k^{(i)}\| \leq \|x_{k+1} - x_k\|$ and $\|v_{k+1}^{(i)} - x_k\| \leq \|x_{k+1} - x_k\|$, we can prove similar results to those in Lemmas 3.3 and 3.4. Then, similar to the proof of Theorem 3.2, we obtain the following result only through some parameters replacement, e.g., $\bar{\eta} + \bar{\beta}_k$ with $\iota + \vartheta_k + L_{11}$ in (3.49) and $(\bar{\eta} + \bar{\beta}_k)^2$ with $K_1 (\iota + \vartheta_k + L_{11})^2$ in (3.51).

Theorem 5.2 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.11). If $\xi \leq \frac{1}{10L_{22}}$, $\hat{c}_k = \frac{19}{20\xi k^{1/4}}, k \geq 1$, $\tau > \max\{\frac{\xi \hat{c}_1^2 (L_{11} + 2\iota - \xi L_{12}^2)}{16 L_{12}^2}, 2\}$, then for any given $\varepsilon > 0$,

$$\mathcal{T}(\varepsilon) \leq \max\left(\left(\frac{2 \cdot 80^2 \xi (\tau - 2) L_{12}^2 r_2 r_3}{19^2 \varepsilon^2} + 1\right)^2, \frac{19^4 \hat{\delta}_y^4}{10^4 \xi^4 \varepsilon^4}\right),$$

where $r_2 = \mathcal{L}_2 - \underline{\mathcal{L}} + (8 \cdot 2^{1/4} + \frac{19}{40}) \frac{\hat{\delta}_y^2}{\xi}$, $r_1 = \frac{8 K_1 \tau^2}{(\tau - 2)^2} + \frac{19^4 \left(2K_1 \left(\xi L_{12}^2 - \iota + L_{11}\right)^2 + 2L_{12}^2\right)}{64 \cdot 20^4 \xi^2 (\tau - 2)^2 L_{12}^4}$, $r_3 = \max\{r_1, \frac{19^2}{1440\sqrt{2}(\tau - 2)\xi^2 L_{12}^2}\}$, $\hat{\delta}_y := \max\{\|y\| \mid y \in \mathcal{Y}\}$, $\mathcal{L}_2 := l(x_2, y_2) + \frac{16 \hat{\delta}_y^2}{\xi^2 \hat{c}_2}$, $\underline{\mathcal{L}} := \underline{l} - \left(2^{13/4} + 15 + \frac{19}{40}\right) \frac{\hat{\delta}_y^2}{\xi}$ with $\underline{l} := \min_{(x,y) \in \mathcal{X} \times \mathcal{Y}} l(x, y)$.

By setting $\xi = \frac{1}{10L_{22}}$, from Theorem 5.2, we conclude that for any given $\varepsilon \in (0, 1)$,

$$\mathcal{T}(\varepsilon) \leq \max\left(\left(\frac{1280(\tau - 2) L_{12}^2 r_2 r_3}{19^2 L_{22} \varepsilon^2} + 1\right)^2, \frac{19^4 \hat{\delta}_y^4 L_{22}^4}{\varepsilon^4}\right) = \mathcal{O}\left(L^4 \varepsilon^{-4}\right).$$

This implies that the iteration complexity of the proposed BAPG algorithm to obtain a point that satisfies $\|\nabla \mathcal{G}(x_k, y_k)\| \leq \varepsilon$ with $\nabla \mathcal{G}(x_k, y_k)$ being defined in Definition 5.1 for nonsmooth block-wise nonconvex-concave minimax problems (BP) is bounded by $\mathcal{O}\left(L^4 \varepsilon^{-4}\right)$.

5.2 Complexity Analysis for (Strongly) Convex-Nonconcave Setting

In this subsection we analyze the iteration complexity of Algorithm 2 for solving (BP) with $K_1 = 1$ under the convex-nonconcave setting.

We first consider the strongly convex-nonconcave setting. Under this setting, $\forall k \geq 1$, we set

$$\beta_k^{(i)} = \frac{1}{\varrho}, \xi_k^{(j)} = \bar{\xi}, \hat{b}_k = \hat{c}_k = 0, \tag{5.13}$$

Lemma 5.3 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.13). If $\forall k, \bar{\xi} > L_{22}$, then we have

$$l(x_{k+1}, y_{k+1}) - l(x_{k+1}, y_k) \geq \frac{\bar{\xi}}{2} \|y_{k+1} - y_k\|^2. \tag{5.14}$$

Proof By the optimality condition for (5.5), $\forall j$, we have

$$\langle \nabla_{y^{(j)}} f(x_{k+1}, w_{k+1}^{(j)}) - \partial g_j(y_{k+1}^{(j)}) - \bar{\xi}(y_{k+1}^{(j)} - y_k^{(j)}), y_k^{(j)} - y_{k+1}^{(j)} \rangle \leq 0. \tag{5.15}$$

By the convexity of $g_j(\cdot)$ and Assumption 2.1, the gradient of f is Lipschitz continuous, implying that

$$\begin{aligned} & l(x_{k+1}, w_{k+1}^{(-j)}, y_{k+1}^{(j)}) - l(x_{k+1}, w_{k+1}^{(-j)}, y_k^{(j)}) \\ \geq & \langle \nabla_{y^{(j)}} f(x_{k+1}, w_{k+1}^{(j)}) - \partial g_j(y_{k+1}^{(j)}), y_{k+1}^{(j)} - y_k^{(j)} \rangle - \frac{L_{22}}{2} \|y_{k+1}^{(j)} - y_k^{(j)}\|^2. \end{aligned} \tag{5.16}$$

Adding (5.15) and (5.16), and then summing it up from $j=1$ to K_2, we have

$$l(x_{k+1},y_{k+1})-l(x_{k+1},y_k)\geq\sum_{j=1}^{K_2}(\bar{\xi}-\frac{L_{22}}{2})\|y_{k+1}^{(j)}-y_k^{(j)}\|^2.$$

The result then follows by the assumption that $\bar{\xi}>L_{22}$. ∎

By Lemma 5.3, similar to the proof of Lemma 4.1-4.2 and Theorem 4.1 in Subsection 4.1, we can prove the following theorem and we omit the details here.

Theorem 5.3 Suppose that Assumption 2.1 holds. Let $\{(x_k,y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.13). If

$$\bar{\xi}>L_{22},\bar{\xi}>L_{21}^2\varrho+\frac{4L_{21}^2}{\varrho\hat{\theta}^2},\ \text{and}\ \varrho\leq\frac{\hat{\theta}}{4L_{11}^2},$$

then $\forall\varepsilon>0$, it holds that

$$\mathcal{T}(\varepsilon)\leq\tfrac{\overline{L}-\hat{L}_1}{\hat{r}_1\varepsilon^2},$$

where $\hat{r}_1:=\frac{\min\left\{\frac{\bar{\xi}}{2}-\frac{\varrho L_{21}^2}{2}-\frac{2L_{21}^2}{\varrho\hat{\theta}^2},\frac{3\hat{\theta}-\varrho L_{11}^2}{2}+\frac{\hat{\theta}-4\varrho L_{11}^2}{2\varrho\hat{\theta}}\right\}}{\max\left\{\frac{1}{\varrho^2}+2K_2L_{12}^2,2K_2(\bar{\xi}+L_{22})^2\right\}}$ and $\overline{L}=\hat{l}+(\hat{\theta}+\frac{7}{2\varrho}-\frac{\varrho L_{11}^2}{2}-\frac{2L_{11}^2}{\hat{\theta}})\delta_x^2$, $\hat{L}_1:=\hat{l}-\frac{4\delta_x^2}{\varrho^2\hat{\theta}}-(\frac{L_{21}\varrho}{2}+\frac{2L_{21}^2}{\hat{\theta}^2\varrho})\|y_2-y_1\|^2$ with $\hat{l}:=\max\limits_{(x,y)\in\mathcal{X}\times\mathcal{Y}}l(x,y)$ and $\delta_x=\max\{\|x_1-x_2\|\mid\forall x_1,x_2\in\mathcal{X}\}$.

Theorem 5.3 implies that the iteration complexity of the proposed BAPG algorithm to obtain an ε-stationary point for smooth strongly convex-nonconcave minimax problems (BP) is bounded by $\mathcal{O}\left(L^2\varepsilon^{-2}\right)$.

We can also analyze the iteration complexity of Algorithm 2 applied to the general convex-nonconcave setting. Under this setting, $\forall k\geq1$, we set

$$\beta_k^{(i)}=\frac{1}{\bar{\varrho}},\quad\xi_k^{(j)}=\bar{\iota}+\bar{\vartheta}_k,\quad\hat{c}_k=0,\tag{5.17}$$

where $\xi_k^{(j)}$ is stepsize parameter to be defined later.

Assumption 5.2 $\{\hat{b}_k\}$ is a nonnegative monotonically decreasing sequence.

Similar to the proof of Lemma 4.3-4.4 and Theorem 4.2 in Subsection 4.2, we can prove the following theorem.

Theorem 5.4 Suppose that Assumption 2.1 holds. Let $\{(x_k,y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.17). If $\bar{\varrho}\leq\frac{1}{10L_{11}}$,$\hat{b}_k=\frac{19}{20\bar{\varrho}k^{1/4}},k\geq1$, $\tau>\max\{\frac{\bar{\varrho}\hat{b}_2^2(L_{22}+2\bar{\iota}-\bar{\varrho}L_{21}^2)}{16L_{21}^2},2\}$, then for any given $\varepsilon>0$,

$$\mathcal{T}(\varepsilon)\leq\max\left(\left(\frac{2\cdot80^2\bar{\varrho}(\tau-2)L_{21}^2\hat{r}_2\hat{r}_3}{19^2\varepsilon^2}+2\right)^2,\frac{19^4\hat{\delta}_x^4}{10^4\bar{\varrho}^4\varepsilon^4}\right),$$

where $\hat{\delta}_x:=\max\{\|x\|\mid x\in\mathcal{X}\}$,

$$\hat{r}_2:=\hat{\mathcal{L}}_0-\hat{\mathcal{L}}_2+\frac{8\cdot2^{1/4}}{\bar{\varrho}}\hat{\delta}_x^2+\frac{19}{40\bar{\varrho}}\hat{\delta}_x^2,$$

$\hat{r}_3:=\max\{\hat{d}_1,\frac{10+20K_2\bar{\varrho}^2L_{12}^2}{9\bar{\varrho}\alpha_2}\}$ with $\hat{d}_1=\frac{16K_2\tau^2}{(\tau-2)^2}+\frac{19^4\left(K_2\left(\bar{\varrho}L_{21}^2-\bar{\iota}+L_{22}\right)^2\right)}{16\cdot20^4\bar{\varrho}^2(\tau-2)^2L_{21}^4}$, $\hat{\mathcal{L}}_0:=\hat{l}+\left(2^{13/4}+15+\frac{19}{40}\right)\frac{\hat{\delta}_x^2}{\bar{\varrho}}$ with $\hat{l}:=\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}}l(x,y)$, $\hat{\mathcal{L}}_2:=L(x_2,y_2)-\frac{16\hat{\delta}_x^2}{\bar{\varrho}^2\hat{b}_2}-\frac{8\hat{\delta}_x^2}{\bar{\varrho}}-\left(\frac{\bar{\varrho}L_{21}^2}{2}+\frac{16L_{21}^2}{\bar{\varrho}(\hat{b}_2)^2}\right)\|y_2-y_1\|^2$.

Theorem 5.4 implies that the iteration complexity of Algorithm 2 to obtain an ε-stationary point for general block-wise nonsmooth convex-nonconcave minimax problems (BP) is bounded by $\mathcal{O}\left(L^4\varepsilon^{-4}\right)$.

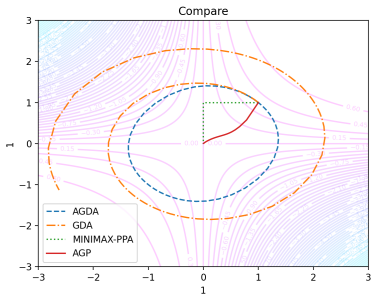


Fig. 1 Comparison of four algorithms for solving the dirac-GAN problem

Table 2 Comparison of iteration number and time for AGP algorithm and MINIMAX-PPA algorithm

	Distance	Total iteration	Time(s)
MINIMAX-PPA	0.01	18812	0.0313
AGP	0.01	144	0.0001

6 Numerical results

In this section, we compare the numerical performance of the proposed AGP algorithm with the gradient descent ascent algorithm (GDA), the alternating gradient descent ascent algorithm (AGDA) and the state-of-the-art nested-looped algorithm, i.e., MINIMAX-PPA in [32] through two representative test problems. The first numerical test is implemented in Python 3.9 and run with an Apple M1 processor, while the second one is carried out on an NVIDIA Tesla P100 GPU.

6.1 **Dirac-GAN Problem**

The Dirac-GAN problem [54] can be formulated as the following nonconvex-concave minimax problem :

$$\min_x \max_y L(x,y) = -\log(1+exp(-xy)) + \log 2, \tag{6.1}$$

where $(0,0)$ is the unique stationary point.

