A Single-Loop Gradient-Decent-Ascent Algorithm with Momentum for Nonconvex Minimax Optimization

Anonymous Authors[1]

Abstract

Gradient-descent-ascent (GDA) has been widely applied in minimax optimization. As the updates of GDA are gradient-based, it is natural to apply momentum to accelerate its practical convergence. However, most of the existing designs of GDA with momentum adopt a nested-loop structure, which introduces extra hyper-parameters and cannot be implemented efficiently in practice. In this paper, we develop a simple and effective momentum scheme for accelerating GDA in nonconvex-strongly-concave optimization. Our algorithm design has the following amenable features: (i) it has a single loop and involves only learning rate and momentum as hyperparameters; (ii) it applies the heavy-ball momentum and Nesterov's momentum to the descent update and ascent update, respectively, to simultaneously boost the convergence of nonconvex optimization and strongly-concave optimization. By leveraging a special Lyapunov function, we prove the convergence of the algorithm to a critical point. Moreover, under a Łojasiewicz gradient-type geometry, we established the convergence rates of the algorithm for different parameterizations of the geometry. Experiments validate the effectiveness of our algorithm.

1. Introduction

Minimax optimization is an emerging and important optimization framework that covers a variety of modern machine learning applications. Some popular application examples include generative adversarial networks (GANs) (Goodfellow et al., 2014), adversarial training (Sinha et al., 2017), game theory (Ferreira et al., 2012), reinforcement learning (Qiu et al., 2020), etc. A standard minimax optimization problem is written as follows, where f is a differentiable bivariate function.

$$\min_{x\in\mathbb{R}^m} \max_{y\in\mathcal{Y}} f(x,y).$$

A basic and popular algorithm for solving the above minimax problem is gradient-descent-ascent (GDA), which alternatively performs a gradient descent update on the variable x and a gradient ascent update on the variable y. Despite the simplicity of GDA, its convergence properties are not well understood until recently. Specifically, many studies have established the convergence of GDA under different global geometries of the objective function, e.g., convex-concave geometry (Nedić & Ozdaglar, 2009) and Polyak-Łojasiewicz (PŁ) geometry (Nouiehed et al., 2019; Yang et al., 2020). Some other works considered stronger global geometries such as (strongly)-convex-strongly-concave geometry (Du & Hu, 2019; Mokhtari et al., 2020; Zhang & Wang, 2020) and bi-linear geometry (Neumann, 1928; Robinson, 1951). However, these global function geometries do not cover modern machine learning problems that are usually nonconvex. Recently, (Lin et al., 2020b; Nouiehed et al., 2019; Xu et al., 2020b) studied the convergence of GDA in the nonconvex-concave setting, whereas (Lin et al., 2020b; Xu et al., 2020b) focused on the nonconvex-strongly-concave setting. In these nonconvex settings, it is shown that GDA converges to a certain stationary point at a sublinear rate.

As GDA can be viewed as a generalization of the gradient descent algorithm to the minimax setting, it is natural to consider applying acceleration techniques that are originally developed for gradient-based algorithms to boost the convergence of GDA. In particular, **momentum** is a popular and widely used technique. In conventional gradient-based optimization, momentum is well known to accelerate the convergence rate of gradient descent in convex optimization (Nesterov, 2014; Tseng, 2010; Beck & Teboulle, 2009), and is also widely applied to boost the practical convergence of gradient descent in nonconvex optimization (Ghadimi & Lan, 2016; Li & Lin, 2015; Li et al., 2017). Recently, several works applied momentum to boost the convergence of GDA. Specifically, for strongly-convex-strongly-concave problems, (Thekumparampil et al., 2019; Wang & Li, 2020; Lin et al., 2020a) developed nested-loop accelerated gradient descent (AGD)-based GDA algorithms with improved

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

complexities, and (Zhang & Wang, 2020) developed a negative momentum-based GDA algorithm and characterized its convergence rate. For convex-concave problems, (Zhu et al., 2020) developed an AGD-based primal-dual algorithm. Moreover, for nonconvex-(strongly)-concave problems, (Lin et al., 2020a) developed a nested-loop AGD-based GDA algorithm that achieves (near)-optimal complexities. Despite that these momentum-accelerated GDA algorithms have convergence guarantees, they are lack of practicability as we elaborate below.

- Many of the existing GDA with momentum algorithms adopt complex algorithm designs that involve a **nested-loop structure**, in which the algorithms take (multiple) inner-loop updates to solve certain subproblems up to a predefined accuracy. Such a nested-loop structure often slows down the practical convergence and introduces many extra hyperparameters that are hard to fine-tune in practice. As a comparison, single-loop algorithms (i.e., without nested-loop) are much easier to implement and usually converge faster than nested-loop algorithms in practice.
- Many of these GDA with momentum algorithms have convergence guarantee only for convex-concave-type problems, which do not cover modern machine learning problems that have nonconvex geometry. On the other hand, although the accelerated GDA developed in (Lin et al., 2020a) has a convergence guarantee in nonconvex scenarios, it only guarantees a certain type of gradient norm convergence. Such a convergence result does not guarantee the desired **variable convergence**, i.e., minimax optimization obtains convergent optimization variables $x_t \to x^*, y_t \to y^*(x^*)$ (see the analysis section).

Hence, as the existing designs of GDA with momentum algorithms are lack of practicability, it is of vital importance to develop a simple, practical and effective momentum scheme that can accelerate GDA and has strong convergence guarantee. We therefore want to ask the following fundamental questions.

- *Q1: Can we develop a single-loop momentum scheme for boosting the convergence of GDA in nonconvex minimax optimization?*
- *Q2: With momentum acceleration, can GDA generate variable sequences that converge to a critical point in nonconvex minimax optimization?*

In this paper, we provide comprehensive answers to these questions. We develop a simple and effective momentum scheme for the proximal-GDA algorithm for solving nonsmooth and nonconvex minimax problems. This algorithm has a single-loop structure and adopts momentum to boost the practical convergence. We summarize our contributions as follows.

1.1. Our Contributions

We consider the following regularized nonconvex-strongly-concave minimax optimization problem.

$$\min_{x\in\mathbb{R}^m} \max_{y\in\mathcal{Y}} f(x,y) + g(x) - h(y), \tag{P}$$

where f is differentiable and nonconvex-strongly-concave, g is a general nonconvex regularizer and h is a convex regularizer. Both g and h can be nonsmooth. To solve the above minimax problem, we propose a proximal-GDA with momentum (referred to as PGDAm) algorithm. Specifically, to minimize the nonconvex part of the objective function, PGDAm applies a proximal gradient descent step with the heavy-ball momentum. On the other hand, to maximize the strongly-concave part, PGDAm applies a proximal gradient ascent step with the Nesterov's momentum. Our study extends the applicability of the conventional momentum schemes (heavy-ball and Nesterov's momentum) for nonconvex minimization to nonconvex minimax optimization. In particular, our PGDAm has a single loop and involves only learning rate and momentum as hyperparameters.

We study the convergence property of PGDAm. Specifically, under standard smoothness assumptions on the objective function f, we show that PGDAm admits a monotonically decreasing Lyapunov function (see Proposition 2). By leveraging this special Lyapunov function, we show that every limit point of the variable sequences generated by PGDAm is a critical point of the minimax problem (P).

Then, we further investigate the convergence of PGDAm under the general nonconvex Kurdyka-Lojasiewicz (KŁ) geometry of the objective function. We prove that the entire variable sequences generated by PGDAm admit a unique limit point, i.e., they converge to a certain critical point $x_t \to x^*, y_t \to y^*(x^*)$ (see the definition of y^* in Section 2). Moreover, we characterize the asymptotic convergence rates of both the variable sequences and function value of PGDAm in different parameterization regimes of the KŁ geometry. Depending on the value of the KŁ parameter θ, we show that PGDAm achieves different types of convergence rates that range from sublinear convergence to finite-step convergence.

1.2. Related Work

GDA algorithms: (Yang et al., 2020) studied an alternating gradient descent-ascent algorithm in which the gradient ascent step uses the current variable x_{t+1} instead of x_t. (Xu et al., 2020b) studied an alternating gradient projection algorithm which applies ℓ_2 regularizer to the local objective function of GDA followed by projection onto the constraint sets. (Mokhtari et al., 2020) also studied an extra-gradient algorithm which applies two-step GDA in each iteration. (Nouiehed et al., 2019) studied multi-step GDA where mul-

tiple gradient ascent steps are performed, and they also studied the momentum-accelerated version. (Cherukuri et al., 2017; Daskalakis & Panageas, 2018; Jin et al., 2020) studied GDA in continuous time dynamics using differential equations. (Adolphs et al., 2019) analyzed a second-order variant of the GDA algorithm.

Many other studies have developed stochastic GDA algorithms. (Lin et al., 2020b; Yang et al., 2020) analyzed stochastic GDA and stochastic AGDA, which are direct extension of GDA and AGDA to the stochastic setting. Variance reduction techniques have been applied to stochastic minimax optimization, including SVRG-based (Du & Hu, 2019; Yang et al., 2020), SPIDER-based (Xu et al., 2020a), STORM (Qiu et al., 2020) and its gradient free version (Huang et al., 2020). (Xie et al., 2020) studied the complexity lower bound of first-order stochastic algorithms for finite-sum minimax problem.

GDA with momentum: For strongly-convex-strongly-concave problems, (Thekumparampil et al., 2019; Wang & Li, 2020; Lin et al., 2020a) developed nested-loop AGD-based GDA algorithms with improved complexities. For convex-concave problems, (Zhu et al., 2020) developed an AGD-based primal-dual algorithm. (Daskalakis & Panageas, 2018; Mokhtari et al., 2020; Zhang & Wang, 2020) analyzed optimistic GDA that applies negative momentum to accelerate GDA. Moreover, for nonconvex-(strongly)-concave problems, (Lin et al., 2020a) developed a nested-loop AGD-based GDA algorithm that achieves (near)-optimal complexities.

KŁ geometry: The Kurdyka-Lojasiewicz (KŁ) geometry was defined in (Bolte et al., 2007). The KŁ geometry has been exploited to study the convergence of various first-order algorithms for solving minimization problems, including gradient descent (Attouch & Bolte, 2009), alternating gradient descent (Bolte et al., 2014), distributed gradient descent (Zhou et al., 2016a; 2018a), accelerated gradient descent (Li et al., 2017). It has also been exploited to study the convergence of second-order algorithms such as Newton's method (Noll & Rondepierre, 2013; Frankel et al., 2015) and cubic regularization method (Zhou et al., 2018b).

2. Problem Formulation and Preliminaries

In this section, we introduce the problem formulation and technical assumptions. Recall the following regularized minimax optimization problem.

$$\min_{x\in\mathbb{R}^m}\max_{y\in\mathcal{Y}} f(x,y)+g(x)-h(y), \tag{P}$$

where $f:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$ is differentiable and nonconvex-strongly-concave, $\mathcal{Y}\subset\mathbb{R}^n$ is a compact and convex set, and g,h are possibly non-smooth regularizers. In particular, define the envelope function $\Phi(x):=\max_{y\in\mathcal{Y}} f(x,y)-h(y)$, then the problem (P) can be rewritten as the minimization problem $\min_{x\in\mathbb{R}^m}\Phi(x)+g(x)$.

Throughout the paper, we adopt the following assumptions on the problem (P). These are standard assumptions that have been considered in the existing literature (Lin et al., 2020b; Chen et al., 2021).

Assumption 1. *The minimax problem* (P) *satisfies:*

1. *Function $f(\cdot,\cdot)$ is L-smooth and function $f(x,\cdot)$ is μ-strongly concave for all x;*
2. *Function $(\Phi+g)(x)$ is bounded below and has compact sub-level sets;*
3. *Function h is proper and convex, and function g is proper and lower semi-continuous.*

In particular, item 3 of the above assumption allows the regularizer h to be any convex function, which include the popular examples such as ℓ_1 norm, nuclear norm, etc. On the other hand, the regularizer g can be any lower semi-continuous nonconvex function, e.g., the ℓ_0 norm, ℓ_p norm with $0<p<1$, matrix rank.

Next, consider the mapping $y^*(x)=\arg\max_{y\in\mathcal{Y}} f(x,y)-h(y)$, which is uniquely defined by the strong concavity of $f(x,\cdot)$. The following proposition is proved in (Lin et al., 2020b; Chen et al., 2021) that characterizes the Lipschitz continuity of the mapping $y^*(x)$ and the smoothness of the function $\Phi(x)$. Throughout the paper, we denote $\kappa=L/\mu>1$ as the condition number, and denote $\nabla_1 f(x,y),\nabla_2 f(x,y)$ as the gradients with respect to the first and the second input argument, respectively.

Proposition 1 ((Lin et al., 2020b; Chen et al., 2021))**.** *Let Assumption 1 hold. Then, the mapping $y^*(x)$ is κ-Lipschitz continuous and the function $\Phi(x)$ is $L(1+\kappa)$-smooth with $\nabla\Phi(x)=\nabla_1 f(x,y^*(x))$.*

Lastly, recall that the minimax problem (P) is equivalent to the minimization problem $\min_{x\in\mathbb{R}^m}\Phi(x)+g(x)$. Therefore, the optimization goal of the minimax problem (P) is equivalent to finding a critical point x^* of the nonconvex function $\Phi(x)+g(x)$ that satisfies the optimality condition $\mathbf{0}\in\partial(\Phi+g)(x^*)$. Here, ∂ denotes the following generalized notion of subdifferential.