Experimental setup. Let α_x and β_y be the stepsize of x and y, respectively, and T denote the outer-loop iteration number for AGP, GDA, AGDA and MINIMAX-PPA algorithms. The initial point of all algorithms is chosen as $(1,1)$.

For the AGP algorithm, we set $T = 72, \alpha_x = \frac{0.8}{\sqrt{k}}$, $\beta_y = 0.3$ and $c_k = \frac{0.5}{k^{1/4}}$. For the GDA and AGDA algorithms, we set $T = 100$, $\alpha_x = 0.3$ and $\beta_y = 0.3$. For MINIMAX-PPA algorithm, we set $T = 30, \epsilon = 0.01, l = 2, \mu_y = 0.000625$ and $g(x,y) = L(x,y) - 0.00125\|y-1\|^2$. Moreover, for MINIMAX-PPA algorithm, we choose $\delta = 1e-6$ and $\tilde{\epsilon} = 5.8e-9$ through grid search in the sub-rountines AG2 and AGD respectively, since the theoretical settings for these two parameters are $\delta = 2.38e-27$ and $\tilde{\epsilon} = 4.06e-60$ respectively, which will cause the algorithm to converge very slowly in practice.

Results. Fig 1 shows the sequences generated by these four algorithms. We find that GDA fails to converge, and AGDA falls into a limit cycle no matter what stepsize is chosen, which is shown in Lemma 2.4 in [54]. Numerical results show that both AGP and MINIMAX-PPA algorithms approach the unique stationary point. Note that there are much more parameters needed to be adjusted in advance in MINIMAX-PPA algorithm than the AGP algorithm. Table 2 shows the total number of iterations and CPU times for both algorithms in solving the dirac-GAN problem. We count the total number of iterations in all the sub-rountines in MINIMAX-PPA in [32]. The "Distance" column reports the Euclidean distance between the point obtained by both algorithms and the unique stationary point. Note that the theoretical settings for $\tilde{\epsilon}$ in the sub-rountine AGD in MINIMAX-PPA is actually $4.06e-60$ which is almost unacceptable in training a neural network with millions of parameters. In this test, we reduce the precision for the sub-rountine AGD in MINIMAX-PPA algorithm in order to obtain a meaningful solution, otherwise it will get stuck. To obtain a solution near $(0,0)$ with the accuracy within 0.01, MINIMAX-PPA algorithm needs millions number of iterations totally, and hence takes much more CPU time than the AGP algorithm.

6.2 Robust learning over multiple domains

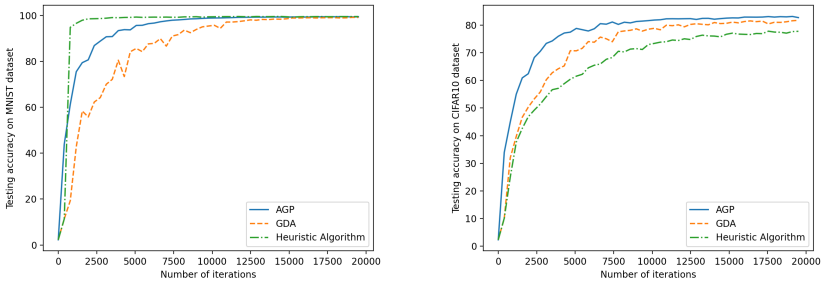
In this subsection, we perform some numerical tests for solving a robust learning problem over multiple domains [46], formulated as a nonconvex-linear minimax problem:

$$\min_{x} \max_{y \in \Delta} y^T F(x), \tag{6.2}$$

where $F(x) := [f_1(x); \ldots; f_M(x)] \in \mathbb{R}^{M\times 1}$ with $f_m(x) = \frac{1}{|\mathcal{S}_m|}\sum_{i=1}^{|\mathcal{S}_m|} \ell\left(d_i^m, l_i^m, x\right)$ and ℓ can be any non-negative loss function, M is the number of tasks and x denotes the network parameters. $y \in \Delta$ describes the weights over different tasks and Δ is the simplex, i.e., $\Delta = \left\{(y_1, \cdots, y_M)^T \mid \sum_{m=1}^{M} y_m = 1, y_m \geq 0\right\}$. We compare the proposed AGP algorithm with the GDA algorithm and a heuristic training algorithm with even weights, i.e., fixing $y_1 = y_2 = 0.5$ when $M = 2$, which is also used as a baseline algorithm in many other existing works [33]. We consider two image classification problems with MNIST [52] and CIFAR10 datasets[5]. MNIST consists of images about handwritten digits while CIFAR10 contains 10 different classes such as airplanes, cars, birds etc. Our goal is to train a neural network that works on these two completely unrelated problems simultaneously. Since cross-entropy loss is popular in multi-class classification problem, we use it as our loss function. We should point out that the quality of the algorithm for solving a robust learning problem over multiple domains is measured by the worst case accuracy over all domains [46].

Data augmentation. Since the data in MNIST are 28×28 gray images while those in CIFAR10 are 32×32 RGB images, we first repeat gray channel to RGB channels and resize the examples in MNIST from 28×28 to 32×32 using bilinear interpolation. We adopt AlexNet [4] as our base model which is the same as the one used in [33]. Then, we convert 32×32 to 224×224 with the same method to fit the input of AlexNet.

Experiment setup. We set $\frac{1}{\beta_k}$, γ_k, and c_k to be $\frac{2}{2+\sqrt{k}}$, 100, $\frac{1}{10+k^{\frac{1}{4}}}$ respectively for the AGP algorithm. In the GDA algorithm, we set $\alpha = 0.02$ and $\beta = 0.01$. In the Heuristic Algorithm, we set the stepsize of x to be 0.02, but fix y as $(0.5, 0.5)^T$. We set the batch size in all tasks to be 128 and run 50 epochs for all algorithms. Moreover, we sample our results every epoch and calculate the average accuracy on these two testing sets to evaluate the performance of different algorithms.



(a) (b)

Fig. 2 Performance of three algorithms for solving robust multi-task learning problem.

Method	MNIST		CIFAR10	
	Training set	Testing set	Training set	Testing set
Heuristic Algorithm	99.976	99.479	99.868	77.694
GDA	99.355	99.128	99.994	81.611
AGP	99.954	99.329	99.998	83.173

Table 3 Comparison the test accuracies for three algorithms on robust multi-task learning

Results. Figure 2(a) shows the testing accuracy on MNIST dataset. All three algorithms achieve high precision, while AGP algorithm takes less time than that of GDA. Figure 2(b) shows the testing accuracy on CIFAR10 dataset. AGP algorithm still performs slightly better than other two algorithms.

Table 3 shows the accuracies of all algorithms. MNIST and CIFAR10 indicate training only on MNIST and CIFAR10 dataset respectively. GDA and Heuristic Algorithm can achieve good performance while AGP algorithm can further improve the performance and provide a more reliable model for multi-task learning.

Remark. Due to hardware (especially memory) limitations, we use mini-batch randomized gradient instead of the exact gradient at each iteration in our numerical experiment, which is the same as in [33]. Since at each iteration solving the inner subproblem will take huge amounts of time in the nested loop algorithms, e.g., MINIMAX-PPA algorithm, as the gradient calculation needs to go through the whole dataset and they need to calculate gradient several times per iteration. Moreover, high requirements for precision in solving subproblems will occupy a lot of memory. Hence, we were not able to compare AGP algorithm with nested-loop algorithm for this test problem.

7 Conclusions and Discussion

In this paper, we propose a unified single-loop algorithm for general smooth one block nonconvex-concave or convex-nonconcave minimax problems. At each iteration, only one gradient projection step is employed for updating x and y respectively. The gradient complexity of the proposed unified AGP algorithm under four different settings are established. We prove that an ε-first order stationary point of f can be obtained in $\mathcal{O}\left(\varepsilon^{-4}\right)$ (resp. $\mathcal{O}\left(\varepsilon^{-2}\right)$) iterations for nonconvex-concave (resp. nonconvex-strongly concave) setting. To the best of our knowledge, they are the best complexity bounds among single-loop algorithms in general nonconvex-(strongly) concave settings, and very simple to be implemented. Numerical results show the efficiency of the proposed AGP algorithm. Nonetheless, there is still a certain gap between the iteration complexity of the proposed AGP algorithm and the best complexity of $\mathcal{O}(\varepsilon^{-2.5})$ among nested-looped algorithms in nonconvex-concave setting, which is also an open problem worth studying in the future.

Moreover, we consider the (strongly) convex-nonconcave setting of (P). Under this setting, the whole problem is convex. However, for any given x, to solve the inner max subproblem, i.e., $\max_{y\in\mathcal{Y}} f(x,y)$, is already NP-hard. Due to this reason, almost all the existing nested-loop algorithms will lose the theoretic guarantee since they all need to solve the inner subproblem exactly, or inexactly but only an error proportional to the accuracy ε is allowed. We show that the proposed unified single-loop AGP algorithm can deal with general convex-nonconcave setting. More specifically, the gradient complexity to obtain an ε-first-order stationary point of f is $\mathcal{O}\left(\varepsilon^{-2}\right)$ (resp., $\mathcal{O}\left(\varepsilon^{-4}\right)$) under the strongly convex-nonconcave (resp., convex-nonconcave) setting. To the best of our knowledge, these theoretical performance guarantees under this setting have not been obtained before in the literature.

Furthermore, we consider more general nonsmooth multi-block nonconvex-(strongly) concave and (strongly) convex-nonconcave minimax problems, which include smooth one block setting as a special case. We propose a block alternating proximal gradient algorithm (BAPG) to solve it. At each iteration, only simple proximal gradient steps are employed for updating one block variable, and for each block of multi-block variables alternatively. We prove the gradient complexity of the proposed BAPG algorithm under four different settings. To the best of our knowledge, these are the state-of-the-art single loop algorithms under nonconvex-concave setting and the first two theoretically guaranteed convergence results reported in the literature under (non)smooth multi-blocks (strongly) convex-nonconcave minimax problems. Our development shows that single loop algorithms do not need to solve the inner subproblem, and thus can be easily generalized. This is also the reason that the iteration complexity analysis will be more difficult than that of nested-looped algorithms. It will also be interesting to study whether the iteration complexity of the proposed AGP algorithm is tightest or not among single loop algorithms.

References

1. S. S. Abadeh, P. M. M. Esfahani, and D. Kuhn. *Distributionally robust logistic regression.* In NeurIPS, pages 1576–1584, 2015.
2. J. Abernethy, K. A. Lai, and A. Wibisono. *Last-iterate convergence rates for min-max optimization.* ArXiv preprint arXiv:1906.02027, 2019.
3. L. Adolphs, H. Daneshmand, A. Lucchi, and T. Hofmann. *Local saddle point optimization: A curvature exploitation approach.* The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, pages 486–495, 2019.
4. A. Krizhevsky, I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, 25:1097–1105, 2012.
5. Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J]. 2009.