Definition 1. *(Subdifferential and critical point, (Rockafellar & Wets, 2009)) The Frechét subdifferential $\widehat{\partial}h$ of function h at $x\in\operatorname{dom}h$ is the set of $u\in\mathbb{R}^d$ defined as*

$$\widehat{\partial}h(x):=\Big\{u:\liminf_{z\neq x,z\to x}\frac{h(z)-h(x)-u^\intercal(z-x)}{\|z-x\|}\geq 0\Big\},$$

and the limiting subdifferential ∂h at $x\in\operatorname{dom}h$ is the graphical closure of $\widehat{\partial}h$ defined as:

$$\partial h(x):=\{u:\exists(x_k,h(x_k))\to(x,h(x)),\widehat{\partial}h(x_k)\ni u_k\xrightarrow{k}u\}.$$

The set of critical points of h is defined as $\{x:\mathbf{0}\in\partial h(x)\}$.

Throughout, the limiting subdifferential is referred to as subdifferential. Subdifferential is a generalization of gradient (when h is differentiable) and subgradient (when h is convex) to the nonconvex and lower semi-continuous setting. In particular, any local minimizer of the function h must be a critical point.

3. Proximal-GDA with Momentum

In this section, we propose a proximal-GDA with momentum (PGDAm) algorithm to solve the regularized minimax problem (P).

We first recall the update rules of the basic proximal-GDA algorithm for solving the problem (P). Specifically, proximal-GDA alternates between the following two proximal-gradient updates (a.k.a. forward-backward splitting updates (Lions & Mercier, 1979)).

(Proximal-GDA):

$$\begin{cases} x_{t+1} \in \mathrm{prox}_{\eta_x g}\big(x_t - \eta_x \nabla_1 f(x_t, y_t)\big), \\ y_{t+1} = \mathrm{prox}_{\eta_y h}\big(y_t + \eta_y \nabla_2 f(x_t, y_t)\big). \end{cases}$$

To elaborate, the first update is a proximal gradient descent update that aims to minimize the nonconvex function $f(x, y_t)+g(x)$ at the current point x_t, and the second update is a proximal gradient ascent update that aims to maximize the strongly-concave function $f(x_t, y) - h(y)$ at the current point y_t. In particular, as g is nonconvex and $\mathrm{prox}_{\eta_x g}$ can be a set-valued mapping, we set x_{t+1} to be any element of the nonconvex proximal mapping. More specifically, the two proximal gradient mappings are formally defined as

$$\begin{aligned} &\mathrm{prox}_{\eta_x g}\big(x_t - \eta_x \nabla_1 f(x_t, y_t)\big) \\ &:= \operatorname*{argmin}_{u \in \mathbb{R}^m} \Big\{ g(u) + \frac{1}{2\eta_x} \|u - x_t + \eta_x \nabla_1 f(x_t, y_t)\|^2 \Big\}, \\ &\mathrm{prox}_{\eta_y h}\big(y_t + \eta_y \nabla_2 f(x_t, y_t)\big) \\ &:= \operatorname*{argmin}_{v \in \mathcal{Y}} \Big\{ h(v) + \frac{1}{2\eta_y} \|v - y_t - \eta_y \nabla_2 f(x_t, y_t)\|^2 \Big\}. \end{aligned}$$

Next, we introduce our design of momentum for the proximal-GDA. As the two proximal gradient steps of proximal-GDA are used to solve two different types of optimization problems, namely, the nonconvex problem $f(x, y_t)+g(x)$ and the strongly-concave problem $f(x_t, y)-h(y)$, we consider applying different momentum schemes to accelerate these proximal gradient updates. Specifically, the proximal gradient descent step minimizes a composite nonconvex function, and we apply the heavy-ball momentum scheme (Polyak, 1964) that was originally designed for accelerating nonconvex optimization. Therefore, we propose the following proximal gradient descent with heavy-ball momentum update for minimizing the nonconvex part of the problem (P).

(Heavy-ball momentum):

$$\begin{cases} \widetilde{x}_t = x_t + \beta(x_t - x_{t-1}), \\ x_{t+1} \in \mathrm{prox}_{\eta_x g}\big(\widetilde{x}_t - \eta_x \nabla_1 f(x_t, y_t)\big). \end{cases}$$

To explain, the first step is an extrapolation step that applies the momentum term $\beta(x_t - x_{t-1})$ (with momentum coefficient $\beta > 0$), and the second proximal gradient step updates the extrapolation point $\widetilde{x}_t$ using the original gradient $\nabla_1 f(x_t, y_t)$. In conventional gradient-based optimization, gradient descent with such a heavy-ball momentum is guaranteed to find a critical point of smooth nonconvex functions (Ochs et al., 2014; Ochs, 2018).

On the other hand, as the proximal gradient ascent step of Proximal-GDA maximizes a composite strongly-concave function, we are motivated to apply the popular Nesterov's momentum scheme, which was originally designed for accelerating strongly-concave (convex) optimization. Specifically, we propose the following proximal gradient ascent with Nesterov's momentum update for maximizing the strongly-concave part of the problem (P).

(Nesterov's momentum):

$$\begin{cases} \widetilde{y}_t = y_t + \gamma(y_t - y_{t-1}), \\ y_{t+1} = \mathrm{prox}_{\eta_y h}\big(\widetilde{y}_t + \eta_y \nabla_2 f(x_t, \widetilde{y}_t)\big). \end{cases} \tag{1}$$

To elaborate, the first step is a regular extrapolation step that involves momentum. The second proximal gradient step is different from that in the heavy-ball scheme. Specifically, the proximal gradient ascent step is updated at the extrapolated point $\widetilde{y}_t$ using the gradient at the extrapolated point, i.e., $\nabla_2 f(x_t, \widetilde{y}_t)$. Hence, the Nesterov's momentum scheme is more greedy than the heavy-ball momentum. Intuitively, this is because the function $f(x_t, y) - h(y)$ has a more amenable strongly-concave optimization geometry, and therefore Nesterov's momentum should be applied to maximize the momentum acceleration effect.

We refer to the above algorithm design as **proximal-GDA with momentum (PGDAm)**, and the algorithm updates are formally presented in Algorithm 1. It can be seen that PGDAm is a simple GDA-type algorithm that has a single loop and adopts momentum acceleration. More importantly, it involves only standard hyper-parameters such as the learning rates and momentum parameters and therefore is easy to implement in practice. In fact, this is probably the most natural design of GDA with momentum that one would implement in practice.

4. Global Convergence Properties of PGDAm

In this section, we analyze the global convergence properties of PGDAm under the standard Assumption 1.

Algorithm 1 Proximal-GDA with Momentum (PGDAm)

Input: Initialization x_0, y_0, learning rates η_x, η_y, momentum parameters β, γ.

for $t = 0, 1, 2, \ldots, T-1$ **do**

$$\begin{aligned} \widetilde{x}_t &= x_t + \beta(x_t - x_{t-1}), \\ x_{t+1} &\in \mathrm{prox}_{\eta_x g}\big(\widetilde{x}_t - \eta_x \nabla_1 f(x_t, y_t)\big), \\ \widetilde{y}_t &= y_t + \gamma(y_t - y_{t-1}), \\ y_{t+1} &= \mathrm{prox}_{\eta_y h}\big(\widetilde{y}_t + \eta_y \nabla_2 f(x_t, \widetilde{y}_t)\big). \end{aligned}$$

end

Output: x_T, y_T.

In the recent work (Chen et al., 2021), it has been shown that the basic proximal-GDA algorithm (without momentum) admits a Lyapunov function that monotonically decreases along the optimization path. Importantly, this decreasing property of the Lyapunov function implies strong convergence guarantees of the proximal-GDA. For our PGDAm, as the momentum schemes are known to cause oscillations in conventional gradient-based optimization, it might appear that PGDAm may not admit a proper Lyapunov function. Interestingly, our next result does identify an intrinsic Lyapunov function for PGDAm that takes a special form.

Proposition 2. *Let Assumption 1 hold. Denote* $z_t := (x_t, y_t, x_{t-1})$ *and define the Lyapunov function*

$$H(z_t) := \Phi(x_t) + g(x_t) + \big(1 - \frac{1}{4\kappa}\big)\|y_t - y^*(x_t)\|^2 + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\|^2.$$

Choose learning rates $\eta_x \le \frac{1}{4(4L\kappa + L^2\kappa + 4\kappa^{\frac{5}{2}})}$, $\eta_y \le \frac{1}{L}$ *and momentum parameters* $\beta \le \frac{1}{4}, \gamma = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. *Then, the variables* $z_t = (x_t, y_t, x_{t-1})$ *generated by PGDAm satisfy, for all* $t = 0, 1, 2, ...$

$$\begin{aligned} H(z_{t+1}) \le & H(z_t) - L\kappa\|x_{t+1} - x_t\|^2 - \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 \\ & - \frac{1}{4\kappa}\big(\|y^*(x_t) - y_t\|^2 + \|y^*(x_{t+1}) - y_{t+1}\|^2\big). \end{aligned} \tag{2}$$

The special Lyapunov function in Proposition 2 tracks the optimization progress made by PGDAm. To elaborate, the Lyapunov function $H(z_t)$ consists of three components: the objective function $\Phi(x_t) + g(x_t)$, the error term $(1 - \frac{1}{4\kappa})\|y_t - y^*(x_t)\|^2$ that tracks the optimality gap of the y update, and the increment term $\frac{\beta}{\eta_x}\|x_t - x_{t-1}\|^2$ that is induced by the momentum updates. As a comparison, the Lyapunov function of the basic proximal-GDA algorithm does not involve the last increment term. For PGDAm, its Lyapunov function implicitly represents the minimax optimization goal. Specifically, if $H(z_t)$ keeps decreasing to a local minimum, then we will reach a local minimizer x^* of $\Phi(x) + g(x)$ and also have $\|y_t - y^*(x^*)\| \to 0, \|x_t - x_{t-1}\| \to 0$.

Remark 1. We note that in Proposition 2, our choices of the momentum parameters, i.e., $\beta \le \frac{1}{4}, \gamma = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ are standard choices for the heavy-ball momentum and Nesterov's momentum, respectively. This implies that the practical implementation of PGDAm is as simple as that of standard proximal gradient algorithms with momentum.

Remark 2. In (Chen et al., 2021), the learning rate of proximal-GDA must satisfy $\eta_x \le O(\kappa^{-3})$ in order to guarantee the decreasing property of the Lyapunov function. As a comparison, PGDAm only requires $\eta_x \le O(\kappa^{-2.5})$ to guarantee the same property. This is because the Nesterov's momentum leads to an orderwise improved convergence rate of the proximal gradient ascent step under strong concavity.

Dependence of η_x on κ: In the existing analysis of GDA in nonconvex-strongly-concave optimization (Lin et al., 2020b), gradient norm convergence is established under the choice of learning rate $\eta_x \le O(\kappa^{-2})$ without relying on the decreasing Lyapunov function. As a comparison, PGDAm needs a smaller learning rate $\eta_x \le O(\kappa^{-2.5})$ to guarantee the monotonic decreasing property of the Lyapunov function, which leads to stronger convergence guarantees as we show later.

By telescoping eq. (2), we obtain the following asymptotic properties of the variable sequences generated by PGDAm.

Corollary 1. *Under the same conditions as those of Proposition 2, the variable sequences* $\{x_t, y_t\}_t$ *generated by PGDAm satisfy*

$$\begin{aligned} \lim_{t\to\infty} \|x_{t+1} - x_t\| &= 0, \\ \lim_{t\to\infty} \|y_{t+1} - y_t\| &= 0, \\ \lim_{t\to\infty} \|y_t - y^*(x_t)\| &= 0. \end{aligned}$$

Corollary 1 establishes the asymptotic stability of the variable sequences generated by PGDAm. Such a result corresponds to the gradient norm convergence result obtained in the existing analysis of GDA in nonconvex minimax optimization (Lin et al., 2020b). However, these properties do not have implications on the convergence of the variable sequences. Next, by leveraging the monotonic decreasing property of the Lyapunov function and the structure of the momentum updates, we establish the following global convergence result of PGDAm.

Theorem 1 (Global convergence)**.** *Let Assumption 1 hold and choose the learning rates as those specified in Proposition 2. Then, the sequences generated by PGDAm satisfies the following global convergence properties.*

1. *The function value sequence* $\{(\Phi+g)(x_t)\}_t$ *converges to a finite limit* $H^* > -\infty$*;*
2. *The sequences* $\{x_t\}_t, \{y_t\}_t$ *are bounded and have compact sets of limit points. Moreover, for any limit point* x^* *of* $\{x_t\}_t$ *we have* $(\Phi+g)(x^*) \equiv H^*$*;*
3. *Every limit point of* $\{x_t\}_t$ *is a critical point of* $(\Phi+g)(x)$.

To elaborate, Theorem 1 shows that the variable sequences generated by PGDAm approach a set of critical points of the minimax problem (item 3), at which the objective function achieves a constant value (items 1 and 2). Therefore, under momentum acceleration, PGDAm will eventually enter a local parameter region around a set of critical points. With this global convergence result, we are further motivated to exploit the local nonconvex function geometry around the critical points to characterize stronger asymptotic convergence behaviors of PGDAm in the next section.

5. Analysis of PGDAm under KŁ Geometry

In this section, we establish stronger convergence guarantees for PGDAm under the nonconvex local Kurdyka-Łojasiewicz (KŁ) geometry of the function. We first introduce the definition of the KŁ geometry.

5.1. Kurdyka-Łojasiewicz Geometry

General nonconvex functions typically do not have a global geometry, yet they may satisfy certain local geometries around the critical points. In particular, the Kurdyka-Łojasiewicz (KŁ) geometry characterizes a broad spectrum of local geometries of nonconvex functions (Bolte et al., 2007; 2014). It generalizes many conventional global geometries such as the strong convexity and Polyak-Łojasiewicz geometry. Hence, it is much desired to study the convergence of PGDAm under the general KŁ geometry, which we elaborate below. Throughout, the point-to-set distance is denoted as $\mathrm{dist}_\Omega(x) := \inf_{u\in\Omega}\|x-u\|$.

Definition 2 (KŁ geometry, (Bolte et al., 2014))**.** *A proper and lower semi-continuous function* h *is said to have the KŁ geometry if for every compact set* $\Omega \subset \mathrm{dom}h$ *on which* h *takes a constant value* $h_\Omega \in \mathbb{R}$*, there exist* $\varepsilon, \lambda > 0$ *such that for all* $\bar{x} \in \Omega$ *and all* $x \in \{z \in \mathbb{R}^m : \mathrm{dist}_\Omega(z) < \varepsilon, h_\Omega < h(z) < h_\Omega + \lambda\}$*, the following condition holds:*

$$\varphi'\left(h(x)-h_\Omega\right)\cdot \mathrm{dist}_{\partial h(x)}(\mathbf{0}) \geq 1, \tag{3}$$

where φ' *is the derivative of function* $\varphi : [0,\lambda) \to \mathbb{R}_+$*, which takes the form* $\varphi(t) = \frac{c}{\theta}t^\theta$ *for certain universal constant* $c > 0$ *and KŁ parameter* $\theta \in (0,1]$.

The KŁ geometry can be understood as a certain type of local gradient dominance when the function h is differentiable. To see this, when h is differentiable so that $\partial h(x) = \nabla h(x)$, the KŁ inequality in eq. (3) reduces to $h(x)-h_\Omega \leq \mathcal{O}(\|\nabla h(x)\|^{\frac{1}{1-\theta}})$, which is a generalization of the Polyak-Łojasiewicz (PL) condition (with KŁ parameter $\theta = \frac{1}{2}$) (Łojasiewicz, 1963; Karimi et al., 2016).

In the existing studies, a large class of functions has been shown to have the local KŁ geometry. Examples include sub-analytic functions, logarithm and exponential functions and semi-algebraic functions, etc. These function classes cover most of the nonconvex objective functions encountered in practical machine learning applications (Zhou et al., 2016b; Yue et al., 2019; Zhou & Liang, 2017). Moreover, the KŁ geometry has been exploited to analyze the variable convergence of various gradient-based algorithms in nonconvex optimization, e.g., gradient descent (Attouch & Bolte, 2009; Li et al., 2017), accelerated gradient method (Zhou et al., 2020), alternating minimization (Bolte et al., 2014) and distributed gradient methods (Zhou et al., 2016a). In these studies, it has been shown that the variable sequences generated by these algorithms converge to a desired critical point under the nonconvex KŁ geometry, and the asymptotic convergence rates critically depend on the parameterization θ of the KŁ geometry. In the following subsections, we provide a comprehensive understanding of the convergence and convergence rate of PGDAm under the KŁ geometry.

5.2. Variable Convergence of PGDAm

We already show in Theorem 1 that every limit point of $\{x_t\}_t$ generated by PGDAm is a critical point of the minimax problem (P). However, it is possible that the variable sequence $\{x_t\}_t$ has multiple limit points and hence diverges. In this subsection, we exploit the KŁ geometry of the Lyapunov function to formally prove the variable convergence of PGDAm.

Throughout this section, we adopt the following assumption that $y^*(x)$ is a sub-differentiable mapping, which has also been considered in (Chen et al., 2021).

Assumption 2. *The function* $v(x) := \|y^*(x)-y\|^2$ *has a non-empty subdifferential, i.e.,* $\partial v(x) \neq \emptyset$.

We note that this is a mild assumption on $y^*(x)$ that allows it to be non-differentiable. Then, we obtain the following variable convergence result of PGDAm under the KŁ geometry.

Theorem 2 (Variable convergence)**.** *Let Assumptions 1 and 2 hold and assume that* $H(z)$ *has the KŁ geometry. Choose the learning rates as specified in Proposition 2. Then, the sequences* $\{(x_t,y_t)\}_t$ *generated by PGDAm converge to a certain critical point* $(x^*, y^*(x^*))$ *of* $(\Phi+g)(x)$*, i.e.,*

$$x_t \xrightarrow{t} x^*, \quad y_t \xrightarrow{t} y^*(x^*).$$

Theorem 2 formally shows that the variable sequences generated by GDA with momentum are guaranteed to converge

to a critical point $(x^*, y^*(x^*))$ of the minimax problem (P). In the proof of this result, we leverage the local KŁ geometry of the Lyapunov function to regularize the trajectory length of PGDAm, i.e., we show that the variable trajectory is absolutely summable, i.e.,

$$\sum_{t=0}^{\infty} \|x_{t+1} - x_t\| < +\infty,$$

which is much stronger than the property $\sum_{t=0}^{\infty} \|x_{t+1} - x_t\|^2 < +\infty$ proved in the proof of Corollary 1. This shows that the local KŁ geometry actually strengthens the asymptotic convergence behavior of the algorithm. To summarize, our result shows that the combination of the Nesterov's momentum and the heavy-ball momentum in PGDAm can provide acceleration with guaranteed variable convergence.

5.3. Convergence Rates of PGDAm

In this subsection, we establish the asymptotic convergence rates of PGDAm under different parameterizations of the KŁ geometry. In the definition of the KŁ geometry, the parameter θ characterizes the sharpness of the local geometry around the critical points. Intuitively, a large θ implies that the local geometry of is sharp and hence should lead to a fast convergence of the algorithm. Throughout, we denote $t_0 \in \mathbb{N}$ as a sufficiently large positive integer, denote $c > 0$ as the constant in Definition 2 and define the following universal constant $M = \max\Big\{\frac{1}{L\kappa}\Big(\frac{1}{\eta_x} + (L+4\kappa^2)(1+\kappa) + \frac{4\beta}{\eta_x}\Big)^2, \frac{2\beta}{\eta_x}, 4\kappa(L+4\kappa)^2\Big\}$.

We first obtain the following asymptotic convergence rates of the Lyapunov function of PGDAm under different parameterizations θ of the KŁ geometry.

Theorem 3 (Function value convergence rate)**.** *Under the same conditions as those of Theorem 2, the Lyapunov function value sequence $\{H(z_t)\}_t$ converges to the limit H^* at the following rates.*

1. *If KŁ geometry holds with $\theta = 1$, then $H(z_t) \downarrow H^*$ within finite number of iterations;*
2. *If $\theta \in (\frac{1}{2}, 1)$, then $H(z_t) \downarrow H^*$ super-linearly: $\forall t \geq t_0$*

$$H(z_t) - H^* \leq (3Mc^2)^{-\frac{1}{2\theta-1}} \exp\Big(-\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_0}\Big);$$

3. *If $\theta = \frac{1}{2}$, then $H(z_t) \downarrow H^*$ linearly: $\forall t \geq t_0$*

$$H(z_t) - H^* \leq \Big(1 + \frac{1}{3Mc^2}\Big)^{t_0-t}(H(z_{t_0}) - H^*);$$

4. *If $\theta \in (0, \frac{1}{2})$, then $H(z_t) \downarrow H^*$ sub-linearly: $\forall t \geq t_0$*

$$H(z_t) - H^* \leq \mathcal{O}\Big((t-t_0)^{-\frac{1}{1-2\theta}}\Big).$$

The above function value convergence rates are orderwise same as those of the proximal-GDA in (Chen et al., 2021). This is expected in the general nonconvex minimax optimization. Regarding the results, under a sharper local KŁ geometry (i.e., a larger value of θ), PGDAm achieves an order-wise faster convergence rate.

Next, we further obtain the following asymptotic convergence rates of the variable sequences generated by PGDAm.

Theorem 4 (Variable convergence rate)**.** *Under the same conditions as those of Theorem 2, the sequences $\{x_t, y_t\}_t$ converge to their limits $x^*, y^*(x^*)$ respectively at the following rates, where we denote $d_t := \max\{\|x_t - x^*\|, \|y_t - y^*(x^*)\|\}$.*

1. *If KŁ geometry holds with $\theta = 1$, then $(x_t, y_t) \to (x^*, y^*(x^*))$ within finite number of iterations;*
2. *If $\theta \in (\frac{1}{2}, 1)$, then $(x_t, y_t) \to (x^*, y^*(x^*))$ super-linearly:*

$$d_t \leq \mathcal{O}\Big(\exp\Big(-\big(\frac{1}{2(1-\theta)}\big)^{t-t_0}\Big)\Big), \forall t \geq t_0;$$

3. *If $\theta = \frac{1}{2}$, then $(x_t, y_t) \to (x^*, y^*(x^*))$ linearly:*

$$d_t \leq \mathcal{O}\Big(\Big(\min\Big\{2, 1 + \frac{1}{2Mc^2}\Big\}\Big)^{(t_0-t)/2}\Big), \forall t \geq t_0;$$

4. *If $\theta \in (0, \frac{1}{2})$, then $(x_t, y_t) \to (x^*, y^*(x^*))$ sub-linearly:*

$$d_t \leq \mathcal{O}\Big((t-t_0)^{-\frac{\theta}{1-2\theta}}\Big), \forall t \geq t_0.$$

6. Experiments

In this section, we compare the performance of PGDAm with that of other GDA-type algorithms via numerical experiments. Specifically, we compare PGDAm with the standard proximal-GDA algorithm (Chen et al., 2021) and the single-loop accelerated GDA algorithm (APDA) (Zhu et al., 2020). All these algorithms have a single-loop structure.

6.1. Synthetic Minimax Optimization Problem

We first consider the following simple strongly-convex-strongly-concave minimax optimization problem, whose Lyapunov function can be analytically evaluated.

$$\min_{x\in\mathbb{R}^{12}} \max_{y\in\mathbb{R}^{12}} f(x,y) = \frac{\|x\|^2}{2} - \frac{\lambda\|y\|^2}{2} + \langle x, y\rangle, \quad (4)$$

where we set $\lambda = 0.1$. For this problem, it is easy to check that $\Phi(x) = \frac{x^2}{2}(1 + \frac{1}{\lambda})$, $y^*(x) = \frac{x}{\lambda}$. Moreover, $f(\cdot,\cdot)$ is $L = 2$-smooth and $f(x,\cdot)$ is $\mu = \lambda$ strongly-concave, and hence the condition number is $\kappa = \frac{2}{\lambda}$. Given these information, we can analytically evaluate the Lyapunov function defined in Proposition 2.

In the experiment, we choose the same learning rates $\eta_x = \kappa^{-2.5}, \eta_y = \frac{1}{L}$ (suggested by our theory) for all the three algorithms. For our PGDAm, we choose momentum $\beta = \frac{1}{4}, \gamma = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ as suggested by our theory. For APDA, we choose the fine-tuned extra hyper-parameter parameter $\eta = 7 \times 10^{-3}$. No proximal mapping is needed for this problem.

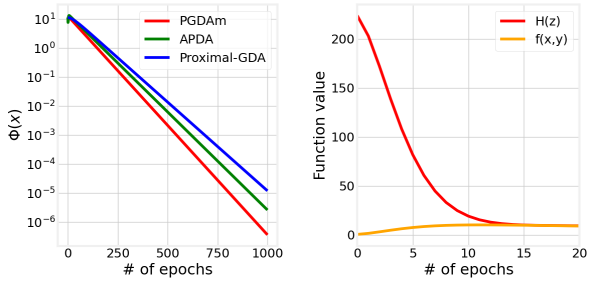

Figure 1. Left: comparison of $\Phi(x)$ of all three algorithms. Right: comparison between $H(z)$ and $f(x,y)$ of PGDAm.

Figure 1 (Left) compares the objective function value $\Phi(x)$ achieved by all the three algorithms. It can be seen that our PGDAm achieves the fastest convergence among these algorithms. Moreover, in Figure 1 (Right), we compare the Lyapunov function value $H(z)$ and the minimax function value $f(x,y)$ in the training process of our PGDAm. It can be seen that $H(z)$ does decreases monotonically while $f(x,y)$ does not. This verifies our theoretical result in Proposition 2.