6. J. Bailey, G. Gidel, and G. Piliouras. *Finite regret and cycles with fixed step-size via alternating gradient descent-ascent*. Proceedings of Thirty Third Conference on Learning Theory, PMLR 125:391–407, 2020.
7. S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
8. A. Chambolle and T. Pock. *On the ergodic convergence rates of a first-order primal-dual algorithm*. Mathematical Programming, 159(1-2):253–287, 2016.
9. Y. Chen, G. Lan, and Y. Ouyang. *Optimal primal-dual methods for a class of saddle point problems*. SIAM Journal on Optimization, 24(4):1779–1814, 2014.
10. Y. Chen, G. Lan, and Y. Ouyang. *Accelerated schemes for a class of variational inequalities*. Mathematical Programming, 165(1):113–149, 2017.
11. C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. *Training gans with optimism*. ArXiv preprint arXiv:1711.00141, 2017.
12. C. Daskalakis and I. Panageas. *The limit points of (optimistic) gradient descent in min-max optimization*. In Advances in Neural Information Processing Systems, pages 9236–9246, 2018.
13. C. D. Dang and G. Lan. *Randomized first-Order methods for saddle point optimization*. Technical Report, Department of Industrial and Systems Engineering, University of Florida, 2015.
14. C. D. Dang and G. Lan. *On the convergence properties of non-Euclidean extragradient methods for variational Inequalities with Generalized Monotone Operators*. Computational Optimization and Applications, 60(2):277–310, 2015.
15. L. Flokas, E. Vlatakis-Gkaragkounis, and G. Piliouras. *Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games*. Advances in Neural Information Processing Systems, 32:10450–10461, 2019.
16. G. B. Giannakis, Q. Ling, G. Mateos, I. D. Schizas, and H. Zhu. *Decentralized learning for wireless communications and networking*. Splitting Methods in Communication, Imaging, Science, and Engineering, Springer, Cham, pages 461–497, 2016.
17. G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. *A variational inequality perspective on generative adversarial networks*. International Conference on Learning Representations, 2018.
18. G. Gidel, R. A. Hemmat, M. Pezeshki, G. Huang, R. Lepriol, S. Lacoste-Julien, and I. Mitliagkas. *Negative momentum for improved game dynamics*. The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, pages 1802–1811, 2019.
19. R. Giordano, T. Broderick, and M. I. Jordan. *Covariances, robustness, and variational bayes*. Journal of Machine Learning Research, 19(51), 2018.
20. Y. He and R. D. C. Monteiro. *An accelerated hpe-type algorithm for a class of composite convex-concave saddle-point problems*. SIAM Journal on Optimization, 26(1):29—56, 2016.
21. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. *Gans trained by a two time-scale update rule converge to a local nash equilibrium*. In NeurIPS, pages 6626—6637, 2017.
22. J. Ho and S. Ermon. *Generative adversarial imitation learning*. In Advances in Neural Information Processing Systems, pages 4565—4573, 2016.
23. Y. Hsieh, C. Liu, and V. Cevher. *Finding mixed nash equilibria of generative adversarial networks*. International Conference on Machine Learning, PMLR, pages 2810–2819, 2019.
24. C. Jin, P. Netrapalli, and M.I. Jordan. *Minmax optimization: stable limit points of gradient descent ascent are locally optimal*. ArXiv preprint arXiv:1902.00618, 2019.
25. W. Kong and R. D. C. Monteiro. *An accelerated inexact proximal point method for solving nonconvex concave min-max problems*. ArXiv preprint arXiv:1905.13433, 2019.
26. G. Lan, *First-order and stochastic optimization methods for machine learning*, Springer-Nature, 2020.
27. G. Lan and R. D. C. Monteiro. *Iteration-complexity of first-order augmented lagrangian methods for convex programming*. Mathematical Programming, 155(1-2):511—547, 2016.
28. A. Letcher, D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. *Differentiable game mechanics*. Journal of Machine Learning Research, 20(84):1—40, 2019.
29. W. Liao, M. Hong, H. Farmanbar, and Z.-Q. Luo. *Semi-asynchronous routing for large scale hierarchical networks*. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2894–2898, 2015.
30. Q. Lin, M. Liu, H. Rafique, and T. Yang. *Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality*. ArXiv preprint arXiv:1810.10207, 2018.
31. T. Lin, C. Jin and M. Jordan. *On gradient descent ascent for nonconvex-concave minimax problems*. International Conference on Machine Learning, PMLR, pages 6083–6093, 2020.
32. T. Lin, C. Jin and M. Jordan. *Near-optimal algorithms for minimax optimization*. Conference on Learning Theory, PMLR, pages 2738–2779, 2020.
33. S. Lu, I. Tsaknakis, M. Hong, and Y. Chen. *Hybrid block successive approximation for one-sided nonconvex min-max problems: Algorithms and applications*. IEEE Transactions on Signal Processing, 68:3676–3691, 2020.
34. G. Mateos, J. A. Bazerque, and G. B. Giannakis. *Distributed sparse linear regression*. IEEE Transactions on Signal Processing, 58(10):5262–5276, 2010.
35. E. V. Mazumdar, M. I. Jordan, and S. S. Sastry. *On finding local nash equilibria (and only local nash equilibria) in zero-sum games*. ArXiv preprint arXiv:1901.00838, 2019.
36. R. D. C. Monteiro and B. F. Svaiter. *On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean*. SIAM Journal on Optimization, 20(6):2755—2787, 2010.
37. R. D. C. Monteiro and B. F. Svaiter. *Complexity of variants of Tseng's modified F-B splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems*. SIAM Journal on Optimization, 21(4):1688—1720, 2011.
38. A. Nedic and A. Ozdaglar. *Subgradient methods for saddle-point problems*. Journal of Optimization Theory and Applications, 142(1):205–228, 2009.
39. A. Nemirovski. *Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems*. SIAM Journal on Optimization, 15(1):229—251, 2004.
40. Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2003.

41. Y. Nesterov. *Dual extrapolation and its applications to solving variational inequalities and related problems*. Mathematical Programming, 109(2-3):319—344, 2007.
42. Y. Nesterov. *Gradient methods for minimizing composite functions*. Mathematical Programming, 140(1):125–161, 2013.
43. M. Nouiehed, M. Sanjabi, T. Huang, J.D. Lee. *Solving a class of nonconvex min-max games using iterative first order methods*. Advances in Neural Information Processing Systems, 32:14934–14942, 2019.
44. Y. Ouyang, Y. Chen, G. Lan, and E. Pasiliao Jr. *An accelerated linearized alternating direction method of multipliers*. SIAM Journal on Imaging Sciences, 8(1):644–681, 2015.
45. Y. Ouyang and Y. Xu. *Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems*. Mathematical Programming, doi:10.1007/s10107-019-01420-0, 2019.
46. Q. Qian, S. Zhu, J. Tang, R. Jin, B. Sun, and H. Li. *Robust optimization over multiple domains*. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):4739–4746, 2019.
47. H. Rafique, M. Liu, Q. Lin, and T. Yang. *Nonconvex min-max optimization: provable algorithms and applications in machine learning*. ArXiv preprint arXiv:1810.02060, 2018.
48. M. Sanjabi, J. Ba, M. Razaviyayn, and J. D. Lee. *On the convergence and robustness of training gans with regularized optimal transport*. In Advances in Neural Information Processing Systems, pages 7091–7101, 2018.
49. K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. *Efficient algorithms for smooth minimax optimization*. In NeurIPS, pages 12659—12670, 2019.
50. P. Tseng. *On accelerated proximal gradient methods for convex-concave optimization*. submitted to SIAM Journal on Optimization, 2:3, 2008.
51. J. Yang, N. Kiyavash, and N. He. *Global convergence and variance-blueuced optimization for a class of nonconvex-nonconcave minimax problems*. ArXiv preprint arXiv:2002.09621, 2020.
52. L. Yann, B. Léon, B. Yoshua, and H. Patrick. *Gradient-based learning applied to document recognition*. Proceedings of the IEEE, 86(11):2278–2324, 1998.
53. B. Grimmer, H. Lu, P. Worah, and V. Mirrokni. *The landscape of nonconvex-nonconcave minimax optimization*. ArXiv preprint arXiv:2006.08667, 2020.
54. L. Mescheder, A. Geiger, and S. Nowozin. *Which training methods for GANs do actually converge?*. International conference on machine learning, pages 3481–3490, 2018.

A Proof of Theorem 5.1

We prove the following two lemmas before giving the proof of Theorem 5.1.

Lemma A.1 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.6). If $\hat{\eta} > L_{11}$, then we have

$$\begin{aligned} & l(x_{k+1}, y_{k+1}) - l(x_k, y_k) \\ \leq & -\left(\tfrac{\hat{\eta}}{2} - \tfrac{L_{12}^2\hat{\rho}}{2}\right)\|x_{k+1} - x_k\|^2 - \left(\tfrac{\hat{\mu}}{2} - \tfrac{1}{\hat{\rho}}\right)\|y_{k+1} - y_k\|^2 - \left(\hat{\mu} - \tfrac{1}{2\hat{\rho}} - \tfrac{\hat{\rho}L_{22}^2}{2}\right)\|y_k - y_{k-1}\|^2. \end{aligned} \tag{A.1}$$

Proof The optimality condition for y_k in (5.5) implies that $\forall y \in \mathcal{Y}$ and $\forall k \geq 1$,

$$\langle \nabla_y f(x_{k+1}, y_k) - \frac{1}{\hat{\rho}}(y_{k+1} - y_k) - \partial g_1(y_{k+1}), y - y_{k+1}\rangle \leq 0. \tag{A.2}$$

By choosing $y = y_k$ in (A.2), we have

$$\langle \nabla_y f(x_{k+1}, y_k) - \frac{1}{\hat{\rho}}(y_{k+1} - y_k) - \partial g_1(y_{k+1}), y_k - y_{k+1}\rangle \leq 0. \tag{A.3}$$

On the other hand, by replacing k with $k-1$ and choosing $y = y_{k+1}$ in (A.2), we obtain

$$\langle \nabla_y f(x_k, y_{k-1}) - \frac{1}{\hat{\rho}}(y_k - y_{k-1}) - \partial g_1(y_k), y_{k+1} - y_k\rangle \leq 0, \tag{A.4}$$

which, in view of the fact that $f(x, y)$ is $\hat{\mu}$-strongly concave w.r.t. y for any given $x \in \mathcal{X}$ and $g_1(y)$ is convex, then implies that

$$\begin{aligned} & l(x_{k+1}, y_{k+1}) - l(x_{k+1}, y_k) \\ \leq & \langle \nabla_y f(x_{k+1}, y_k) - \partial g_1(y_k), y_{k+1} - y_k\rangle - \frac{\hat{\mu}}{2}\|y_{k+1} - y_k\|^2 \\ \leq & \langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_{k-1}), y_{k+1} - y_k\rangle + \frac{1}{\hat{\rho}}\langle y_k - y_{k-1}, y_{k+1} - y_k\rangle - \frac{\hat{\mu}}{2}\|y_{k+1} - y_k\|^2. \end{aligned} \tag{A.5}$$

Denoting $v_{k+1} := (y_{k+1} - y_k) - (y_k - y_{k-1})$, we can write the first inner product term in the r.h.s. of (A.5) as

$$\begin{aligned} & \langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_{k-1}), y_{k+1} - y_k\rangle \\ = & \langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_k), y_{k+1} - y_k\rangle + \langle \nabla_y f(x_k, y_k) - \nabla_y f(x_k, y_{k-1}), v_{k+1}\rangle \\ & + \langle \nabla_y f(x_k, y_k) - \nabla_y f(x_k, y_{k-1}), y_k - y_{k-1}\rangle. \end{aligned} \tag{A.6}$$

Next, we estimate the three terms in the right hand side of (A.6) respectively. By Assumption 2.1 and the Cauchy-Schwarz inequality, we can bound the first two terms according to

$$\langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_k), y_{k+1} - y_k \rangle \leq \frac{L_{12}^2 \hat{\rho}}{2} \|x_{k+1} - x_k\|^2 + \frac{1}{2\hat{\rho}} \|y_{k+1} - y_k\|^2, \tag{A.7}$$

and

$$\langle \nabla_y f(x_k, y_k) - \nabla_y f(x_k, y_{k-1}), v_{k+1} \rangle \leq \frac{\hat{\rho} L_{22}^2}{2} \|y_k - y_{k-1}\|^2 + \frac{1}{2\hat{\rho}} \|v_{k+1}\|^2. \tag{A.8}$$

For the third term, by $\hat{\mu}$-strongly-concavity of f with respect to y,

$$\langle \nabla_y f(x_k, y_k) - \nabla_y f(x_k, y_{k-1}), y_k - y_{k-1} \rangle \leq -\hat{\mu} \|y_k - y_{k-1}\|^2. \tag{A.9}$$

Moreover, it can be easily checked that

$$\langle y_k - y_{k-1}, y_{k+1} - y_k \rangle = \frac{1}{2} \|y_k - y_{k-1}\|^2 + \frac{1}{2} \|y_{k+1} - y_k\|^2 - \frac{1}{2} \|v_{k+1}\|^2. \tag{A.10}$$

Plugging (A.6)-(A.10) into (A.5) and rearranging the terms, we conclude that

$$\begin{aligned} & l(x_{k+1}, y_{k+1}) - l(x_{k+1}, y_k) \\ \leq & \frac{L_{12}^2 \hat{\rho}}{2} \|x_{k+1} - x_k\|^2 - (\hat{\mu} - \frac{1}{2\hat{\rho}} - \frac{\hat{\rho} L_{22}^2}{2}) \|y_k - y_{k-1}\|^2 - (\frac{\hat{\mu}}{2} - \frac{1}{\hat{\rho}}) \|y_{k+1} - y_k\|^2. \end{aligned} \tag{A.11}$$

The proof is completed by combining (A.11) with (5.7) in Lemma 5.1. ∎

Lemma A.2 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.6). Also let us denote

$$l_{k+1} := l(x_{k+1}, y_{k+1}), \ S_{k+1} := \frac{2}{\hat{\rho}^2 \hat{\mu}} \|y_{k+1} - y_k\|^2,$$
$$F_{k+1} := l_{k+1} + S_{k+1} - (\hat{\mu} + \frac{7}{2\hat{\rho}} - \frac{\hat{\rho} L_{22}^2}{2} - \frac{2L_{22}^2}{\hat{\mu}}) \|y_{k+1} - y_k\|^2.$$

If $\hat{\eta} > L_{11}$, then

$$F_{k+1} - F_k \leq -\left(\frac{\hat{\eta}}{2} - \frac{\hat{\rho} L_{12}^2}{2} - \frac{2L_{12}^2}{\hat{\rho} \hat{\mu}^2} \right) \|x_{k+1} - x_k\|^2 - \left(\frac{3\hat{\mu} - \hat{\rho} L_{22}^2}{2} + \frac{\hat{\mu} - 4\hat{\rho} L_{22}^2}{2\hat{\rho}\hat{\mu}} \right) \|y_{k+1} - y_k\|^2. \tag{A.12}$$