6.2. Regularized Wasserstein Robustness Model

We consider solving the following regularized Wasserstein robustness model (WRM) (Sinha et al., 2018) using the MNIST dataset (Lecun et al., 1998).

$$\min_{\theta} \max_{\{\xi_i\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^{n}\Big[\ell(h_\theta(\xi_i), y_i) - \lambda\|\xi_i - x_i\|^2\Big] - \lambda_1 \sum_{i=1}^{n}\|\xi_i\|_1 + \frac{\lambda_2}{2}\|\theta\|^2, \quad (5)$$

where $n = 60$k is the number of training samples, θ is the model parameter, (x_i, y_i) corresponds to the i-th data sample and label, respectively, and ξ_i is the adversarial sample corresponding to x_i. We choose the cross-entropy loss function ℓ. We add an ℓ_1 regularization to impose sparsity on the adversarial examples, and add an ℓ_2^2 regularization to prevent divergence of the model parameters.

We set $\lambda = 1.3$ that suffices to make the maximization part be strongly-concave, and set $\lambda_1 = \lambda_2 = 10^{-4}$. We use a convolution network that consists of two convolution blocks followed by two fully connected layers. Specifically, each convolution block contains a convolution layer, a max-pooling layer with stride step 2, and a ReLU activation layer. The convolution layers in the two blocks have $1, 10$ input channels and $10, 20$ output channels, respectively, and both of them have kernel size 5, stride step 1 and no padding. The two fully connected layers have input dimensions $320, 50$ and output dimensions $50, 10$, respectively.

We implement all three algorithms using stochastic gradients with batch size 1k, as it requires too much memory to find adversarial examples for all 60k samples in one iteration. We choose the same learning rates $\eta_x = \eta_y = 10^{-3}$ for all algorithms. For PGDAm, we choose momentum $\beta = \gamma = 0.75$. For APDA, we choose the fine-tuned $\eta = 2\eta_x$. As the function $\Phi(x)$ cannot be exactly evaluated, we run 100 steps of stochastic gradient ascent updates with learning rate 0.1 to maximize $f(x_t, y) - h(y)$ and obtain an approximated $y^*(x_t)$, which is used to estimate $\Phi(x) + g(x)$.

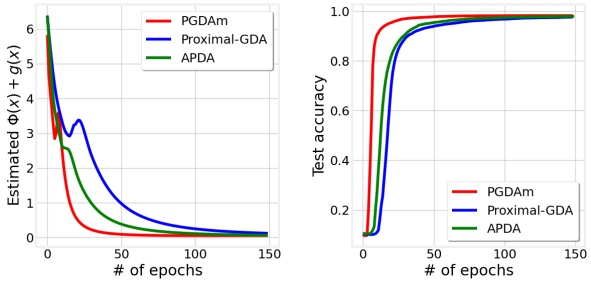

Figure 2. Left: comparison of $\Phi(x) + g(x)$ of all three algorithms. Right: comparison of the corresponding test accuracy.

Figure 2 (Left) compares the estimated objective function value achieved by all the three algorithms. It can be seen that PGDAm achieves the fastest convergence among these algorithms and is significantly faster than the proximal-GDA. This demonstrates the effectiveness of our simple momentum scheme. Figure 2 (Right) further demonstrates the advantage of PGDAm in terms of the test accuracy.

7. Conclusion

We develop a simple momentum scheme for boosting the convergence of proximal-GDA in nonconvex minimax optimization. Our momentum design has a single-loop structure and can be efficiently implemented in practice. We show that PGDAm admits an intrinsic Lyapunov function that monotonically decreases in the minimax optimization process. Such a key property further implies the variable convergence of PGDAm to a critical point under the KŁ geometry. Our study extends the applicability of the conventional momentum schemes (heavy-ball and Nesterov's momentum) for nonconvex minimization algorithms to nonconvex minimax algorithms. We expect that such a simple and effective algorithm will be widely applied in machine learning practice to accelerate solving minimax optimization tasks.

References

Adolphs, L., Daneshmand, H., Lucchi, A., and Hofmann, T. Local saddle point optimization: A curvature exploitation approach. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 486–495, 2019.

Attouch, H. and Bolte, J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009. ISSN 0025-5610.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, March 2009.

Bolte, J., Daniilidis, A., and Lewis, A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17:1205–1223, 2007.

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

Chen, Z., Zhou, Y., Xu, T., and Liang, Y. Proximal gradient descent-ascent: Variable convergence under kł geometry. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

Cherukuri, A., Gharesifard, B., and Cortes, J. Saddle-point dynamics: conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, 55 (1):486–511, 2017.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9236–9246, 2018.

Du, S. S. and Hu, W. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 196–205, 2019.

Ferreira, M. A. M., Andrade, M., Matos, M. C. P., Filipe, J. A., and Coelho, M. P. Minimax theorem and nash equilibrium. 2012.

Frankel, P., Garrigos, G., and Peypouquet, J. Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, Jun 2015.

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, March 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2672–2680, 2014.

Huang, F., Gao, S., Pei, J., and Huang, H. Accelerated zeroth-order momentum methods from mini to minimax optimization. *ArXiv:2008.08170*, 2020.

Jin, C., Netrapalli, P., and Jordan, M. I. What is local optimality in nonconvex-nonconcave minimax optimization? pp. 4880–4889, 2020.

Karimi, H., Nutini, J., and Schmidt, M. *Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition*, pp. 795–811. 2016.

Kruger, A. Y. On fréchet subdifferentials. *Journal of Mathematical Sciences*, 116(3):3325–3358, 2003.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pp. 2278–2324, 1998.

Li, H. and Lin, Z. Accelerated proximal gradient methods for nonconvex programming. In *Proc. International Conference on Neural Information Processing Systems (Neurips)*, pp. 379–387, 2015.

Li, Q., Zhou, Y., Liang, Y., and Varshney, P. K. Convergence analysis of proximal gradient with momentum for nonconvex optimization. In *Proc. 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 2111–2119, Aug 2017.

Lin, T., Jin, C., and Jordan, M. I. Near-optimal algorithms for minimax optimization. In *Proc. Annual Conference on Learning Theory (COLT)*, pp. 2738–2779, 2020a.

Lin, T., Jin, C., and Jordan, M. I. On gradient descent ascent for nonconvex-concave minimax problems. pp. 6083–6093, 2020b.

Lions, P.-L. and Mercier, B. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

Łojasiewicz, S. A topological property of real analytic subsets. *Coll. du CNRS, Les equations aux derivees partielles*, pp. 87—-89, 1963.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In

Proc. International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1497–1507, 2020.

Nedić, A. and Ozdaglar, A. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2014.

Neumann, J. v. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

Noll, D. and Rondepierre, A. Convergence of linesearch and trust-region methods using the Kurdyka–Łojasiewicz inequality. In *Proc. Computational and Analytical Mathematics*, pp. 593–611, 2013.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14934–14942, 2019.

Ochs, P. Local convergence of the heavy-ball method and iPiano for non-convex optimization. *Journal of Optimization Theory and Applications*, 177(1):153–180, Apr 2018.

Ochs, P., Chen, Y., Brox, T., and Pock, T. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

Polyak, B. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Qiu, S., Yang, Z., Wei, X., Ye, J., and Wang, Z. Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear td learning. *ArXiv:2008.10103*, 2020.

Robinson, J. An iterative method of solving a game. *Annals of mathematics*, 54(2):296–301, 1951.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Sinha, A., Namkoong, H., and Duchi, J. C. Certifying some distributional robustness with principled adversarial training. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

Sinha, A., Namkoong, H., and Duchi, J. Certifiable distributional robustness with principled adversarial training. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Efficient algorithms for smooth minimax optimization. In *Proc. Advances in Neural Information Processing Systems (Neurips)*, pp. 12680–12691. 2019.

Tseng, P. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, Oct 2010.

Wang, Y. and Li, J. Improved algorithms for convex-concave minimax optimization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Xie, G., Luo, L., Lian, Y., and Zhang, Z. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. pp. 10504–10513, 2020.

Xu, T., Wang, Z., Liang, Y., and Poor, H. V. Enhanced first and zeroth order variance reduced algorithms for min-max optimization. *arXiv:2006.09361*, 2020a.

Xu, Z., Zhang, H., Xu, Y., and Lan, G. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *ArXiv:2006.02032*, 2020b.

Yang, J., Kiyavash, N., and He, N. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yue, M.-C., Zhou, Z., and Man-Cho So, A. On the quadratic convergence of the cubic regularization method under a local error bound condition. *SIAM Journal on Optimization*, 29(1):904–932, 2019.

Zhang, G. and Wang, Y. On the suboptimality of negative momentum for minimax optimization. *ArXiv:2008.07459*, 2020.

Zhou, Y. and Liang, Y. Characterization of Gradient Dominance and Regularity Conditions for Neural Networks. *ArXiv:1710.06910v2*, Oct 2017.

Zhou, Y., Yu, Y., Dai, W., Liang, Y., and Xing, P. On convergence of model parallel proximal gradient algorithm for stale synchronous parallel system. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51, pp. 713–722, May 2016a.

Zhou, Y., Zhang, H., and Liang, Y. Geometrical properties and accelerated gradient solvers of non-convex phase retrieval. In *Proc. 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 331–335, 2016b.

Zhou, Y., Liang, Y., Yu, Y., Dai, W., and Xing, E. P. Distributed Proximal Gradient Algorithm for Partially Asynchronous Computer Clusters. *Journal of Machine Learning Research (JMLR)*, 19(19):1–32, 2018a.

Zhou, Y., Wang, Z., and Liang, Y. Convergence of cubic regularization for nonconvex optimization under kl property. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3760–3769, 2018b.

Zhou, Y., Wang, Z., Ji, K., Liang, Y., and Tarokh, V. Proximal gradient algorithm with momentum and flexible parameter restart for nonconvex optimization. In *Proc. International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 1445–1451, 7 2020.

Zhu, Y., Liu, D., and Tran-Dinh, Q. Accelerated Primal-Dual Algorithms for a Class of Convex-Concave Saddle-Point Problems with Non-Bilinear Coupling Term. *arXiv:2006.09263*, June 2020.

Appendix

Table of Contents

A. Proof of Proposition 2

Proposition 2. *Let Assumption 1 hold. Denote $z_t := (x_t, y_t, x_{t-1})$ and define the Lyapunov function*

$$H(z_t) := \Phi(x_t) + g(x_t) + \big(1 - \frac{1}{4\kappa}\big)\|y_t - y^*(x_t)\|^2 + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\|^2.$$

Choose learning rates $\eta_x \leq \frac{1}{4(4L\kappa + L^2\kappa + 4\kappa^{\frac{5}{2}})}$, $\eta_y \leq \frac{1}{L}$ and momentum parameters $\beta \leq \frac{1}{4}, \gamma = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. Then, the variables $z_t = (x_t, y_t, x_{t-1})$ generated by PGDAm satisfy, for all $t = 0, 1, 2, ...$

$$\begin{aligned} H(z_{t+1}) \leq & H(z_t) - L\kappa\|x_{t+1} - x_t\|^2 - \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 \\ & - \frac{1}{4\kappa}\big(\|y^*(x_t) - y_t\|^2 + \|y^*(x_{t+1}) - y_{t+1}\|^2\big). \end{aligned} \tag{2}$$

Proof. Consider the t-th iteration of PGDAm. As the function Φ is $L(1+\kappa)$-smooth, we obtain that

$$\Phi(x_{t+1}) \leq \Phi(x_t) + \langle x_{t+1} - x_t, \nabla\Phi(x_t)\rangle + \frac{L(1+\kappa)}{2}\|x_{t+1} - x_t\|^2. \tag{6}$$