Proof First by (A.3) and (A.4), we have

$$\frac{1}{\hat{\rho}} \langle v_{k+1}, y_{k+1} - y_k \rangle \leq \langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_{k-1}) - \partial g_1(y_{k+1}) + \partial g_1(y_k), y_{k+1} - y_k \rangle, \tag{A.13}$$

By the $\langle \partial g_1(y_{k+1}) - \partial g_1(y_k), y_{k+1} - y_k \rangle \geq 0$, and together with (A.6) then imply that

$$\begin{aligned} \frac{1}{\hat{\rho}} \langle v_{k+1}, y_{k+1} - y_k \rangle \leq & \langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_k), y_{k+1} - y_k \rangle \\ & + \langle \nabla_y f(x_k, y_k) - \nabla_y f(x_k, y_{k-1}), v_{k+1} \rangle \\ & + \langle \nabla_y f(x_k, y_k) - \nabla_y f(x_k, y_{k-1}), y_k - y_{k-1} \rangle. \end{aligned} \tag{A.14}$$

Similar to (A.7), we can easily see that

$$\langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_k), y_{k+1} - y_k \rangle \leq \frac{L_{12}^2}{2\hat{\mu}} \|x_{k+1} - x_k\|^2 + \frac{\hat{\mu}}{2} \|y_{k+1} - y_k\|^2. \tag{A.15}$$

By plugging (A.8),(A.9),(A.15) into (A.14), and using the identity $\frac{1}{\hat{\rho}} \langle v_{k+1}, y_{k+1} - y_k \rangle = \frac{1}{2\hat{\rho}} \|y_{k+1} - y_k\|^2 + \frac{1}{2\hat{\rho}} \|v_{k+1}\|^2 - \frac{1}{2\hat{\rho}} \|y_k - y_{k-1}\|^2$, we conclude that

$$\begin{aligned} & \frac{1}{2\hat{\rho}} \|y_{k+1} - y_k\|^2 + \frac{1}{2\hat{\rho}} \|v_{k+1}\|^2 - \frac{1}{2\hat{\rho}} \|y_k - y_{k-1}\|^2 \\ \leq & \frac{L_{12}^2}{2\hat{\mu}} \|x_{k+1} - x_k\|^2 + \frac{\hat{\mu}}{2} \|y_{k+1} - y_k\|^2 + \frac{\hat{\rho} L_{22}^2}{2} \|y_k - y_{k-1}\|^2 + \frac{1}{2\hat{\rho}} \|v_{k+1}\|^2 - \hat{\mu} \|y_k - y_{k-1}\|^2. \end{aligned} \tag{A.16}$$

Rearranging the terms of (A.16), we have

$$\begin{aligned} & \frac{1}{2\hat{\rho}} \|y_{k+1} - y_k\|^2 - \frac{1}{2\hat{\rho}} \|y_k - y_{k-1}\|^2 \\ \leq & \frac{L_{12}^2}{2\hat{\mu}} \|x_{k+1} - x_k\|^2 + \frac{\hat{\mu}}{2} \|y_{k+1} - y_k\|^2 - \left(\hat{\mu} - \frac{\hat{\rho} L_{22}^2}{2} \right) \|y_k - y_{k-1}\|^2. \end{aligned} \tag{A.17}$$

Multiplying $\frac{4}{\hat{\rho}\hat{\mu}}$ on both sides of (A.17) and using the definition of S_{k+1}, we obtain

$$S_{k+1} - S_k \leq \frac{2L_{12}^2}{\hat{\mu}^2\hat{\rho}}\|x_{k+1} - x_k\|^2 + \frac{2}{\hat{\rho}}\|y_{k+1} - y_k\|^2 - \left(\frac{4}{\hat{\rho}} - \frac{2L_{22}^2}{\hat{\mu}}\right)\|y_k - y_{k-1}\|^2.$$

It then follows from (A.1) in Lemma A.1 and the definition of F_k that

$$F_{k+1} - F_k \leq -\left(\frac{\hat{\eta}}{2} - \frac{\hat{\rho}L_{12}^2}{2} - \frac{2L_{12}^2}{\hat{\rho}\hat{\mu}^2}\right)\|x_{k+1} - x_k\|^2 - \left(\frac{3\hat{\mu} - \hat{\rho}L_{22}^2}{2} + \frac{\hat{\mu} - 4\hat{\rho}L_{22}^2}{2\hat{\rho}\hat{\mu}}\right)\|y_{k+1} - y_k\|^2.$$

■

The proof of Theorem 5.1

Proof Noting that (5.4) is equivalent to $x_{k+1}^{(i)} = \mathrm{Prox}_{h_i,\mathcal{X}_i}^{\hat{\eta}}\left(x_k^{(i)} - \frac{1}{\hat{\eta}}\nabla_{x^{(i)}} f\left(v_{k+1}^{(i)}, y_k\right)\right)$, we immediately obtain

$$\begin{aligned}
&\|(\nabla\mathcal{G}_k)_{x^{(i)}}\| \\
\leq&\hat{\eta}\|\mathrm{Prox}_{h_i,\mathcal{X}_i}^{\hat{\eta}}\left(x_k^{(i)} - \frac{1}{\hat{\eta}}\nabla_{x^{(i)}} f\left(v_{k+1}^{(i)}, y_k\right)\right) - \mathrm{Prox}_{h_i,\mathcal{X}_i}^{\hat{\eta}}\left(x_k^{(i)} - \frac{1}{\hat{\eta}}\nabla_{x^{(i)}} f\left(x_k, y_k\right)\right)\| \\
&+ \hat{\eta}\|x_{k+1}^{(i)} - x_k^{(i)}\| \\
\leq&\hat{\eta}\|x_{k+1}^{(i)} - x_k^{(i)}\| + \|\nabla_{x^{(i)}} f\left(v_{k+1}^{(i)}, y_k\right) - \nabla_{x^{(i)}} f\left(x_k, y_k\right)\| \\
\leq&\hat{\eta}\|x_{k+1}^{(i)} - x_k^{(i)}\| + L_{11}\|v_{k+1}^{(i)} - x_k\| \\
\leq&\left(\hat{\eta} + L_{11}\right)\|x_{k+1} - x_k\|,
\end{aligned} \tag{A.18}$$

where the second inequality holds by the nonexpansiveness of the proximal operator $\mathrm{Prox}_{h_i,\mathcal{X}_i}^{\hat{\eta}}$, the last inequality hold since $\|x_{k+1}^{(i)} - x_k^{(i)}\| \leq \|x_{k+1} - x_k\|$, $\|v_{k+1}^{(i)} - x_k\| \leq \|x_{k+1} - x_k\|$ by the definition of $x^{(i)}$ and $v_{k+1}^{(i)}$. On the other hand, since (5.5) is equivalent to $y_{k+1} = \mathrm{Prox}_{g_1,\mathcal{Y}}^{1/\hat{\rho}}\left(y_k + \hat{\rho}\nabla_y f\left(x_{k+1}, y_k\right)\right)$, we conclude from the triangle inequality that

$$\begin{aligned}
&\|(\nabla\mathcal{G}_k)_y\| \\
\leq&\frac{1}{\hat{\rho}}\|\mathrm{Prox}_{g_1,\mathcal{Y}}^{1/\hat{\rho}}\left(y_k + \hat{\rho}\nabla_y f\left(x_{k+1}, y_k\right)\right) - \mathrm{Prox}_{g_1,\mathcal{Y}}^{1/\hat{\rho}}\left(y_k + \hat{\rho}\nabla_y f\left(x_k, y_k\right)\right)\| \\
&+ \frac{1}{\hat{\rho}}\|y_{k+1} - y_k\| \\
\leq&\frac{1}{\hat{\rho}}\|y_{k+1} - y_k\| + L_{12}\|x_{k+1} - x_k\|.
\end{aligned} \tag{A.19}$$

By combining (A.18)-(A.19), and using Cauchy-Schwarz inequality, we obtain

$$\|\nabla\mathcal{G}_k\|^2 \leq \left(K_1\left(\hat{\eta} + L_{11}\right)^2 + 2L_{12}^2\right)\|x_{k+1} - x_k\|^2 + \frac{2}{\hat{\rho}^2}\|y_{k+1} - y_k\|^2. \tag{A.20}$$

Observing that $r_1 > 0$. Multiplying both sides of (A.20) by r_1, and using (A.12) in Lemma A.2, we have

$$r_1\|\nabla\mathcal{G}_k\|^2 \leq F_k - F_{k+1}. \tag{A.21}$$

Summing up the above inequalities from $k = 1$ to $k = \mathcal{T}(\varepsilon)$ and using the definition of L_1, we obtain

$$\sum_{k=1}^{\mathcal{T}(\varepsilon)} r_1\|\nabla\mathcal{G}_k\|^2 \leq F_1 - F_{\mathcal{T}(\varepsilon)+1} \leq L_1 - F_{\mathcal{T}(\varepsilon)+1}. \tag{A.22}$$

Note that by the definition of F_{k+1} in Lemma A.2, we have

$$\begin{aligned}
F_{\mathcal{T}(\varepsilon)+1} &= l_{\mathcal{T}(\varepsilon)+1} + S_{\mathcal{T}(\varepsilon)+1} - (\hat{\mu} + \frac{7}{2\hat{\rho}} - \frac{\hat{\rho}L_{22}^2}{2} - \frac{2L_{22}^2}{\hat{\mu}})\|y_{\mathcal{T}(\varepsilon)+1} - y_{\mathcal{T}(\varepsilon)}\|^2 \\
&\geq \underline{l} - (\hat{\mu} + \frac{7}{2\hat{\rho}} - \frac{\hat{\rho}L_{22}^2}{2} - \frac{2L_{22}^2}{\hat{\mu}})\delta_y^2 = \underline{L},
\end{aligned}$$

where the inequality follows from the definitions of $\underline{l}$ and δ_y, and the facts that $S_k \geq 0$ ($\forall k \geq 1$) and $\hat{\mu} + \frac{7}{2\hat{\rho}} - \frac{\hat{\rho}L_{22}^2}{2} - \frac{2L_{22}^2}{\hat{\mu}} \geq 0$ due to the selection of $\hat{\rho}$. We then conclude from (A.22) that

$$\sum_{k=1}^{\mathcal{T}(\varepsilon)} r_1\|\nabla\mathcal{G}_k\|^2 \leq L_1 - F_{\mathcal{T}(\varepsilon)+1} \leq L_1 - \underline{L},$$

which, in view of the definition of $\mathcal{T}(\varepsilon)$, implies that $\varepsilon^2 \leq (L_1 - \underline{L})/(\mathcal{T}(\varepsilon)\cdot r_1)$ or equivalently, $\mathcal{T}(\varepsilon) \leq (L_1 - \underline{L})/(r_1\varepsilon^2)$. ■

B Proof of Theorem 5.2

We prove the following two lemmas before giving the proof of Theorem 5.2.

Lemma B.1 Suppose that Assumptions 2.1 and 5.1 hold. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.11). If $\forall k, \vartheta_k > L_{11}$ and $\xi \leq \frac{2}{L'_{22}+\hat{c}_1}$, then

$$\begin{aligned} & l(x_{k+1}, y_{k+1}) - l(x_k, y_k) \\ \leq & -\left(\iota + \frac{\vartheta_k}{2} - \frac{\xi L_{12}^2}{2}\right)\|x_{k+1} - x_k\|^2 + \frac{1}{\xi}\|y_{k+1} - y_k\|^2 + \frac{1}{2\xi}\|y_k - y_{k-1}\|^2 \\ & + \frac{\hat{c}_{k-1}}{2}(\|y_{k+1}\|^2 - \|y_k\|^2). \end{aligned} \tag{B.1}$$

Proof The optimality condition for y_k in (5.5) implies that, $\forall y \in \mathcal{Y}$, $\forall k \geq 1$,

$$\langle \nabla_y \bar{f}_k(x_{k+1}, y_k) - \frac{1}{\xi}(y_{k+1} - y_k) - \partial g_1(y_{k+1}), y - y_{k+1}\rangle \leq 0. \tag{B.2}$$

By choosing $y = y_k$ in (B.2), we have

$$\langle \nabla_y \bar{f}_k(x_{k+1}, y_k) - \frac{1}{\xi}(y_{k+1} - y_k) - \partial g_1(y_{k+1}), y_k - y_{k+1}\rangle \leq 0. \tag{B.3}$$

On the other hand, by replacing k with $k-1$ and choosing $y = y_{k+1}$ in (B.2), we obtain

$$\langle \nabla_y \bar{f}_{k-1}(x_k, y_{k-1}) - \frac{1}{\xi}(y_k - y_{k-1}) - \partial g_1(y_k), y_{k+1} - y_k\rangle \leq 0. \tag{B.4}$$

The concavity of $\bar{f}_k(x_{k+1}, y)$ w.r.t. y together with (B.4) and $g_1(y)$ is convex, then imply that

$$\begin{aligned} & \bar{l}_k(x_{k+1}, y_{k+1}) - \bar{l}_k(x_{k+1}, y_k) \\ \leq & \langle \nabla_y \bar{f}_k(x_{k+1}, y_k) - \nabla_x \bar{f}_{k-1}(x_k, y_{k-1}), y_{k+1} - y_k\rangle + \frac{1}{\xi}\langle y_k - y_{k-1}, y_{k+1} - y_k\rangle \\ = & \langle \nabla_y \bar{f}_k(x_{k+1}, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_k), y_{k+1} - y_k\rangle + \langle \nabla_y \bar{f}_{k-1}(x_k, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1}), v_{k+1}\rangle \\ & + \langle \nabla_y \bar{f}_{k-1}(x_k, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1}), y_k - y_{k-1}\rangle + \frac{1}{\xi}\langle y_k - y_{k-1}, y_{k+1} - y_k\rangle, \end{aligned} \tag{B.5}$$

where $v_{k+1} = y_{k+1} - y_k - (y_k - y_{k-1})$. We now provide bounds on the inner product terms of (B.5). Firstly, by the definition of $\bar{f}_k(x_{k+1}, y_k)$ and $\bar{f}_{k-1}(x_k, y_k)$, Assumptions 2.1 and 5.1, and the Cauchy-Schwarz inequality, we have

$$\begin{aligned} & \langle \nabla_y \bar{f}_k(x_{k+1}, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_k), y_{k+1} - y_k\rangle \\ = & \langle \nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_k), y_{k+1} - y_k\rangle - (\hat{c}_k - \hat{c}_{k-1})\langle y_k, y_{k+1} - y_k\rangle \\ \leq & \frac{\xi L_{12}^2}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{2\xi}\|y_{k+1} - y_k\|^2 - \frac{\hat{c}_k - \hat{c}_{k-1}}{2}(\|y_{k+1}\|^2 - \|y_k\|^2) \\ & + \frac{\hat{c}_k - \hat{c}_{k-1}}{2}\|y_{k+1} - y_k\|^2 \\ \leq & \frac{\xi L_{12}^2}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{2\xi}\|y_{k+1} - y_k\|^2 - \frac{\hat{c}_k - \hat{c}_{k-1}}{2}(\|y_{k+1}\|^2 - \|y_k\|^2). \end{aligned} \tag{B.6}$$

Secondly, by the Cauchy-Schwarz inequality,

$$\langle \nabla_y \bar{f}_{k-1}(x_k, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1}), v_{k+1}\rangle \leq \frac{\xi}{2}\|\nabla_y \bar{f}_{k-1}(x_k, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1})\|^2 + \frac{1}{2\xi}\|v_{k+1}\|^2. \tag{B.7}$$

Thirdly, it follows from the concavity of $\bar{f}_{k-1}(x_k, y)$ w.r.t. y that

$$\begin{aligned} & \langle \nabla_y \bar{f}_{k-1}(x_k, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1}), y_k - y_{k-1}\rangle \\ \leq & -\frac{1}{L'_{22} + \hat{c}_{k-1}}\|\nabla_y \bar{f}_{k-1}(x_k, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1})\|^2 - \frac{\hat{c}_{k-1} L'_{22}}{L'_{22} + \hat{c}_{k-1}}\|y_k - y_{k-1}\|^2 \\ \leq & -\frac{1}{L'_{22} + \hat{c}_{k-1}}\|\nabla_y \bar{f}_{k-1}(x_k, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1})\|^2. \end{aligned} \tag{B.8}$$

Also observe that

$$\frac{1}{\xi}\langle y_{k+1} - y_k, y_k - y_{k-1}\rangle = \frac{1}{2\xi}\|y_{k+1} - y_k\|^2 + \frac{1}{2\xi}\|y_k - y_{k-1}\|^2 - \frac{1}{2\xi}\|v_{k+1}\|^2. \tag{B.9}$$

Plugging (B.6)-(B.9) into (B.5), and using the definition of $\bar{l}_k(x_{k+1}, y_{k+1})$ and $\bar{l}_k(x_{k+1}, y_k)$ and the assumption $\frac{\xi}{2} \leq \frac{1}{L_{22}^{'}+\hat{c}_1}$, we obtain

$$l(x_{k+1}, y_{k+1}) - l(x_{k+1}, y_k) \leq \frac{\xi L_{12}^2}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{\xi}\|y_{k+1} - y_k\|^2 + \frac{1}{2\xi}\|y_k - y_{k-1}\|^2 + \frac{\hat{c}_{k-1}}{2}(\|y_{k+1}\|^2 - \|y_k\|^2). \tag{B.10}$$

The result in (B.1) the follows by adding (B.10) and (5.12) in Lemma 5.2. ∎

Lemma B.2 Suppose that Assumptions 2.1 and 5.1 hold. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.11). Also let us denote

$$\mathcal{S}_{k+1} := \frac{8}{\xi^2 \hat{c}_{k+1}}\|y_{k+1} - y_k\|^2 + \frac{8}{\xi}\left(1 - \frac{\hat{c}_k}{\hat{c}_{k+1}}\right)\|y_{k+1}\|^2,$$
$$\mathcal{F}_{k+1} := l(x_{k+1}, y_{k+1}) + \mathcal{S}_{k+1} - \frac{15}{2\xi}\|y_{k+1} - y_k\|^2 - \frac{\hat{c}_k}{2}\|y_{k+1}\|^2.$$

If

$$\vartheta_k > L_{11},\ \frac{1}{\hat{c}_{k+1}} - \frac{1}{\hat{c}_k} \leq \frac{\xi}{5},\ \xi \leq \frac{2}{L_{22}^{'} + \hat{c}_1}, \tag{B.11}$$

then $\forall k \geq 2$,

$$\mathcal{F}_{k+1} - \mathcal{F}_k \leq -\left(\iota + \frac{\vartheta_k}{2} - \frac{\xi L_{12}^2}{2} - \frac{16 L_{12}^2}{\xi \hat{c}_k^2}\right)\|x_{k+1} - x_k\|^2 + \frac{\hat{c}_{k-1} - \hat{c}_k}{2}\|y_{k+1}\|^2 \tag{B.12}$$
$$- \frac{9}{10\xi}\|y_{k+1} - y_k\|^2 + \frac{8}{\xi}\left(\frac{\hat{c}_{k-1}}{\hat{c}_k} - \frac{\hat{c}_k}{\hat{c}_{k+1}}\right)\|y_{k+1}\|^2. \tag{B.13}$$

Proof By (B.3) and (B.4), we have

$$\frac{1}{\xi}\langle v_{k+1}, y_{k+1} - y_k\rangle \leq \langle \nabla_y \bar{f}_k(x_{k+1}, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1}) - \partial g_1(y_{k+1}) + \partial g_1(y_k), y_{k+1} - y_k\rangle. \tag{B.14}$$

By the $\langle \partial g_1(y_{k+1}) - \partial g_1(y_k), y_{k+1} - y_k\rangle \geq 0$, similar to (B.5) in Lemma B.1, (B.14) can be rewritten as

$$\begin{aligned}
&\frac{1}{\xi}\langle v_{k+1}, y_{k+1} - y_k\rangle \\
\leq &\langle \nabla_y \bar{f}_k(x_{k+1}, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_k), y_{k+1} - y_k\rangle + \langle \nabla_y \bar{f}_{k-1}(x_k, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1}), v_{k+1}\rangle \\
&+ \langle \nabla_y \bar{f}_{k-1}(x_k, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1}), y_k - y_{k-1}\rangle.
\end{aligned}$$

Using an argument similar to the proof of (B.6)-(B.9), the concavity of $\bar{f}_{k-1}(x_k, y)$ w.r.t. y, and the Cauchy-Schwarz inequality, we conclude from the above inequality that

$$\begin{aligned}
&\frac{1}{2\xi}\|y_{k+1} - y_k\|^2 + \frac{1}{2\xi}\|v_{k+1}\|^2 - \frac{1}{2\xi}\|y_k - y_{k-1}\|^2 \\
\leq &\frac{L_{12}^2}{2a_k}\|x_{k+1} - x_k\|^2 + \frac{a_k}{2}\|y_{k+1} - y_k\|^2 - \frac{\hat{c}_k - \hat{c}_{k-1}}{2}(\|y_{k+1}\|^2 - \|y_k\|^2) \\
&+ \frac{\xi}{2}\|\nabla_y \bar{f}_{k-1}(x_k, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1})\|^2 - \frac{1}{L_{22}^{'} + \hat{c}_{k-1}}\|\nabla_y \bar{f}_{k-1}(x_k, y_k) - \nabla_y \bar{f}_{k-1}(x_k, y_{k-1})\|^2 \\
&+ \frac{1}{2\xi}\|v_{k+1}\|^2 - \frac{\hat{c}_{k-1} L_{22}^{'}}{L_{22}^{'} + \hat{c}_{k-1}}\|y_k - y_{k-1}\|^2 + \frac{\hat{c}_k - \hat{c}_{k-1}}{2}\|y_{k+1} - y_k\|^2,
\end{aligned} \tag{B.15}$$

for any $a_k > 0$. Observing by $\hat{c}_1 \leq L_{22}^{'}$, and Assumption 5.1, we have

$$-\frac{\hat{c}_{k-1} L_{22}^{'}}{\hat{c}_{k-1} + L_{22}^{'}} \leq -\frac{\hat{c}_{k-1} L_{22}^{'}}{2L_{22}^{'}} = -\frac{\hat{c}_{k-1}}{2} \leq -\frac{\hat{c}_k}{2}.$$

Combining $\xi \leq \frac{2}{L_{22}^{'}+\hat{c}_1}$ and rearranging the terms in (B.15), we obtain

$$\begin{aligned}
&\frac{1}{2\xi}\|y_{k+1} - y_k\|^2 + \frac{\hat{c}_k - \hat{c}_{k-1}}{2}\|y_{k+1}\|^2 \\
\leq &\frac{1}{2\xi}\|y_k - y_{k-1}\|^2 + \frac{\hat{c}_k - \hat{c}_{k-1}}{2}\|y_k\|^2 + \frac{L_{12}^2}{2a_k}\|x_{k+1} - x_k\|^2 + \frac{a_k}{2}\|y_{k+1} - y_k\|^2 \\
&- \frac{\hat{c}_k}{2}\|y_k - y_{k-1}\|^2.
\end{aligned}$$

By multiplying $\frac{16}{\xi\hat{c}_k}$ on both sides of the above inequality, we then obtain

$$\begin{aligned}&\frac{8}{\xi^2\hat{c}_k}\|y_{k+1}-y_k\|^2+\frac{8}{\xi}\left(1-\frac{\hat{c}_{k-1}}{\hat{c}_k}\right)\|y_{k+1}\|^2\\ \leq&\frac{8}{\xi^2\hat{c}_k}\|y_k-y_{k-1}\|^2+\frac{8}{\xi}(1-\frac{\hat{c}_{k-1}}{\hat{c}_k})\|y_k\|^2+\frac{8L_{12}^2}{\xi\hat{c}_k a_k}\|x_{k+1}-x_k\|^2\\ &+\frac{8a_k}{\xi\hat{c}_k}\|y_{k+1}-y_k\|^2-\frac{8}{\xi}\|y_k-y_{k-1}\|^2.\end{aligned}\tag{B.16}$$

Setting $a_k=\frac{\hat{c}_k}{2}$ in the above inequality, and using the definition of $\mathcal{S}_{k+1}$ and (B.11), we have

$$\begin{aligned}\mathcal{S}_{k+1}-\mathcal{S}_k\leq&\frac{8}{\xi}\left(\frac{\hat{c}_{k-1}}{\hat{c}_k}-\frac{\hat{c}_k}{\hat{c}_{k+1}}\right)\|y_{k+1}\|^2+\frac{16L_{12}^2}{\xi\hat{c}_k^2}\|x_{k+1}-x_k\|^2+\frac{28}{5\xi}\|y_{k+1}-y_k\|^2\\ &-\frac{8}{\xi}\|y_k-y_{k-1}\|^2.\end{aligned}\tag{B.17}$$

Combining (B.17) and (B.1) in Lemma B.1, and using the definition of $\mathcal{F}_{k+1}$, we conclude

$$\begin{aligned}\mathcal{F}_{k+1}-\mathcal{F}_k\leq&-\left(\iota+\frac{\vartheta_k}{2}-\frac{\xi L_{12}^2}{2}-\frac{16L_{12}^2}{\xi\hat{c}_k^2}\right)\|x_{k+1}-x_k\|^2+\frac{\hat{c}_{k-1}-\hat{c}_k}{2}\|y_{k+1}\|^2\\ &-\frac{9}{10\xi}\|y_{k+1}-y_k\|^2+\frac{8}{\xi}\left(\frac{\hat{c}_{k-1}}{\hat{c}_k}-\frac{\hat{c}_k}{\hat{c}_{k+1}}\right)\|y_{k+1}\|^2.\end{aligned}$$

■

The proof of Theorem 5.2

Proof By $\xi\leq\frac{1}{10L_{22}}$ and $\hat{c}_k=\frac{19}{20\xi k^{1/4}},\forall k\geq 1$, let us denote $\vartheta_k=\xi L_{12}^2+\frac{16\tau L_{12}^2}{\xi\hat{c}_k^2}-2\iota$, $\alpha_k=\frac{8(\tau-2)L_{12}^2}{\xi\hat{c}_k^2}$, we can easily see that the relations in (B.11) are satisfied. It follows from the selection of ϑ_k and α_k that

$$\iota+\frac{\vartheta_k}{2}-\frac{\xi L_{12}^2}{2}-\frac{16L_{12}^2}{\xi\hat{c}_k^2}=\alpha_k.$$