On the other hand, by the definition of the proximal gradient step of x_t, we have that

$$g(x_{t+1}) + \frac{1}{2\eta_x}\|x_{t+1} - \widetilde{x}_t + \eta_x\nabla_1 f(x_t, y_t)\|^2 \leq g(x_t) + \frac{1}{2\eta_x}\|x_t - \widetilde{x}_t + \eta_x\nabla_1 f(x_t, y_t)\|^2,$$

which further simplifies to

$$\begin{aligned} g(x_{t+1}) &\leq g(x_t) + \frac{1}{2\eta_x}\|x_t - \widetilde{x}_t\|^2 + \langle x_t - \widetilde{x}_t, \nabla_1 f(x_t, y_t)\rangle \\ &\quad - \frac{1}{2\eta_x}\|x_{t+1} - \widetilde{x}_t\|^2 - \langle x_{t+1} - \widetilde{x}_t, \nabla_1 f(x_t, y_t)\rangle \\ &\overset{(i)}{\leq} g(x_t) + \frac{\beta^2}{2\eta_x}\|x_t - x_{t-1}\|^2 - \frac{1}{2\eta_x}\|x_{t+1} - x_t - \beta(x_t - x_{t-1})\|^2 \\ &\quad + \langle x_t - x_{t+1}, \nabla_1 f(x_t, y_t)\rangle \end{aligned}$$

$$
\begin{aligned}
&= g(x_t) + \frac{\beta^2}{2\eta_x}\|x_t - x_{t-1}\|^2 - \frac{1}{2\eta_x}\|x_{t+1} - x_t\|^2 - \frac{\beta^2}{2\eta_x}\|x_t - x_{t-1}\|^2 \\
&\quad + \frac{\beta}{\eta_x}\langle x_{t+1} - x_t, x_t - x_{t-1}\rangle + \langle x_t - x_{t+1}, \nabla_1 f(x_t, y_t)\rangle \\
&\le g(x_t) - \frac{1}{2\eta_x}\|x_{t+1} - x_t\|^2 \\
&\quad + \frac{\beta}{2\eta_x}\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 + \langle x_t - x_{t+1}, \nabla_1 f(x_t, y_t)\rangle, \qquad (7)
\end{aligned}
$$

where (i) uses the fact that $x_t - \widetilde{x}_t = \beta(x_{t-1} - x_t)$. Adding up eq. (6) and eq. (7) yields that

$$
\begin{aligned}
&\Phi(x_{t+1}) + g(x_{t+1}) \\
&\le \Phi(x_t) + g(x_t) - \Big(\frac{1}{2\eta_x} - \frac{L(1+\kappa)}{2}\Big)\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 \\
&\quad + \langle x_{t+1} - x_t, \nabla\Phi(x_t) - \nabla_1 f(x_t, y_t)\rangle \\
&\le \Phi(x_t) + g(x_t) - \Big(\frac{1}{2\eta_x} - \frac{L(1+\kappa)}{2}\Big)\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 \\
&\quad + \|x_{t+1} - x_t\|\|\nabla\Phi(x_t) - \nabla_1 f(x_t, y_t)\| \\
&\le \Phi(x_t) + g(x_t) - \Big(\frac{1}{2\eta_x} - \frac{L(1+\kappa)}{2}\Big)\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 \\
&\quad + \|x_{t+1} - x_t\|\|\nabla_1 f(x_t, y^*(x_t)) - \nabla_1 f(x_t, y_t)\| \\
&\le \Phi(x_t) + g(x_t) - \Big(\frac{1}{2\eta_x} - \frac{L(1+\kappa)}{2}\Big)\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 \\
&\quad + L\|x_{t+1} - x_t\|\|y^*(x_t) - y_t\| \\
&\le \Phi(x_t) + g(x_t) - \Big(\frac{1-\beta}{2\eta_x} - \frac{L(1+\kappa)}{2} - \frac{L^2\kappa}{2}\Big)\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 \\
&\quad + \frac{1}{2\kappa}\|y^*(x_t) - y_t\|^2. \qquad (8)
\end{aligned}
$$

Next, consider the term $\|y^*(x_t) - y_t\|$ in the above inequality. Note that $y^*(x_t)$ is the unique minimizer of the strongly concave function $f(x_t, y) - h(y)$, and y_{t+1} is obtained by applying one proximal gradient step with Nesterov's momentum. Hence, by the convergence rate of accelerated proximal gradient ascent algorithm under strong concavity with the choices of $\eta_y \le \frac{1}{L}$, $\gamma = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ (Nesterov, 2014), we conclude that

$$
\|y_{t+1} - y^*(x_t)\|^2 \le \big(1 - \kappa^{-\frac{1}{2}}\big)\|y_t - y^*(x_t)\|^2. \qquad (9)
$$

Hence, we further obtain that

$$
\begin{aligned}
\|y^*(x_{t+1}) - y_{t+1}\|^2 &\le \big(1 + \kappa^{-\frac{1}{2}}\big)\|y_{t+1} - y^*(x_t)\|^2 + (1 + \kappa^{\frac{1}{2}})\|y^*(x_{t+1}) - y^*(x_t)\|^2 \\
&\le \big(1 - \kappa^{-1}\big)\|y_t - y^*(x_t)\|^2 + \kappa^2(1 + \kappa^{\frac{1}{2}})\|x_{t+1} - x_t\|^2. \qquad (10)
\end{aligned}
$$

Adding up eqs. (8) & (10) yields that

$$
\begin{aligned}
&\Phi(x_{t+1}) + g(x_{t+1}) \\
&\le \Phi(x_t) + g(x_t) - \Big(\frac{1-\beta}{2\eta_x} - \frac{L(1+\kappa)}{2} - \frac{L^2\kappa}{2} - \kappa^2(1 + \kappa^{\frac{1}{2}})\Big)\|x_{t+1} - x_t\|^2 \\
&\quad + \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 + \Big(1 - \frac{1}{2\kappa}\Big)\|y^*(x_t) - y_t\|^2 - \|y^*(x_{t+1}) - y_{t+1}\|^2.
\end{aligned}
$$

Rearranging the equation above and recalling the definition of the Lyapunov function $H(z_t) = \Phi(x_t) + g(x_t) + \Big(1 - \frac{1}{4\kappa}\Big)\|y_t - y^*(x_t)\|^2 + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\|^2$, we have

$$\begin{aligned} H(z_{t+1}) \leq & H(z_t) - \Big(\frac{1-3\beta}{2\eta_x} - \frac{L(1+\kappa)}{2} - \frac{L^2\kappa}{2} - \kappa^2(1+\kappa^{\frac{1}{2}})\Big)\|x_{t+1} - x_t\|^2 \\ & - \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 - \frac{1}{4\kappa}(\|y^*(x_t) - y_t\|^2 + \|y^*(x_{t+1}) - y_{t+1}\|^2). \end{aligned} \tag{11}$$

Choosing $\beta < \frac{1}{4}, \eta_x < \frac{1}{4(4L\kappa + L^2\kappa + 4\kappa^{\frac{5}{2}})}$ and using $\kappa \geq 1$ yields that

$$\begin{aligned} & \frac{1-3\beta}{2\eta_x} - \frac{L(1+\kappa)}{2} - \frac{L^2\kappa}{2} - \kappa^2(1+\kappa^{\frac{1}{2}}) \\ & \geq \frac{1}{8\eta_x} - L\kappa - \frac{L^2\kappa}{2} - 2\kappa^{\frac{5}{2}} \\ & \geq L\kappa. \end{aligned} \tag{12}$$

As a result, eq. (2) can be concluded by substituting eq. (12) into eq. (11).

□

B. Proof of Corollary 1

Corollary 1. *Under the same conditions as those of Proposition 2, the variable sequences $\{x_t, y_t\}_t$ generated by PGDAm satisfy*

$$\begin{aligned} \lim_{t\to\infty} \|x_{t+1} - x_t\| &= 0, \\ \lim_{t\to\infty} \|y_{t+1} - y_t\| &= 0, \\ \lim_{t\to\infty} \|y_t - y^*(x_t)\| &= 0. \end{aligned}$$

Proof. To prove the first and third items of Corollary 1, summing the inequality of Proposition 2 over $t = 0, 1, ..., T-1$, we obtain that for all $T \geq 1$,

$$\begin{aligned} & \sum_{t=0}^{T-1} \Big[L\kappa\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 + \frac{1}{4\kappa}(\|y_{t+1} - y^*(x_{t+1})\|^2 + \|y_t - y^*(x_t)\|^2)\Big] \\ & \leq H(z_0) - H(z_T) \\ & \leq H(z_0) - \big[\Phi(x_T) + g(x_T)\big] \\ & \leq H(z_0) - \inf_{x\in\mathbb{R}^m}\big(\Phi(x) + g(x)\big) < +\infty. \end{aligned}$$

Letting $T \to \infty$, we conclude that

$$\sum_{t=0}^{\infty} \Big[L\kappa\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 + \frac{1}{4\kappa}(\|y_{t+1} - y^*(x_{t+1})\|^2 + \|y_t - y^*(x_t)\|^2)\Big] < +\infty.$$

Therefore, we must have $\lim_{t\to\infty}\|x_{t+1} - x_t\| = \lim_{t\to\infty}\|y_t - y^*(x_t)\| = 0$.

To prove the second item, note that

$$\|y_{t+1} - y_t\| \leq \|y_{t+1} - y^*(x_t)\| + \|y_t - y^*(x_t)\| \overset{eq.\ (9)}{\leq} (\sqrt{1-\kappa^{-\frac{1}{2}}} + 1)\|y_t - y^*(x_t)\| \xrightarrow{t} 0.$$

□

C. Proof of Theorem 1

Theorem 1 (Global convergence)**.** *Let Assumption 1 hold and choose the learning rates as those specified in Proposition 2. Then, the sequences generated by PGDAm satisfies the following global convergence properties.*

1. *The function value sequence $\{(\Phi+g)(x_t)\}_t$ converges to a finite limit $H^* > -\infty$;*
2. *The sequences $\{x_t\}_t, \{y_t\}_t$ are bounded and have compact sets of limit points. Moreover, for any limit point x^* of $\{x_t\}_t$ we have $(\Phi+g)(x^*) \equiv H^*$;*
3. *Every limit point of $\{x_t\}_t$ is a critical point of $(\Phi+g)(x)$.*

Proof. We first prove some useful results on the Lyapunov function $H(z_t)$. By Assumption 1 we know that $\Phi+g$ is bounded below and have compact sub-level sets. Also, note that the Lyapunov function $H(z_t)$ is the objective function $\Phi+g$ plus some quadratic regularization terms, we therefore conclude that $H(z_t)$ is also bounded below and have compact sub-level set.

We first show that $\{(\Phi+g)(x_t)\}_t$ has a finite limit. We have shown in Proposition 2 that $\{H(z_t)\}_t$ is monotonically decreasing. Since $H(z)$ is bounded below, we conclude that $\{H(z_t)\}_t$ has a finite limit $H^* > -\infty$, i.e., $\lim_{t\to\infty}(\Phi+g)(x_t) + \Big(1-\frac{1}{4\kappa}\Big)\|y_t - y^*(x_t)\|^2 + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\|^2 = H^*$. Moreover, since $\|x_t - x_{t-1}\| \xrightarrow{t} 0, \|y_t - y^*(x_t)\| \xrightarrow{t} 0$, we further conclude that $\lim_{t\to\infty}(\Phi+g)(x_t) = H^*$.

Next, we prove the second item. Since $\{H(z_t)\}_t$ is monotonically decreasing and $H(z)$ has compact sub-level set, we conclude that $\{x_t\}_t, \{y_t\}_t$ are bounded and hence have compact sets of limit points. Next, we derive a bound on the subdifferential. By the optimality condition of the proximal gradient update of x_{t+1} we have

$$\begin{aligned}\mathbf{0} &\in \partial g(x_{t+1}) + \frac{1}{\eta_x}\big(x_{t+1} - \widetilde{x}_t + \eta_x \nabla_1 f(x_t, y_t)\big)\\ &= \partial g(x_{t+1}) + \frac{1}{\eta_x}\big(x_{t+1} - x_t - \beta(x_t - x_{t-1}) + \eta_x \nabla_1 f(x_t, y_t)\big).\end{aligned}$$

Then, we obtain that

$$\frac{1}{\eta_x}\big(x_t - x_{t+1}\big) - \nabla_1 f(x_t, y_t) + \frac{\beta}{\eta_x}(x_t - x_{t-1}) + \nabla\Phi(x_{t+1}) \in \partial(\Phi+g)(x_{t+1}), \tag{13}$$

which further implies that

$$\begin{aligned}\mathrm{dist}_{\partial(\Phi+g)(x_{t+1})}(\mathbf{0}) &\le \frac{1}{\eta_x}\|x_{t+1} - x_t\| + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\| + \|\nabla_1 f(x_t, y_t) - \nabla\Phi(x_{t+1})\|\\ &\le \frac{1}{\eta_x}\|x_{t+1} - x_t\| + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\| + \|\nabla_1 f(x_t, y_t) - \nabla_1 f(x_{t+1}, y^*(x_{t+1}))\|\\ &\le \frac{1}{\eta_x}\|x_{t+1} - x_t\| + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\| + L(\|x_{t+1} - x_t\| + \|y^*(x_{t+1}) - y_t\|)\\ &\le \Big(\frac{1}{\eta_x} + L\Big)\|x_{t+1} - x_t\| + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\|\\ &\quad + L\big(\|y^*(x_{t+1}) - y^*(x_t)\| + \|y^*(x_t) - y_t\|\big)\\ &\le \Big(\frac{1}{\eta_x} + L(1+\kappa)\Big)\|x_{t+1} - x_t\| + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\| + L\|y^*(x_t) - y_t\|.\end{aligned}$$

Since we have shown that $\|x_{t+1} - x_t\| \xrightarrow{t} 0, \|y^*(x_t) - y_t\| \xrightarrow{t} 0$, we conclude from the above inequality that $\mathrm{dist}_{\partial(\Phi+g)(x_t)}(\mathbf{0}) \xrightarrow{t} 0$. Therefore, we have shown that

$$\partial(\Phi+g)(x_t) \ni \frac{1}{\eta_x}\big(x_t - x_{t+1}\big) - \nabla_1 f(x_t, y_t) + \frac{\beta}{\eta_x}(x_t - x_{t-1}) + \nabla\Phi(x_{t+1}) \xrightarrow{t} \mathbf{0}. \tag{14}$$

Now consider any limit point x^* of x_t so that $x_{t(j)} \xrightarrow{j} x^*$ along a subsequence. By the proximal update of $x_{t(j)}$, we have

$$g(x_{t(j)}) + \frac{1}{2\eta_x}\|x_{t(j)} - \widetilde{x}_{t(j)-1}\|^2 + \langle x_{t(j)} - \widetilde{x}_{t(j)-1}, \nabla_1 f(x_{t(j)-1}, y_{t(j)-1})\rangle$$

$$\leq g(x^*) + \frac{1}{2\eta_x}\|x^* - \widetilde{x}_{t(j)-1}\|^2 + \langle x^* - \widetilde{x}_{t(j)-1}, \nabla_1 f(x_{t(j)-1}, y_{t(j)-1})\rangle.$$

Taking limsup on both sides of the above inequality and noting that $\{x_t\}_t, \{y_t\}_t$ are bounded, ∇f is Lipschitz, $\|x_t - \widetilde{x}_{t-1}\| \leq \|x_t - x_{t-1}\| + \beta\|x_{t-1} - x_{t-2}\| \overset{t}{\to} 0$ and $x_{t(j)} \to x^*$, we conclude that $\limsup_j g(x_{t(j)}) \leq g(x^*)$. Since g is lower-semicontinuous, we know that $\liminf_j g(x_{t(j)}) \geq g(x^*)$. Combining these two inequalities yields that $\lim_j g(x_{t(j)}) = g(x^*)$. By continuity of Φ, we further conclude that $\lim_j (\Phi + g)(x_{t(j)}) = (\Phi + g)(x^*)$. Since we have shown that the entire sequence $\{(\Phi + g)(x_t)\}_t$ converges to a certain finite limit H^*, we conclude that $(\Phi + g)(x^*) \equiv H^*$ for all the limit points x^* of $\{x_t\}_t$.