This observation, in view of Lemma B.2, then immediately implies that

$$\begin{aligned}\alpha_k\|x_{k+1}-x_k\|^2+\frac{9}{10\xi}\|y_{k+1}-y_k\|^2\leq&\mathcal{F}_k-\mathcal{F}_{k+1}+\frac{8}{\xi}\left(\frac{\hat{c}_{k-1}}{\hat{c}_k}-\frac{\hat{c}_k}{\hat{c}_{k+1}}\right)\|y_{k+1}\|^2\\ &+\frac{\hat{c}_{k-1}-\hat{c}_k}{2}\|y_{k+1}\|^2.\end{aligned}\tag{B.18}$$

We can easily check from the definition of $\bar{f}_k(x_k,y_k)$ that

$$\|\nabla\mathcal{G}_k\|-\|\nabla\bar{\mathcal{G}}_k\|\leq\hat{c}_k\|y_k\|.$$

By replacing f with $\bar{f}_k$, $\hat{\eta}$ with $\vartheta_k+\iota$, similar to (A.18) and (A.19), we immediately obtain that

$$\|(\nabla\bar{\mathcal{G}}_k)_{x^{(i)}}\|\leq(\vartheta_k+\iota+L_{11})\|x_{k+1}-x_k\|,\tag{B.19}$$

and

$$\|(\nabla\bar{\mathcal{G}}_k)_y\|\leq\frac{1}{\xi}\|y_{k+1}-y_k\|+L_{12}\|x_{k+1}-x_k\|.\tag{B.20}$$

Combining (B.19) and (B.20), and using the Cauchy-Schwarz inequality, we have

$$\|\nabla\bar{\mathcal{G}}_k\|^2\leq\left(K_1\left(\vartheta_k+\iota+L_{11}\right)^2+2L_{12}^2\right)\|x_{k+1}-x_k\|^2+\frac{2}{\xi^2}\|y_{k+1}-y_k\|^2.\tag{B.21}$$

Since both α_k and ϑ_k are in the same order when k becomes large enough, it then follows from the definition of r_1 that $\forall k\geq 1$,

$$\begin{aligned}\frac{K_1\left(\vartheta_k+\iota+L_{11}\right)^2+2L_{12}^2}{\alpha_k^2}=&\frac{K_1\left(\xi L_{12}^2+\frac{16\tau L_{12}^2}{\xi(\hat{c}_k)^2}-\iota+L_{11}\right)^2+2L_{12}^2}{\alpha_k^2}\\ \leq&\frac{2K_1\left(\frac{16\tau L_{12}^2}{\xi(\hat{c}_k)^2}\right)^2+2K_1\left(\xi L_{12}^2-\iota+L_{11}\right)^2+2L_{12}^2}{\alpha_k^2}\\ =&\frac{8K_1\tau^2}{(\tau-2)^2}+\frac{19^4\left(2K_1\left(\xi L_{12}^2-\iota+L_{11}\right)^2+2L_{12}^2\right)}{64\cdot 20^4\xi^2(\tau-2)^2L_{12}^4k}\\ \leq&\, r_1\end{aligned}\tag{B.22}$$

Combining the previous two inequalities in (B.22) and (B.21), we obtain

$$\|\nabla\bar{\mathcal{G}}_k\|^2 \le r_1(\alpha_k)^2\|x_{k+1}-x_k\|^2 + \frac{2}{\xi^2}\|y_{k+1}-y_k\|^2. \tag{B.23}$$

Denote $r_k^{(2)} = \frac{1}{\max\left\{r_1\alpha_k, \frac{20}{9\xi}\right\}}$. By multiplying $r_k^{(2)}$ on the both sides of (B.23), and using (B.18), we have

$$r_k^{(2)}\|\nabla\bar{\mathcal{G}}_k\|^2 \le \mathcal{F}_k - \mathcal{F}_{k+1} + \frac{8}{\xi}\left(\frac{\hat{c}_{k-1}}{\hat{c}_k} - \frac{\hat{c}_k}{\hat{c}_{k+1}}\right)\|y_{k+1}\|^2 + \frac{\hat{c}_{k-1}-\hat{c}_k}{2}\|y_{k+1}\|^2, \tag{B.24}$$

where the last inequality follows since $r_k^{(2)}$ is a decreasing sequence. Denoting

$$\bar{\mathcal{T}}(\varepsilon) := \min\{k \mid \|\nabla\bar{\mathcal{G}}(x_k,y_k)\| \le \frac{\varepsilon}{2}, k\ge 2\},$$

Summing both sides of (B.24) from $k=2$ to $k=\bar{\mathcal{T}}(\varepsilon)$, we then obtain

$$\begin{aligned}
&\sum_{k=2}^{\bar{\mathcal{T}}(\varepsilon)} r_k^{(2)}\|\nabla\bar{\mathcal{G}}_k\|^2 \\
\le& \mathcal{F}_2 - \mathcal{F}_{\bar{\mathcal{T}}(\varepsilon)} + \frac{8}{\xi}\left(\frac{\hat{c}_1}{\hat{c}_2} - \frac{\hat{c}_{\bar{\mathcal{T}}(\varepsilon)}}{\hat{c}_{\bar{\mathcal{T}}(\varepsilon)+1}}\right)\hat{\delta}_y^2 + \frac{\hat{c}_1-\hat{c}_{\bar{\mathcal{T}}(\varepsilon)}}{2}\hat{\delta}_y^2 \\
\le& \mathcal{F}_2 - \mathcal{F}_{\bar{\mathcal{T}}(\varepsilon)} + \frac{8\hat{c}_1}{\xi\hat{c}_2}\hat{\delta}_y^2 + \frac{\hat{c}_1}{2}\hat{\delta}_y^2 \\
=& \mathcal{F}_2 - \mathcal{F}_{\bar{\mathcal{T}}(\varepsilon)} + \left(8\cdot 2^{1/4} + \frac{19}{40}\right)\frac{\hat{\delta}_y^2}{\xi}.
\end{aligned} \tag{B.25}$$

Note that by the definition of $\mathcal{F}_{k+1}$ in Lemma B.2, we have

$$\begin{aligned}
\mathcal{F}_{\bar{\mathcal{T}}(\varepsilon)} &\ge \underline{l} - \frac{8\hat{\delta}_y^2}{\xi}\left(1+\frac{1}{\bar{\mathcal{T}}(\varepsilon)}\right)^{1/4} - \frac{15\hat{\delta}_y^2}{\xi} - \frac{\hat{c}_1\hat{\delta}_y^2}{2} \\
&= \underline{l} - \left(2^{13/4} + 15 + \frac{19}{40}\right)\frac{\hat{\delta}_y^2}{\xi} = \underline{\mathcal{L}},
\end{aligned}$$

where $\underline{l} := \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} l(x,y)$. By the definition of $\mathcal{L}_2$, we then conclude from (B.25) that

$$\begin{aligned}
\sum_{k=2}^{\bar{\mathcal{T}}(\varepsilon)} r_k^{(2)}\|\nabla\bar{\mathcal{G}}_k\|^2 &\le \mathcal{F}_2 - \underline{\mathcal{L}} + \left(8\cdot 2^{1/4} + \frac{19}{40}\right)\frac{\hat{\delta}_y^2}{\xi} \\
&\le \mathcal{L}_2 - \underline{\mathcal{L}} + \left(8\cdot 2^{1/4} + \frac{19}{40}\right)\frac{\hat{\delta}_y^2}{\xi} = r_2.
\end{aligned} \tag{B.26}$$

We can see from the selection of r_3 that $r_3 = \max\{r_1, \frac{20}{9\xi\alpha_2}\}$. Observe that α_k is an increasing sequence, when $k\ge 2$, we have that $r_3 \ge \max\{r_1, \frac{20}{9\xi\alpha_k}\}$, which implies that $r_k^{(2)} \ge \frac{1}{r_3\alpha_k}$, by multiplying r_3 on the both sides of (B.26), and combining the definition of r_2, we have $\sum_{k=2}^{\bar{\mathcal{T}}(\varepsilon)} \frac{1}{\alpha_k}\|\nabla\bar{\mathcal{G}}_k\|^2 \le r_2 r_3$, which, by the definition of $\bar{\mathcal{T}}(\varepsilon)$, implies that

$$\frac{\varepsilon^2}{4} \le \frac{r_2 r_3}{\sum_{k=2}^{\bar{\mathcal{T}}(\varepsilon)}\frac{1}{\alpha_k}}. \tag{B.27}$$

Note that when $\hat{c}_k = \frac{19}{20\xi k^{1/4}}$, $\alpha_k = \frac{2\cdot 40^2\xi(\tau-2)L_{12}^2\sqrt{k}}{19^2}$. By using the fact $\sum_{k=2}^{\bar{\mathcal{T}}(\varepsilon)} 1/\sqrt{k} \ge \sqrt{\bar{\mathcal{T}}(\varepsilon)} - 1$ and (B.27), we conclude $\frac{\varepsilon^2}{4} \le \frac{2\cdot 40^2\xi(\tau-2)L_{12}^2 r_2 r_3}{19^2\left(\sqrt{\bar{\mathcal{T}}(\varepsilon)}-1\right)}$ or equivalently,

$$\bar{T}(\varepsilon) \le \left(\frac{2\cdot 80^2\xi(\tau-2)L_{12}^2 r_2 r_3}{19^2\varepsilon^2} + 1\right)^2.$$

On the other hand, if $k \ge \frac{19^4\hat{\delta}_y^4}{10^4\xi^4\varepsilon^4}$, then $c_k = \frac{19}{20\xi k^{1/4}} \le \frac{\varepsilon}{2\hat{\delta}_y}$. This inequality together with the definition of $\hat{\delta}_y$ then imply that $\hat{c}_k\|y_k\| \le \frac{\varepsilon}{2}$. Therefore, there exists a

$$\begin{aligned}
\mathcal{T}(\varepsilon) &\le \max\left(\bar{\mathcal{T}}(\varepsilon), \frac{19^4\hat{\delta}_y^4}{10^4\xi^4\varepsilon^4}\right) \\
&\le \max\left(\left(\frac{2\cdot 80^2\xi(\tau-2)L_{12}^2 r_2 r_3}{19^2\varepsilon^2} + 1\right)^2, \frac{19^4\hat{\delta}_y^4}{10^4\xi^4\varepsilon^4}\right),
\end{aligned}$$

such that $\|\nabla\mathcal{G}_k\| \le \|\nabla\bar{\mathcal{G}}_k\| + \hat{c}_k\|y_k\| \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$. ∎

C Proof of Theorem 5.3

We prove the following two lemmas before giving the proof of Theorem 5.3.

Lemma C.1 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.13). If $\bar{\xi} > L_{22}$, then we have

$$\begin{aligned} & l(x_{k+1}, y_{k+1}) - l(x_k, y_k) \\ \geq & \frac{\bar{\xi}}{2}\|y_{k+1} - y_k\|^2 - \frac{L_{21}^2\varrho}{2}\|y_k - y_{k-1}\|^2 + (\hat{\theta} - \frac{1}{2\varrho} - \frac{\varrho L_{11}^2}{2})\|x_k - x_{k-1}\|^2 \\ & + (\frac{\hat{\theta}}{2} - \frac{1}{\varrho})\|x_{k+1} - x_k\|^2. \end{aligned} \tag{C.1}$$

Proof Similar to (A.2)-(A.4) in the proof of Lemma A.1, by the optimality condition for x_k in (5.4) implies that $\forall x \in \mathcal{X}$ and $\forall k \geq 1$,

$$\langle \nabla_x f(x_k, y_k) + \partial h_1(x_{k+1}) + \frac{1}{\varrho}(x_{k+1} - x_k), x - x_{k+1}\rangle \geq 0, \tag{C.2}$$

$$\langle \nabla_x f(x_k, y_k) + \partial h_1(x_{k+1}) + \frac{1}{\varrho}(x_{k+1} - x_k), x_k - x_{k+1}\rangle \geq 0, \tag{C.3}$$

$$\langle \nabla_x f(x_{k-1}, y_{k-1}) + \partial h_1(x_k) + \frac{1}{\varrho}(x_k - x_{k-1}), x_{k+1} - x_k\rangle \geq 0, \tag{C.4}$$

which, in view of the fact that $f(x, y)$ is $\hat{\theta}$-strongly convex w.r.t. x for any given $y \in \mathcal{Y}$ and $h_1(x)$ is convex, then implies that

$$\begin{aligned} l(x_{k+1}, y_k) - l(x_k, y_k) \geq & \langle \nabla_x f(x_k, y_k) + \partial h_1(x_k), x_{k+1} - x_k\rangle + \frac{\hat{\theta}}{2}\|x_{k+1} - x_k\|^2 \\ \geq & \langle \nabla_x f(x_k, y_k) - \nabla_x f(x_{k-1}, y_{k-1}), x_{k+1} - x_k\rangle \\ & - \frac{1}{\varrho}\langle x_k - x_{k-1}, x_{k+1} - x_k\rangle + \frac{\hat{\theta}}{2}\|x_{k+1} - x_k\|^2. \end{aligned} \tag{C.5}$$