Next, we prove the third item. To this end, we have shown that for every subsequence $x_{t(j)} \overset{j}{\to} x^*$, we have that $(\Phi + g)(x_{t(j)}) \overset{j}{\to} (\Phi + g)(x^*)$ and there exists $u_t \in \partial(\Phi + g)(x_t) \overset{t}{\to} \mathbf{0}$ (by eq. (14)). Recall the definition of limiting sub-differential, we conclude that every limit point x^* of $\{x_t\}_t$ is a critical point of $(\Phi + g)(x)$, i.e., $\mathbf{0} \in \partial(\Phi + g)(x^*)$.

□

D. Proof of Theorem 2

Theorem 2 (Variable convergence)**.** *Let Assumptions 1 and 2 hold and assume that $H(z)$ has the KŁ geometry. Choose the learning rates as specified in Proposition 2. Then, the sequences $\{(x_t, y_t)\}_t$ generated by PGDAm converge to a certain critical point $(x^*, y^*(x^*))$ of $(\Phi + g)(x)$, i.e.,*

$$x_t \overset{t}{\to} x^*, \quad y_t \overset{t}{\to} y^*(x^*).$$

Proof. We first derive a bound on $\partial H(z)$. Recall that $H(z) = \Phi(x) + g(x) + \Big(1 - \frac{1}{4\kappa}\Big)\|y - y^*(x)\|^2 + \frac{\beta}{\eta_x}\|x - x'\|^2$ and $\|y^*(x) - y\|^2$ is sub-differentiable. We therefore have

$$\begin{aligned}
\partial_x H(z) &\supset \partial(\Phi + g)(x) + \Big(1 - \frac{1}{4\kappa}\Big)\partial_x(\|y^*(x) - y\|^2) + \frac{2\beta}{\eta_x}(x - x'),\\
\nabla_{x'} H(z) &= \frac{2\beta}{\eta_x}(x' - x),\\
\nabla_y H(z) &= \Big(2 - \frac{1}{2\kappa}\Big)\big(y - y^*(x)\big),
\end{aligned}$$

where the first inclusion follows from the scalar multiplication rule and sum rule of sub-differential (Kruger, 2003). Moreover, in the proof of Theorem 2 in (Chen et al., 2021), it has been proved that $\mathrm{dist}_{\partial_x(\|y^*(x)-y\|^2)}(\mathbf{0}) \leq 2\kappa\|y^*(x) - y\|$. Therefore, utilizing the characterization of $\partial(\Phi + g)(x_{t+1})$ in eq. (13), we obtain that

$$\begin{aligned}
&\mathrm{dist}_{\partial H(z_{t+1})}(\mathbf{0})\\
&\leq \mathrm{dist}_{\partial_x H(z_{t+1})}(\mathbf{0}) + \|\nabla_{x'} H(z_{t+1})\| + \|\nabla_y H(z_{t+1})\|\\
&\leq \frac{1}{\eta_x}\|x_{t+1} - x_t\| + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\| + \|\nabla_1 f(x_t, y_t) - \nabla\Phi(x_{t+1})\|\\
&\quad + \Big(2 - \frac{1}{2\kappa}\Big)\kappa\|y^*(x_{t+1}) - y_{t+1}\| + \frac{2\beta}{\eta_x}\|x_{t+1} - x_t\| + \frac{2\beta}{\eta_x}\|x_{t+1} - x_t\|\\
&\quad + \Big(2 - \frac{1}{2\kappa}\Big)\|y_{t+1} - y^*(x_{t+1})\|\\
&\overset{(i)}{\leq} \Big(\frac{1}{\eta_x} + L + \frac{4\beta}{\eta_x}\Big)\|x_{t+1} - x_t\| + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\| + L\|y^*(x_{t+1}) - y_t\|\\
&\quad + \big(2 - \frac{1}{2\kappa}\big)\big(1 + \kappa\big)\|y^*(x_{t+1}) - y_{t+1}\|\\
&\overset{(ii)}{\leq} \Big(\frac{1}{\eta_x} + L + \frac{4\beta}{\eta_x}\Big)\|x_{t+1} - x_t\| + \frac{\beta}{\eta_x}\|x_t - x_{t-1}\| + L(\|y^*(x_t) - y_t\| + \kappa\|x_{t+1} - x_t\|)\\
&\quad + 2(1 + \kappa)\big[\sqrt{1 - \kappa^{-1}}\|y^*(x_t) - y_t\| + \kappa\sqrt{1 + \kappa^{\frac{1}{2}}}\|x_{t+1} - x_t\|\big]
\end{aligned}$$

$$\overset{(iii)}{\leq} \Big(\frac{1}{\eta_x}+(L+4\kappa^2)(1+\kappa)+\frac{4\beta}{\eta_x}\Big)\|x_{t+1}-x_t\|+\frac{\beta}{\eta_x}\|x_t-x_{t-1}\|+(L+4\kappa)\|y^*(x_t)-y_t\|, \tag{15}$$

where (i) uses $\nabla\Phi(x_{t+1})=\nabla_1 f(x_{t+1},y^*(x_{t+1}))$, (ii) uses eq. (10) and the inequality that $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ $(a,b\geq 0)$ and (iii) uses $\kappa\geq 1$.

Next, we prove the convergence of the sequence under the assumption that $H(z)$ is a KŁ function. Recall that we have shown in the proof of Theorem 1 that: 1) $\{H(z_t)\}_t$ decreases monotonically to the finite limit H^*; 2) for any limit point x^*, y^* of $\{x_t\}_t, \{y_t\}_t$, $H(x^*,y^*)$ has the constant value H^*. Hence, the KŁ inequality (see Definition 2) holds after sufficiently large number of iterations, i.e., there exists $t_0\in\mathbb{N}^+$ such that for all $t\geq t_0$,

$$\varphi'(H(z_t)-H^*)\mathrm{dist}_{\partial H(z_t)}(\mathbf{0})\geq 1.$$

Rearranging the above inequality and utilizing eq. (15), we obtain that for all $t\geq t_0$,

$$\begin{aligned}
&\varphi'(H(z_t)-H^*)\\
&\geq \frac{1}{\mathrm{dist}_{\partial H(z_t)}(\mathbf{0})}\\
&\geq \Big[\Big(\frac{1}{\eta_x}+(L+4\kappa^2)(1+\kappa)+\frac{4\beta}{\eta_x}\Big)\|x_t-x_{t-1}\|+\frac{\beta}{\eta_x}\|x_{t-1}-x_{t-2}\|+(L+4\kappa)\|y^*(x_{t-1})-y_{t-1}\|\Big]^{-1}
\end{aligned} \tag{16}$$

By concavity of the function φ (see Definition 2), we know that

$$\begin{aligned}
&\varphi(H(z_t)-H^*)-\varphi(H(z_{t+1})-H^*)\\
&\geq \varphi'(H(z_t)-H^*)(H(z_t)-H(z_{t+1}))\\
&\overset{(i)}{\geq} \frac{L\kappa\|x_{t+1}-x_t\|^2+\frac{\beta}{2\eta_x}\|x_t-x_{t-1}\|^2+\frac{1}{4\kappa}\|y_t-y^*(x_t)\|^2}{\Big(\frac{1}{\eta_x}+(L+4\kappa^2)(1+\kappa)+\frac{4\beta}{\eta_x}\Big)\|x_t-x_{t-1}\|+\frac{\beta}{\eta_x}\|x_{t-1}-x_{t-2}\|+(L+4\kappa)\|y^*(x_{t-1})-y_{t-1}\|}\\
&\overset{(ii)}{\geq} \frac{\frac{1}{3}\Big[\sqrt{\kappa L}\|x_{t+1}-x_t\|+\sqrt{\frac{\beta}{2\eta_x}}\|x_t-x_{t-1}\|+\frac{1}{2\sqrt{\kappa}}\|y_t-y^*(x_t)\|\Big]^2}{\Big(\frac{1}{\eta_x}+(L+4\kappa^2)(1+\kappa)+\frac{4\beta}{\eta_x}\Big)\|x_t-x_{t-1}\|+\frac{\beta}{\eta_x}\|x_{t-1}-x_{t-2}\|+(L+4\kappa)\|y^*(x_{t-1})-y_{t-1}\|},
\end{aligned} \tag{17}$$

where (i) uses Proposition 2 and eq. (16), (ii) uses the inequality that $a^2+b^2+c^2\geq\frac{1}{3}(a+b+c)^2$.

Rearranging the above inequality yields that

$$\begin{aligned}
&\Big[\sqrt{\kappa L}\|x_{t+1}-x_t\|+\sqrt{\frac{\beta}{2\eta_x}}\|x_t-x_{t-1}\|+\frac{1}{2\sqrt{\kappa}}\|y_t-y^*(x_t)\|\Big]^2\\
&\leq 3[\varphi(H(z_t)-H^*)-\varphi(H(z_{t+1})-H^*)]\\
&\quad\Big[\Big(\frac{1}{\eta_x}+(L+4\kappa^2)(1+\kappa)+\frac{4\beta}{\eta_x}\Big)\|x_t-x_{t-1}\|+\frac{\beta}{\eta_x}\|x_{t-1}-x_{t-2}\|+(L+4\kappa)\|y^*(x_{t-1})-y_{t-1}\|\Big]\\
&\leq \frac{3}{2}\Big[C[\varphi(H(z_t)-H^*)-\varphi(H(z_{t+1})-H^*)]\\
&\quad+\frac{1}{C}\Big(\frac{1}{\eta_x}+(L+4\kappa^2)(1+\kappa)+\frac{4\beta}{\eta_x}\Big)\|x_t-x_{t-1}\|+\frac{\beta}{C\eta_x}\|x_{t-1}-x_{t-2}\|+\frac{1}{C}(L+4\kappa)\|y^*(x_{t-1})-y_{t-1}\|\Big]^2
\end{aligned}$$

where the final step uses the inequality that $3ab\leq\frac{3}{2}(Ca+\frac{b}{C})^2$ for any $a,b\geq 0$ and $C>0$ (the value of C will be assigned later). Taking square root of both sides of the above inequality, denoting $A:=\min\{\sqrt{\kappa L},\sqrt{\frac{\beta}{2\eta_x}},\frac{1}{2\sqrt{\kappa}}\}$ and telescoping over $t=t_0,\ldots,T-1$, we obtain that

$$\frac{2A}{3}\Big(\sum_{t=t_0}^{T-1}\|x_{t+1}-x_t\|+\sum_{t=t_0}^{T-1}\|x_t-x_{t-1}\|+\sum_{t=t_0}^{T-1}\|y_t-y^*(x_t)\|\Big)$$

$$\leq C\varphi(H(z_{t_0})-H^*)-C\varphi(H(z_T)-H^*)+\frac{1}{C}\Big(\frac{1}{\eta_x}+(L+4\kappa^2)(1+\kappa)+\frac{4\beta}{\eta_x}\Big)\sum_{t=t_0}^{T-1}\|x_t-x_{t-1}\|$$
$$+\frac{\beta}{C\eta_x}\sum_{t=t_0}^{T-1}\|x_{t-1}-x_{t-2}\|+\frac{1}{C}(L+4\kappa)\sum_{t=t_0}^{T-1}\|y^*(x_{t-1})-y_{t-1}\|$$
$$\leq \frac{Cc}{\theta}[H(z_{t_0})-H^*]^\theta+\frac{1}{C}\Big(\frac{1}{\eta_x}+(L+4\kappa^2)(1+\kappa)+\frac{4\beta}{\eta_x}\Big)\sum_{t=t_0-1}^{T-2}\|x_{t+1}-x_t\|$$
$$+\frac{\beta}{C\eta_x}\sum_{t=t_0-1}^{T-2}\|x_t-x_{t-1}\|+\frac{1}{C}(L+4\kappa)\sum_{t=t_0-1}^{T-2}\|y^*(x_t)-y_t\|$$

where the final steps uses $\varphi(s)=\frac{c}{\theta}s^\theta$ and the fact that $H(z_T)-H^*\geq 0$. Since the value of $C>0$ is arbitrary, we can select large enough C such that $\frac{1}{C}\Big(\frac{1}{\eta_x}+(L+4\kappa)^2(1+\kappa)+\frac{4\beta}{\eta_x}\Big)<\frac{A}{3}$ and $\frac{\beta}{C\eta_x}\leq\frac{2A}{3}$, $\frac{1}{C}(L+4\kappa)<\frac{2A}{3}$. Hence, the inequality above further implies that

$$\frac{A}{3}\sum_{t=t_0}^{T-1}\|x_{t+1}-x_t\|\leq\frac{Cc}{\theta}[H(z_{t_0})-H^*]^\theta+\frac{A}{3}\|x_{t_0}-x_{t_0-1}\|+\frac{2A}{3}(\|x_{t_0-1}-x_{t_0-2}\|)$$
$$+\frac{2A}{3}\|y^*(x_{t_0-1})-y_{t_0-1}\|<+\infty.$$

Letting $T\to\infty$, we conclude that

$$\sum_{t=1}^{\infty}\|x_{t+1}-x_t\|<+\infty.$$

Moreover, this implies that $\{x_t\}_t$ is a Cauchy sequence and therefore converges to a certain limit, i.e., $x_t\xrightarrow{t}x^*$. We have shown in Theorem 1 that any such limit point must be a critical point of $\Phi+g$. Hence, we conclude that $\{x_t\}_t$ converges to a certain critical point x^* of $(\Phi+g)(x)$. Also, note that $\|y^*(x_t)-y_t\|\xrightarrow{t}0$, $x_t\xrightarrow{t}x^*$ and y^* is a Lipschitz mapping, so we conclude that $\{y_t\}_t$ converges to $y^*(x^*)$.

□

E. Proof of Theorem 3

Theorem 3 (Function value convergence rate)**.** *Under the same conditions as those of Theorem 2, the Lyapunov function value sequence $\{H(z_t)\}_t$ converges to the limit H^* at the following rates.*

1. *If KŁ geometry holds with $\theta=1$, then $H(z_t)\downarrow H^*$ within finite number of iterations;*
2. *If $\theta\in(\frac{1}{2},1)$, then $H(z_t)\downarrow H^*$ super-linearly: $\forall t\geq t_0$*

$$H(z_t)-H^*\leq(3Mc^2)^{-\frac{1}{2\theta-1}}\exp\Big(-\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_0}\Big);$$

3. *If $\theta=\frac{1}{2}$, then $H(z_t)\downarrow H^*$ linearly: $\forall t\geq t_0$*

$$H(z_t)-H^*\leq\Big(1+\frac{1}{3Mc^2}\Big)^{t_0-t}(H(z_{t_0})-H^*);$$

4. *If $\theta\in(0,\frac{1}{2})$, then $H(z_t)\downarrow H^*$ sub-linearly: $\forall t\geq t_0$*

$$H(z_t)-H^*\leq\mathcal{O}\Big((t-t_0)^{-\frac{1}{1-2\theta}}\Big).$$

Proof. Note that eq. (15) implies that

$$\mathrm{dist}_{\partial H(z_{t+1})}(\mathbf{0})^2\leq 3\Big(\frac{1}{\eta_x}+(L+4\kappa^2)(1+\kappa)+\frac{4\beta}{\eta_x}\Big)^2\|x_{t+1}-x_t\|^2$$

$$+\frac{3\beta^2}{\eta_x^2}\|x_t-x_{t-1}\|^2+3(L+4\kappa)^2\|y^*(x_t)-y_t\|^2. \tag{18}$$

Recall that we have shown that for all $t\geq t_0$, the KŁ property holds and we have

$$\big[\varphi'(H(z_t)-H^*)\big]^2\mathrm{dist}^2_{\partial H(z_t)}(\mathbf{0})\geq 1.$$

Throughout the rest of the proof, we assume $t\geq t_0$. Substituting eq. (18) into the above bound yields that

$$\begin{aligned}1\leq&3\big[\varphi'(H(z_t)-H^*)\big]^2\Big[\Big(\frac{1}{\eta_x}+(L+4\kappa^2)(1+\kappa)+\frac{4\beta}{\eta_x}\Big)^2\|x_t-x_{t-1}\|^2\\&+\frac{\beta^2}{\eta_x^2}\|x_{t-1}-x_{t-2}\|^2+(L+4\kappa)^2\|y^*(x_{t-1})-y_{t-1}\|^2\Big].\\\leq&3M\big[\varphi'(H(z_t)-H^*)\big]^2\Big[L\kappa\|x_t-x_{t-1}\|^2+\frac{\beta}{2\eta_x}\|x_{t-1}-x_{t-2}\|^2+\frac{1}{4\kappa}\|y^*(x_{t-1})-y_{t-1}\|^2\Big]\end{aligned} \tag{19}$$

where the second inequality uses the definition of M in Section 5.3.

Substituting eq. (2) and $\varphi'(s)=cs^{\theta-1}$ ($c>0$) into eq. (19) and rearranging, we further obtain that

$$\big[c(H(z_t)-H^*)^{\theta-1}\big]^{-2}\leq 3M[H(z_{t-1})-H(z_t)]$$

Defining $d_t=H(z_t)-H^*$, the above inequality further becomes

$$d_{t-1}-d_t\geq\frac{1}{3Mc^2}d_t^{2(1-\theta)}. \tag{20}$$

Next, we prove the convergence rates case by case.

(Case 1) If $\theta=1$, then eq. (20) implies that $d_{t-1}-d_t\geq\frac{1}{3Mc^2}>0$ whenever $d_t>0$. Hence, d_t achieves 0 (i.e., $H(z_t)$ achieves H^*) within finite number of iterations.

(Case 2) If $\theta\in(\frac{1}{2},1)$, since $d_t\geq 0$, eq. (20) implies that

$$d_{t-1}\geq\frac{1}{3Mc^2}d_t^{2(1-\theta)}, \tag{21}$$

which is equivalent to that

$$(3Mc^2)^{\frac{1}{2\theta-1}}d_t\leq\Big[(3Mc^2)^{\frac{1}{2\theta-1}}d_{t-1}\Big]^{\frac{1}{2(1-\theta)}} \tag{22}$$

Since $d_t\downarrow 0$, $(3Mc^2)^{\frac{1}{2\theta-1}}d_{t_1}\leq e^{-1}$ for sufficiently large $t_1\in\mathbb{N}^+$ and $t_1\geq t_0$. Hence, eq. (22) implies that for $t\geq t_1$

$$\begin{aligned}(3Mc^2)^{\frac{1}{2\theta-1}}d_t\leq&\Big[(3Mc^2)^{\frac{1}{2\theta-1}}d_{t_1}\Big]^{\left[\frac{1}{2(1-\theta)}\right]^{t-t_1}}\\\leq&\exp\Big\{-\Big[\frac{1}{2(1-\theta)}\Big]^{t-t_1}\Big\}.\end{aligned}$$

Note that $\theta\in(\frac{1}{2},1)$ implies that $\frac{1}{2(1-\theta)}>1$, and thus the inequality above implies that $H(z_t)\downarrow H^*$ at a super-linear rate.

(Case 3) If $\theta=\frac{1}{2}$,

$$d_{t-1}-d_t\geq\frac{1}{3Mc^2}d_t, \tag{23}$$

which implies that $d_t\leq\Big(1+\frac{1}{3Mc^2}\Big)^{-1}d_{t-1}$. Therefore, $d_t\downarrow 0$ (i.e., $H(z_t)\downarrow H^*$) at a linear rate.

(Case 4) If $\theta \in (0, \frac{1}{2})$, consider the following two subcases.

If $d_{t-1} \leq 2d_t$, denote $\psi(s) = \frac{1}{1-2\theta} s^{-(1-2\theta)}$, then

$$\begin{aligned}\psi(d_t) - \psi(d_{t-1}) = \int_{d_t}^{d_{t-1}} -\psi'(s)ds = \int_{d_t}^{d_{t-1}} s^{-2(1-\theta)}ds \overset{(i)}{\geq} d_{t-1}^{-2(1-\theta)}(d_{t-1} - d_t) \\ \overset{(ii)}{\geq} \frac{1}{3Mc^2}\Big(\frac{d_t}{d_{t-1}}\Big)^{2(1-\theta)} \geq \frac{1}{3^{3-2\theta}Mc^2} \geq \frac{1}{27Mc^2}\end{aligned} \tag{24}$$

where (i) uses $d_t \leq d_{t-1}$ and $-2(1-\theta) < -1$, and (ii) uses eq. (20).

If $d_{t-1} > 2d_t$

$$\begin{aligned}\psi(d_t) - \psi(t_{t-1}) =& \frac{1}{1-2\theta}(d_t^{-(1-2\theta)} - d_{t-1}^{-(1-2\theta)}) \geq \frac{1}{1-2\theta}(d_t^{-(1-2\theta)} - (2d_t)^{-(1-2\theta)}) \\ \geq& \frac{1-2^{-(1-2\theta)}}{1-2\theta} d_t^{-(1-2\theta)} \geq \frac{1-2^{-(1-2\theta)}}{1-2\theta} d_{t_0}^{-(1-2\theta)}.\end{aligned} \tag{25}$$

where we use $-(1-2\theta) < 0$, $d_{t-1} > 2d_t$ and $d_t \leq d_{t_0}$.

Hence,

$$\psi(d_t) - \psi(d_{t-1}) \geq \min\Big[\frac{1}{27Mc^2}, \frac{1-2^{-(1-2\theta)}}{1-2\theta} d_{t_0}^{-(1-2\theta)}\Big] \overset{\text{def}}{=} \frac{C}{1-2\theta} > 0, \tag{26}$$

which implies that

$$\psi(d_t) \geq \psi(d_{t_0}) + \frac{C}{1-2\theta}(t - t_0) \geq \frac{C}{1-2\theta}(t - t_0)$$

By substituing the definition of ψ, the inequality above implies that $H(z_t) \downarrow H^*$ in a sub-linear rate. □

F. Proof of Theorem 4

Theorem 4 (Variable convergence rate)**.** *Under the same conditions as those of Theorem 2, the sequences $\{x_t, y_t\}_t$ converge to their limits $x^*, y^*(x^*)$ respectively at the following rates, where we denote $d_t := \max\{\|x_t - x^*\|, \|y_t - y^*(x^*)\|\}$.*

1. *If KŁ geometry holds with $\theta = 1$, then $(x_t, y_t) \to (x^*, y^*(x^*))$ within finite number of iterations;*
2. *If $\theta \in (\frac{1}{2}, 1)$, then $(x_t, y_t) \to (x^*, y^*(x^*))$ super-linearly:*

$$d_t \leq \mathcal{O}\Big(\exp\Big(-\big(\frac{1}{2(1-\theta)}\big)^{t-t_0}\Big)\Big), \forall t \geq t_0;$$

3. *If $\theta = \frac{1}{2}$, then $(x_t, y_t) \to (x^*, y^*(x^*))$ linearly:*

$$d_t \leq \mathcal{O}\Big(\Big(\min\Big\{2, 1 + \frac{1}{2Mc^2}\Big\}\Big)^{(t_0-t)/2}\Big), \forall t \geq t_0;$$

4. *If $\theta \in (0, \frac{1}{2})$, then $(x_t, y_t) \to (x^*, y^*(x^*))$ sub-linearly:*

$$d_t \leq \mathcal{O}\Big((t - t_0)^{-\frac{\theta}{1-2\theta}}\Big), \forall t \geq t_0.$$

Proof. (Case 1) If $\theta = 1$, then based on the first case of Appendix E, $H(z_t) \equiv H^*$ after finite number of iterations. Hence, for large enough t, Proposition 2 yields that

$$\begin{aligned}&L\kappa\|x_{t+1} - x_t\|^2 + \frac{\beta}{2\eta_x}\|x_t - x_{t-1}\|^2 + \frac{1}{4\kappa}\big(\|y_{t+1} - y^*(x_{t+1})\|^2 + \|y_t - y^*(x_t)\|^2\big) \\ &\leq H(z_t) - H(z_{t+1}) = 0,\end{aligned} \tag{27}$$

which implies that $x_{t+1}=x_t$ and $y_t=y^*(x_t)$ for large enougth t. Hence, $x_t\to x^*$ and $y_t\to y^*(x^*)$ within finite number of iterations.