The rest proof is the same with that of Lemma 4.1 except replacing ζ by ϱ and θ by $\hat{\theta}$. We omit the details here. ∎

Lemma C.2 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.13). Denote

$$l_{k+1} := l(x_{k+1}, y_{k+1}), \ \hat{S}_{k+1} := -\frac{2}{\varrho^2\hat{\theta}}\|x_{k+1} - x_k\|^2,$$

$$\hat{F}_{k+1} := l_{k+1} + \hat{S}_{k+1} + (\hat{\theta} + \frac{7}{2\varrho} - \frac{\varrho L_{11}^2}{2} - \frac{2L_{11}^2}{\hat{\theta}})\|x_{k+1} - x_k\|^2 - (\frac{L_{21}^2\varrho}{2} + \frac{2L_{21}^2}{\hat{\theta}^2\varrho})\|y_{k+1} - y_k\|^2.$$

If $\bar{\xi} > L_{22}$, then $\forall k \geq 1$,

$$\hat{F}_{k+1} - \hat{F}_k \geq \left(\frac{\bar{\xi}}{2} - \frac{\varrho L_{21}^2}{2} - \frac{2L_{21}^2}{\varrho\hat{\theta}^2}\right)\|y_{k+1} - y_k\|^2 + \left(\frac{3\hat{\theta} - \varrho L_{11}^2}{2} + \frac{\hat{\theta} - 4\varrho L_{11}^2}{2\varrho\hat{\theta}}\right)\|x_{k+1} - x_k\|^2. \tag{C.6}$$

Proof First by (C.3) and (C.4), we have

$$\frac{1}{\varrho}\langle m_{k+1}, x_k - x_{k+1}\rangle \geq \langle \nabla_x f(x_k, y_k) - \nabla_x f(x_{k-1}, y_{k-1}) + \partial h_1(x_{k+1}) - \partial h_1(x_k), x_{k+1} - x_k\rangle, \tag{C.7}$$

where $m_{k+1} := (x_{k+1} - x_k) - (x_k - x_{k-1})$, together with $\langle \partial h_1(x_{k+1}) - \partial h_1(x_k), x_{k+1} - x_k\rangle \geq 0$ then imply that

$$\frac{1}{\varrho}\langle m_{k+1}, x_k - x_{k+1}\rangle \geq \langle \nabla_x f(x_k, y_k) - \nabla_x f(x_{k-1}, y_{k-1}), x_{k+1} - x_k\rangle. \tag{C.8}$$

The rest proof is the same with that of Lemma 4.2 except replacing ζ by ϱ and θ by $\hat{\theta}$. We omit the details here. ∎

The proof of Theorem 5.3

Proof Noting that (5.5) is equivalent to $y_{k+1}^{(j)} = \mathrm{Prox}_{g_j, \mathcal{Y}_j}^{\bar{\xi}}\left(y_k^{(j)} + \frac{1}{\bar{\xi}}\nabla_{y^{(j)}} f\left(x_{k+1}, w_{k+1}^{(j)}\right)\right)$, we immediately obtain

$$\begin{aligned} & \|(\nabla \mathcal{G}_k)_{x^{(j)}}\| \\ \leq & \bar{\xi}\|\mathrm{Prox}_{g_j, \mathcal{Y}_j}^{\bar{\xi}}\left(y_k^{(j)} + \frac{1}{\bar{\xi}}\nabla_{y^{(j)}} f\left(x_{k+1}, w_{k+1}^{(j)}\right)\right) - \mathrm{Prox}_{g_j, \mathcal{Y}_j}^{\bar{\xi}}\left(y_k^{(j)} + \frac{1}{\bar{\xi}}\nabla_{y^{(j)}} f(x_k, y_k)\right)\| \\ & + \bar{\xi}\|y_{k+1}^{(j)} - y_k^{(j)}\| \\ \leq & \bar{\xi}\|y_{k+1}^{(j)} - y_k^{(j)}\| + \|\nabla_{y^{(j)}} f\left(x_{k+1}, w_{k+1}^{(j)}\right) - \nabla_{y^{(j)}} f(x_k, y_k)\| \\ \leq & \bar{\xi}\|y_{k+1}^{(j)} - y_k^{(j)}\| + L_{12}\|x_{k+1} - x_k\| + L_{22}\|w_{k+1}^{(j)} - y_k\| \\ \leq & \left(\bar{\xi} + L_{22}\right)\|y_{k+1} - y_k\| + L_{12}\|x_{k+1} - x_k\|, \end{aligned} \tag{C.9}$$

where the second inequality holds by the nonexpansiveness of the proximal operator $\mathrm{Prox}_{g_j,\mathcal{Y}_j}^{\bar{\xi}}$, the last inequality hold since $\|y_{k+1}^{(j)} - y_k^{(j)}\| \le \|y_{k+1} - y_k\|$, $\|w_{k+1}^{(j)} - y_k\| \le \|y_{k+1} - y_k\|$ by the definition of $y^{(j)}$ and $w_{k+1}^{(j)}$. On the other hand, since (5.4) is equivalent to $x_{k+1} = \mathrm{Prox}_{\mathcal{X},h_1}^{1/\varrho}\left(x_k - \varrho \nabla_x f\left(x_k, y_k\right)\right)$, we conclude from the triangle inequality that

$$\|(\nabla \mathcal{G}_k)_x\| = \frac{1}{\varrho}\|x_{k+1} - x_k\|. \tag{C.10}$$

By combining (C.9)-(C.10), and using Cauchy-Schwarz inequality, we obtain

$$\|\nabla \mathcal{G}_k\|^2 \le \left(2K_2\left(\bar{\xi} + L_{22}\right)^2\right)\|y_{k+1} - y_k\|^2 + \left(\frac{1}{\varrho^2} + 2K_2 L_{12}^2\right)\|x_{k+1} - x_k\|^2, \tag{C.11}$$

Observing that $\hat{r}_1 > 0$. Multiplying both sides of (C.11) by $\hat{r}_1$, and using (A.12) in Lemma C.2, we have

$$\hat{r}_1\|\nabla \mathcal{G}_k\|^2 \le \hat{F}_{k+1} - \hat{F}_k. \tag{C.12}$$

Summing up the above inequalities from $k = 1$ to $k = \mathcal{T}(\varepsilon)$ and using the definition of $\hat{L}_1$, we obtain

$$\sum_{k=1}^{\mathcal{T}(\varepsilon)} \hat{r}_1\|\nabla \mathcal{G}_k\|^2 \le \hat{F}_{\mathcal{T}(\varepsilon)+1} - \hat{F}_1 \le \hat{F}_{\mathcal{T}(\varepsilon)+1} - \hat{L}_1. \tag{C.13}$$

Note that by the definition of $\hat{F}_{k+1}$ in Lemma C.2, we have

$$\hat{F}_{\mathcal{T}(\varepsilon)+1} \le \hat{l} + (\hat{\theta} + \frac{7}{2\varrho} - \frac{\varrho L_{11}^2}{2} - \frac{2L_{11}^2}{\hat{\theta}})\delta_x^2 = \bar{L}.$$

We then conclude from (C.13) that

$$\sum_{k=1}^{\mathcal{T}(\varepsilon)} \hat{r}_1\|\nabla \mathcal{G}_k\|^2 \le \bar{L} - \hat{L}_1,$$

which, in view of the definition of $\mathcal{T}(\varepsilon)$, implies that $\varepsilon^2 \le (\bar{L} - \hat{L}_1)/(\mathcal{T}(\varepsilon) \cdot \hat{r}_1)$ or equivalently, $\mathcal{T}(\varepsilon) \le (\bar{L} - \hat{L}_1)/(\hat{r}_1\varepsilon^2)$. ■

D Proof of Theorem 5.4

We prove the following three lemmas before giving the proof of Theorem 5.4.

Lemma D.1 Suppose that Assumption 2.1 holds. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.17), If $\forall k, \bar{\vartheta}_k > L_{22}$, we have

$$l(x_{k+1}, y_{k+1}) - l(x_{k+1}, y_k) \ge \left(\bar{\iota} + \frac{\bar{\vartheta}_k}{2}\right)\|y_{k+1} - y_k\|^2. \tag{D.1}$$

Proof The proof is the same with that of Lemma 5.3 except replacing $\xi_k^{(j)}$ by $\bar{\iota} + \bar{\vartheta}_k$. We omit the details here. ■

Lemma D.2 Suppose that Assumption 2.1 and 5.2 hold. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.17). If $\forall k, \bar{\vartheta}_k > L_{22}$ and $\bar{\varrho} \le \frac{2}{L_{11}' + \hat{b}_1}$, then

$$\begin{aligned}
& l(x_{k+1}, y_{k+1}) - l(x_k, y_k) \\
\ge & \left(\bar{\iota} + \frac{\bar{\vartheta}_k}{2}\right)\|y_{k+1} - y_k\|^2 - \frac{L_{21}^2\bar{\varrho}}{2}\|y_k - y_{k-1}\|^2 - \frac{1}{\bar{\varrho}}\|x_{k+1} - x_k\|^2 - \frac{1}{2\bar{\varrho}}\|x_k - x_{k-1}\|^2 \\
& - \frac{\hat{b}_{k-1}}{2}(\|x_{k+1}\|^2 - \|x_k\|^2).
\end{aligned} \tag{D.2}$$

Proof Similar to (C.2)-(C.4) in the proof of Lemma C.1, by replacing l with $\bar{l}_k$, $f(x_k, y_k)$ with $\bar{f}_k$, ϱ with $\bar{\varrho}$ respectively, and setting $\hat{\theta} = 0$, we obtain that

$$\begin{aligned}
\bar{l}_k(x_{k+1}, y_k) - \bar{l}_k(x_k, y_k) \ge & \langle \nabla_x \bar{f}_k(x_k, y_k) + \partial h_1(x_k), x_{k+1} - x_k\rangle \\
\ge & \langle \nabla_x \bar{f}_k(x_k, y_k) - \nabla_x \bar{f}_{k-1}(x_{k-1}, y_{k-1}), x_{k+1} - x_k\rangle \\
& - \frac{1}{\bar{\varrho}}\langle x_k - x_{k-1}, x_{k+1} - x_k\rangle.
\end{aligned} \tag{D.3}$$

The rest proof is the same with Lemma 4.3 except replacing $\bar{\zeta}$ by $\bar{\varrho}$ and q_k by $\hat{b}_k$. We omit the details here.

■

Lemma D.3 Suppose that Assumptions 2.1 and 5.2 hold. Let $\{(x_k, y_k)\}$ be a sequence generated by Algorithm 2 with parameter settings in (5.17). Denote

$$\hat{\mathcal{S}}_{k+1} := -\frac{8}{\bar{\varrho}^2 \hat{b}_{k+1}}\|x_{k+1} - x_k\|^2 - \frac{8}{\bar{\varrho}}\left(1 - \frac{\hat{b}_k}{\hat{b}_{k+1}}\right)\|x_{k+1}\|^2,$$

$$\begin{aligned}\hat{\mathcal{F}}_{k+1} := {} & l(x_{k+1}, y_{k+1}) + \hat{\mathcal{S}}_{k+1} + \frac{15}{2\bar{\varrho}}\|x_{k+1} - x_k\|^2 + \frac{\hat{b}_k}{2}\|x_{k+1}\|^2 \\ & - \left(\frac{\bar{\varrho}L_{21}^2}{2} + \frac{16L_{21}^2}{\bar{\varrho}(\hat{b}_{k+1})^2}\right)\|y_{k+1} - y_k\|^2.\end{aligned}$$

If

$$\bar{\vartheta}_k > L_{22},\ \frac{1}{\hat{b}_{k+1}} - \frac{1}{\hat{b}_k} \le \frac{\bar{\varrho}}{5},\ \bar{\varrho} \le \frac{2}{L_{11}' + \hat{b}_1}, \tag{D.4}$$

then $\forall k \ge 1$,

$$\begin{aligned}& \hat{\mathcal{F}}_{k+1} - \hat{\mathcal{F}}_k \\ \ge {} & \left(\bar{\iota} + \frac{\bar{\vartheta}_k}{2} - \frac{\bar{\varrho}L_{21}^2}{2} - \frac{16L_{21}^2}{\bar{\varrho}(\hat{b}_{k+1})^2}\right)\|y_{k+1} - y_k\|^2 + \frac{\hat{b}_k - \hat{b}_{k-1}}{2}\|x_{k+1}\|^2 \\ & + \frac{9}{10\bar{\varrho}}\|x_{k+1} - x_k\|^2 + \frac{8}{\bar{\varrho}}\left(\frac{\hat{b}_k}{\hat{b}_{k+1}} - \frac{\hat{b}_{k-1}}{\hat{b}_k}\right)\|x_{k+1}\|^2.\end{aligned}$$

Proof Similar to (C.7) in the proof of Lemma C.2, by replacing $f(x_k, y_k)$ with $\bar{f}_k$, ϱ with $\bar{\varrho}$ respectively, we have

$$\frac{1}{\bar{\varrho}}\langle m_{k+1}, x_k - x_{k+1}\rangle \ge \langle \nabla_x \bar{f}_k(x_k, y_k) - \nabla_x \bar{f}_{k-1}(x_{k-1}, y_{k-1}), x_{k+1} - x_k\rangle,$$

The rest proof is the same with Lemma 4.4 except replacing $\bar{\zeta}$ by $\bar{\varrho}$ and q_k by $\hat{b}_k$. We omit the details here. ∎

The proof of Theorem 5.4

Proof By $\bar{\varrho} \le \frac{1}{10L_{11}}$,$\hat{b}_k = \frac{19}{20\bar{\varrho}k^{1/4}}$, $\forall k \ge 1$, let us denote $\bar{\vartheta}_k = \bar{\varrho}L_{21}^2 + \frac{16\tau L_{21}^2}{\bar{\varrho}(\hat{b}_{k+1})^2} - 2\bar{\iota}$, $p_k = \frac{8(\tau-2)L_{21}^2}{\bar{\varrho}(\hat{b}_{k+1})^2}$, it can be easily checked that the relations in (D.4) are satisfied. It follows from the selection of $\bar{\vartheta}_k$ and p_k that

$$\bar{\iota} + \frac{\bar{\vartheta}_k}{2} - \frac{\bar{\varrho}L_{21}^2}{2} - \frac{16L_{21}^2}{\bar{\varrho}(\hat{b}_{k+1})^2} = p_k.$$

This observation, in view of Lemma D.3, then immediately implies that

$$\begin{aligned}p_k\|y_{k+1} - y_k\|^2 + \frac{9}{10\bar{\varrho}}\|x_{k+1} - x_k\|^2 \le {} & \hat{\mathcal{F}}_{k+1} - \hat{\mathcal{F}}_k + \frac{8}{\bar{\varrho}}\left(\frac{\hat{b}_{k-1}}{\hat{b}_k} - \frac{\hat{b}_k}{\hat{b}_{k+1}}\right)\|x_{k+1}\|^2 \\ & + \frac{\hat{b}_{k-1} - \hat{b}_k}{2}\|x_{k+1}\|^2.\end{aligned} \tag{D.5}$$

We can easily check from the definition of $\bar{f}_k(x_k, y_k)$ that

$$\|\nabla \mathcal{G}_k\| - \|\nabla \bar{\mathcal{G}}_k\| \le \hat{b}_k\|x_k\|.$$

Similar to (C.9) and (C.10) in the proof of Theorem 5.3, by replacing f with $\bar{f}_k$, $\nabla\mathcal{G}_k$ with $\nabla\bar{\mathcal{G}}_k$, ϱ with $\bar{\varrho}$, $\bar{\xi}$ with $\bar{\vartheta}_k + \bar{\iota}$ respectively, we conclude that

$$\|\nabla\bar{\mathcal{G}}_k\|^2 \le \left(2K_2\left(\bar{\vartheta}_k + \bar{\iota} + L_{22}\right)^2\right)\|y_{k+1} - y_k\|^2 + \left(\frac{1}{\bar{\varrho}^2} + 2K_2L_{12}^2\right)\|x_{k+1} - x_k\|^2. \tag{D.6}$$

Since both p_k and $\bar{\vartheta}_k$ are in the same order when k becomes large enough, it then follows from the definition of $\hat{d}_1$ that $\forall k \ge 1$,

$$\begin{aligned}\frac{2K_2\left(\bar{\vartheta}_k + \bar{\iota} + L_{22}\right)^2}{p_k^2} &= \frac{2K_1\left(\bar{\varrho}L_{21}^2 + \frac{16\tau L_{21}^2}{\bar{\varrho}(\hat{b}_{k+1})^2} - \bar{\iota} + L_{22}\right)^2}{p_k^2} \\ &\le \frac{4K_2\left(\frac{16\tau L_{21}^2}{\bar{\varrho}(\hat{b}_{k+1})^2}\right)^2 + 4K_2\left(\bar{\varrho}L_{21}^2 - \bar{\iota} + L_{22}\right)^2}{p_k^2} \\ &= \frac{16K_2\tau^2}{(\tau-2)^2} + \frac{19^4\left(4K_2\left(\bar{\varrho}L_{21}^2 - \bar{\iota} + L_{22}\right)^2\right)}{64\cdot 20^4\bar{\varrho}^2(\tau-2)^2L_{21}^4k} \\ &\le \hat{d}_1.\end{aligned} \tag{D.7}$$

Combining the previous two inequalities in (D.6) and (D.7), we obtain

$$\|\nabla\bar{\mathcal{G}}_k\|^2 \le \hat{d}_1(p_k)^2\|y_{k+1}-y_k\|^2 + \left(\frac{1}{\bar{\varrho}^2} + 2K_2L_{12}^2\right)\|x_{k+1}-x_k\|^2. \tag{D.8}$$

Denote $\hat{d}_k^{(2)} = \frac{1}{\max\left\{\hat{d}_1 p_k, \frac{10+20K_2\bar{\varrho}^2L_{12}^2}{9\bar{\varrho}}\right\}}$. By multiplying $\hat{d}_k^{(2)}$ on the both sides of (D.8), and using (D.5), we get

$$\hat{d}_k^{(2)}\|\nabla\bar{\mathcal{G}}_k\|^2 \le \hat{\mathcal{F}}_{k+1} - \hat{\mathcal{F}}_k + \frac{8}{\bar{\varrho}}\left(\frac{\hat{b}_{k-1}}{\hat{b}_k} - \frac{\hat{b}_k}{\hat{b}_{k+1}}\right)\|x_{k+1}\|^2 + \frac{\hat{b}_{k-1}-\hat{b}_k}{2}\|x_{k+1}\|^2. \tag{D.9}$$

Denoting

$$\bar{\mathcal{T}}(\varepsilon) := \min\{k \mid \|\nabla\bar{\mathcal{G}}(x_k,y_k)\| \le \frac{\varepsilon}{2}, k\ge 2\},$$

Summing both sides of (D.9) from $k=2$ to $k=\bar{\mathcal{T}}(\varepsilon)$, we then obtain

$$\begin{aligned}
&\sum_{k=2}^{\bar{\mathcal{T}}(\varepsilon)} \hat{d}_k^{(2)}\|\nabla\bar{\mathcal{G}}_k\|^2 \\
\le& \hat{\mathcal{F}}_{\bar{\mathcal{T}}(\varepsilon)+1} - \hat{\mathcal{F}}_2 + \frac{8}{\bar{\varrho}}\left(\frac{\hat{b}_1}{\hat{b}_2} - \frac{\hat{b}_{\bar{\mathcal{T}}(\varepsilon)}}{\hat{b}_{\bar{\mathcal{T}}(\varepsilon)+1}}\right)\hat{\delta}_x^2 + \frac{\hat{b}_1-\hat{b}_{\bar{\mathcal{T}}(\varepsilon)}}{2}\hat{\delta}_x^2 \\
\le& \hat{\mathcal{F}}_{\bar{\mathcal{T}}(\varepsilon)+1} - \hat{\mathcal{F}}_2 + \frac{8\hat{b}_1}{\bar{\varrho}\hat{b}_2}\hat{\delta}_x^2 + \frac{\hat{b}_1}{2}\hat{\delta}_x^2 \\
=& \hat{\mathcal{F}}_{\bar{\mathcal{T}}(\varepsilon)+1} - \hat{\mathcal{F}}_2 + \frac{8\cdot 2^{1/4}}{\bar{\varrho}}\hat{\delta}_x^2 + \frac{19}{40\bar{\varrho}}\hat{\delta}_x^2.
\end{aligned} \tag{D.10}$$

Note that by the definition of $\hat{\mathcal{F}}_{k+1}$ in Lemma D.3, we have

$$\begin{aligned}
\hat{\mathcal{F}}_{\bar{\mathcal{T}}(\varepsilon)+1} &\le \hat{l} + \frac{8\hat{\delta}_x^2}{\bar{\varrho}}\left(1+\frac{1}{\bar{\mathcal{T}}(\varepsilon)}\right)^{1/4} + \frac{15\hat{\delta}_x^2}{\bar{\varrho}} + \frac{q_1\hat{\delta}_x^2}{2} \\
&= \hat{l} + \left(2^{13/4} + 15 + \frac{19}{40}\right)\frac{\hat{\delta}_x^2}{\bar{\varrho}} = \hat{\mathcal{L}}_0,
\end{aligned}$$

where $\hat{l} := \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} l(x,y)$. By the definition of $\hat{\mathcal{L}}_2$, we then conclude from (D.10) that

$$\begin{aligned}
\sum_{k=2}^{\bar{\mathcal{T}}(\varepsilon)} \hat{d}_k^{(2)}\|\nabla\bar{\mathcal{G}}_k\|^2 &\le \hat{\mathcal{L}}_0 - \hat{\mathcal{F}}_2 + \frac{8\cdot 2^{1/4}}{\bar{\varrho}}\hat{\delta}_x^2 + \frac{19}{40\bar{\varrho}}\hat{\delta}_x^2 \\
&\le \hat{\mathcal{L}}_0 - \hat{\mathcal{L}}_2 + \frac{8\cdot 2^{1/4}}{\bar{\varrho}}\hat{\delta}_x^2 + \frac{19}{40\bar{\varrho}}\hat{\delta}_x^2 = \hat{r}_2.
\end{aligned} \tag{D.11}$$

Note that $\hat{r}_3 = \max\{\hat{d}_1, \frac{10+20K_2\bar{\varrho}^2L_{12}^2}{9\bar{\varrho}p_2}\}$. Since p_k is an increasing sequence when $k\ge 2$, we have that $\hat{r}_3 \ge \max\{\hat{d}_1, \frac{10+20\bar{\varrho}^2L_{12}^2}{9\bar{\varrho}p_k}\}$, which implies that $\hat{d}_k^{(2)} \ge \frac{1}{\hat{r}_3 p_k}$. By multiplying $\hat{r}_3$ on the both sides of (D.11), and using the definition of $\hat{r}_2$, we have $\sum_{k=2}^{\bar{\mathcal{T}}(\varepsilon)}\frac{1}{p_k}\|\nabla\bar{\mathcal{G}}_k\|^2 \le \hat{r}_2\hat{r}_3$, which, by the definition of $\bar{\mathcal{T}}(\varepsilon)$, implies that

$$\frac{\varepsilon^2}{4} \le \frac{\hat{r}_2\hat{r}_3}{\sum_{k=2}^{\bar{\mathcal{T}}(\varepsilon)}\frac{1}{p_k}}. \tag{D.12}$$

Using the assumptions $\hat{b}_k = \frac{19}{20\bar{\varrho}k^{1/4}}$ and $p_k = \frac{2\cdot 40^2\bar{\varrho}(\tau-2)L_{21}^2\sqrt{k+1}}{19^2}$, (D.12) and the fact $\sum_{k=2}^{\bar{\mathcal{T}}(\varepsilon)} 1/\sqrt{k+1} \ge \sqrt{\bar{\mathcal{T}}(\varepsilon)} - 2$, we conclude that $\frac{\varepsilon^2}{4} \le \frac{2\cdot 40^2\bar{\varrho}(\tau-2)L_{21}^2\hat{r}_2\hat{r}_3}{19^2\left(\sqrt{\bar{\mathcal{T}}(\varepsilon)}-2\right)}$ or equivalently,

$$\bar{\mathcal{T}}(\varepsilon) \le \left(\frac{2\cdot 80^2\bar{\varrho}(\tau-2)L_{21}^2\hat{r}_2\hat{r}_3}{19^2\varepsilon^2} + 2\right)^2.$$

On the other hand, if $k > \frac{19^4\hat{\delta}_x^4}{10^4\bar{\varrho}^4\varepsilon^4}$, then $q_k = \frac{19}{20\bar{\varrho}k^{1/4}} \le \frac{\varepsilon}{2\hat{\delta}_x}$, this inequality together with the definition of $\hat{\delta}_x$ then imply that $\hat{b}_k\|x_k\| \le \frac{\varepsilon}{2}$. Therefore, there exists a

$$\mathcal{T}(\varepsilon) \le \max\left(\bar{\mathcal{T}}(\varepsilon), \frac{19^4\hat{\delta}_x^4}{10^4\bar{\varrho}^4\varepsilon^4}\right) \le \max\left(\left(\frac{2\cdot 80^2\bar{\varrho}(\tau-2)L_{21}^2\hat{r}_2\hat{r}_3}{19^2\varepsilon^2} + 2\right)^2, \frac{19^4\hat{\delta}_x^4}{10^4\bar{\varrho}^4\varepsilon^4}\right),$$

such that $\|\nabla\mathcal{G}_k\| \le \|\nabla\bar{\mathcal{G}}_k\| + \hat{b}_k\|x_k\| \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$. ∎