(Case 2) If $\theta\in(\frac{1}{2},1)$, denote $A_t=\sqrt{\kappa L}\|x_{t+1}-x_t\|+\sqrt{\frac{\beta}{2\eta_x}}\|x_t-x_{t-1}\|+\frac{1}{2\sqrt{\kappa}}\|y_t-y^*(x_t)\|$. Then, based on the definition of M in Section 5.3, we have

$$\Big(\frac{1}{\eta_x}+(L+4\kappa^2)(1+\kappa)+\frac{4\beta}{\eta_x}\Big)\|x_t-x_{t-1}\|+\frac{\beta}{\eta_x}\|x_{t-1}-x_{t-2}\|+(L+4\kappa)\|y^*(x_{t-1})-y_{t-1}\|\le\sqrt{M}A_{t-1}. \quad (28)$$

Hence, eqs. (16) & (28) and $\varphi'(s)=cs^{\theta-1}$ imply that

$$c(H(z_t)-H^*)^{\theta-1}\ge(\sqrt{M}A_{t-1})^{-1},$$

which along with $\theta-1<0$ implies

$$H(z_t)-H^*\le(c\sqrt{M}A_{t-1})^{\frac{1}{1-\theta}}. \quad (29)$$

Then, eqs. (17) & (28) imply that

$$\varphi(H(z_t)-H^*)-\varphi(H(z_{t+1})-H^*)\ge\frac{A_t^2}{3\sqrt{M}A_{t-1}}.$$

Recalling the definition of A_t and $\varphi(s)=\frac{c}{\theta}s^\theta$, the above inequality further implies that

$$\frac{c}{\theta}(H(z_t)-H^*)^\theta-\frac{c}{\theta}(H(z_{t+1})-H^*)^\theta\ge\frac{A_t^2}{3\sqrt{M}A_{t-1}}. \quad (30)$$

Substituting eq. (29) into eq. (30) and using $H(z_{t+1})-H^*\ge0$ yield that

$$A_t^2\le\frac{3}{\theta}(c\sqrt{M}A_{t-1})^{\frac{1}{1-\theta}},$$

which is equivalent to that

$$C_1A_t\le(C_1A_{t-1})^{\frac{1}{2(1-\theta)}}, \quad (31)$$

where

$$C_1=(3/\theta)^{\frac{1-\theta}{2\theta-1}}(c\sqrt{M})^{\frac{1}{2\theta-1}}.$$

Note that eq. (31) holds for $t\ge t_0$. Since $A_t\to0$, there exists $t_1\ge t_0$ such that $C_1A_{t_1}\le e^{-1}$. Hence, by iterating eq. (31) from $t=t_1+1$, we obtain

$$C_1A_t\le\exp\Big[-\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_1}\Big],\quad\forall t\ge t_1+1.$$

Hence, for any $t\ge t_1+1$,

$$\begin{aligned}\sum_{s=t}^{\infty}A_s\le&\frac{1}{C_1}\sum_{s=t}^{\infty}\exp\Big[-\Big(\frac{1}{2(1-\theta)}\Big)^{s-t_1}\Big]\\=&\frac{1}{C_1}\exp\Big[-\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_1}\Big]\sum_{s=t}^{\infty}\exp\Big[\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_1}-\Big(\frac{1}{2(1-\theta)}\Big)^{s-t_1}\Big]\\=&\frac{1}{C_1}\exp\Big[-\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_1}\Big]\sum_{s=t}^{\infty}\exp\Big\{\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_1}\Big[1-\Big(\frac{1}{2(1-\theta)}\Big)^{s-t}\Big]\Big\}\end{aligned}$$

$$\overset{(i)}{\leq}\frac{1}{C_1}\exp\Big[-\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_1}\Big]\sum_{s=t}^{\infty}\exp\Big[1-\Big(\frac{1}{2(1-\theta)}\Big)^{s-t}\Big]$$
$$=\frac{1}{C_1}\exp\Big[-\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_1}\Big]\sum_{s=0}^{\infty}\exp\Big[1-\Big(\frac{1}{2(1-\theta)}\Big)^{s}\Big]$$
$$\overset{(ii)}{\leq}\mathcal{O}\Big\{\exp\Big[-\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_1}\Big]\Big\}, \tag{32}$$

where (i) uses the inequalities that $\frac{1}{2(1-\theta)} > 1$ and that $s \geq t \geq t_1 + 1$, and (ii) uses the fact that $\sum_{s=0}^{\infty}\exp\Big[1-\Big(\frac{1}{2(1-\theta)}\Big)^{s}\Big] < +\infty$ is a positive constant independent from t. Therefore,

$$\|x_t - x^*\| = \limsup_{T\to\infty}\|x_t - x_T\| \leq \limsup_{T\to\infty}\sum_{s=t}^{T-1}\|x_{s+1}-x_s\|$$
$$\leq\frac{1}{\sqrt{\kappa L}}\limsup_{T\to\infty}\sum_{s=t}^{T-1}A_s \leq \mathcal{O}\Big\{\exp\Big[-\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_1}\Big]\Big\}, \tag{33}$$

and

$$\|y_t - y^*(x^*)\| \leq \|y_t - y^*(x_t)\| + \|y^*(x_t) - y^*(x^*)\| \overset{(i)}{\leq} 2\sqrt{\kappa}A_t + \kappa\|x_t - x^*\|$$
$$\leq 2\sqrt{\kappa}\sum_{s=t}^{\infty}A_s + \kappa\|x_t - x^*\| \overset{(ii)}{\leq} \mathcal{O}\Big\{\exp\Big[-\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_1}\Big]\Big\},$$

where (i) uses the Lipschitz property of y^* in Proposition 1, and (ii) uses eqs. (32) & (33). Hence, $d_t := \max\{\|x_t - x^*\|, \|y_t - y^*(x^*)\|\} \leq \mathcal{O}\Big\{\exp\Big[-\Big(\frac{1}{2(1-\theta)}\Big)^{t-t_1}\Big]\Big\}$.

(Case 3 & 4) Notice that eq. (30) still holds if $\theta \in \big(0, \frac{1}{2}\big]$. Hence, if $A_t \geq \frac{1}{2}A_{t-1}$, then eq. (30) implies that

$$A_t \leq \frac{6c\sqrt{M}}{\theta}\big[(H(z_t) - H^*)^\theta - (H(z_{t+1}) - H^*)^\theta\big].$$

Otherwise, $A_t \leq \frac{1}{2}A_{t-1}$. Combining these two inequalities yields that

$$A_t \leq \frac{6c\sqrt{M}}{\theta}\big[(H(z_t) - H^*)^\theta - (H(z_{t+1}) - H^*)^\theta\big] + \frac{1}{2}A_{t-1}.$$

Notice that the inequality above holds whenever $t \geq t_0$. Hence, telescoping the inequality above yields

$$\sum_{s=t}^{T}A_s \leq \frac{6c\sqrt{M}}{\theta}\big[(H(z_t) - H^*)^\theta - (H(z_{T+1}) - H^*)^\theta\big] + \frac{1}{2}\sum_{s=t-1}^{T-1}A_s, \quad \forall t \geq t_0, \tag{34}$$

which along with $A_T \geq 0$, $H(z_{T+1}) - H^* \geq 0$ implies that

$$\frac{1}{2}\sum_{s=t}^{T}A_s \leq \frac{6c\sqrt{M}}{\theta}(H(z_t) - H^*)^\theta + \frac{1}{2}A_{t-1},$$

Letting $t = t_0$ and $T \to \infty$ in the above inequality yields that $\sum_{s=t_0}^{\infty}A_s < +\infty$. Hence, by letting $T \to \infty$ and denoting $S_t = \sum_{s=t}^{\infty}A_s$ in eq. (34), we obtain that

$$S_t \leq \frac{6c\sqrt{M}}{\theta}(H(z_t) - H^*)^\theta + \frac{1}{2}S_{t-1}, \quad \forall t \geq t_0,$$

which further implies that

$$S_t \leq \frac{1}{2^{t-t_0}} S_{t_0} + \frac{6c\sqrt{M}}{\theta} \sum_{s=t_0+1}^{t} \frac{1}{2^{t-s}} (H(z_s) - H^*)^\theta \tag{35}$$

(Case 3) If $\theta = 1/2$, then substituting the case 3 of Theorem 3 into eq. (35) yields that

$$\begin{aligned} S_t \leq & \frac{1}{2^{t-t_0}} S_{t_0} + 12c\sqrt{M[H(z_{t_0}) - H^*]} \sum_{s=t_0+1}^{t} \frac{1}{2^{t-s}} \Big(1 + \frac{1}{3Mc^2}\Big)^{(t_0-s)/2} \\ = & \frac{1}{2^{t-t_0}} S_{t_0} + \frac{C_2}{2^t} \sum_{s=t_0+1}^{t} \Big(\frac{1}{4} + \frac{1}{12Mc^2}\Big)^{-s/2} \end{aligned} \tag{36}$$

where

$$C_2 = 12c\sqrt{M[H(z_{t_0}) - H^*]} \Big(1 + \frac{1}{3Mc^2}\Big)^{t_0/2} \tag{37}$$

is a positive constant independent of t.

Notice that when $\frac{1}{4} + \frac{1}{8Mc^2} \geq 1$,

$$\sum_{s=t_0+1}^{t} \Big(\frac{1}{4} + \frac{1}{8Mc^2}\Big)^{-s/2} \leq t - t_0$$

and when $\frac{1}{4} + \frac{1}{8Mc^2} < 1$,

$$\sum_{s=t_0+1}^{t} \Big(\frac{1}{4} + \frac{1}{8Mc^2}\Big)^{-s/2} = \Big(\frac{1}{4} + \frac{1}{8Mc^2}\Big)^{-t/2} \frac{1 - \Big(\frac{1}{4} + \frac{1}{8Mc^2}\Big)^{(t-t_0)/2}}{1 - \Big(\frac{1}{4} + \frac{1}{8Mc^2}\Big)^{1/2}} \leq \mathcal{O}\Big[\Big(\frac{1}{4} + \frac{1}{8Mc^2}\Big)^{-t/2}\Big]$$

Since either of the two above inequalities holds, combining them yields that

$$\sum_{s=t_0+1}^{t} \Big(\frac{1}{4} + \frac{1}{8Mc^2}\Big)^{-s/2} \leq \mathcal{O}\Big\{ \max\Big[t - t_0, \Big(\frac{1}{4} + \frac{1}{8Mc^2}\Big)^{-t/2}\Big]\Big\}$$

Substituing the above inequality into eq. (36) yields that

$$\begin{aligned} S_t \leq & \frac{1}{2^{t-t_0}} S_{t_0} + \mathcal{O}\Big\{ \max\Big[2^{-t}(t - t_0), \Big(1 + \frac{1}{2Mc^2}\Big)^{-t/2}\Big]\Big\} \\ \leq & \mathcal{O}\Big\{\Big[\min\Big(2, 1 + \frac{1}{2Mc^2}\Big)\Big]^{-t/2}\Big\}. \end{aligned}$$

Hence,

$$\|x_t - x^*\| \overset{(i)}{\leq} \sum_{s=t}^{\infty} A_s = S_t \leq \mathcal{O}\Big\{\Big[\min\Big(2, 1 + \frac{1}{2Mc^2}\Big)\Big]^{-t/2}\Big\},$$

where (i) comes from eq. (33). Then,

$$\begin{aligned} \|y_t - y^*(x^*)\| \leq & \|y_t - y^*(x_t)\| + \|y^*(x_t) - y^*(x^*)\| \leq 2\sqrt{\kappa} A_t + \kappa \|x_t - x^*\| \\ \leq & 2\sqrt{\kappa} S_t + \kappa \|x_t - x^*\| \leq \mathcal{O}\Big\{\Big[\min\Big(2, 1 + \frac{1}{2Mc^2}\Big)\Big]^{-t/2}\Big\}. \end{aligned}$$

The two above inequalities imply that $d_t := \max\{\|x_t - x^*\|, \|y_t - y^*(x^*)\|\} \leq \mathcal{O}\Big\{\Big[\min\Big(2, 1+\frac{1}{2Mc^2}\Big)\Big]^{-t/2}\Big\}$.

(Case 4) If $\theta \in (0, \frac{1}{2})$, then substituting the case 4 of Theorem 3 into eq. (35) yields that for some constant $C_3 > 0$,

$$\begin{aligned}
S_t \leq& \frac{1}{2^{t-t_0}} S_{t_0} + \frac{6c\sqrt{M}}{\theta} \sum_{s=t_0+1}^{t} \frac{C_3}{2^{t-s}} (s-t_0)^{-\frac{\theta}{1-2\theta}} \\
\leq& \frac{1}{2^{t-t_0}} S_{t_0} + \frac{6c\sqrt{M}}{2^{t-t_0}\theta} \sum_{s=1}^{t-t_0} 2^s s^{-\frac{\theta}{1-2\theta}} \\
\overset{(i)}{=}& \frac{1}{2^{t-t_0}} S_{t_0} + \frac{6c\sqrt{M}}{2^{t-t_0}\theta} \sum_{s=1}^{t_2} 2^s s^{-\frac{\theta}{1-2\theta}} + \frac{6c\sqrt{M}}{2^{t-t_0}\theta} \sum_{s=t_2+1}^{t-t_0} 2^s s^{-\frac{\theta}{1-2\theta}} \\
\leq& \frac{1}{2^{t-t_0}} S_{t_0} + \frac{6c\sqrt{M}}{2^{t-t_0}\theta} \sum_{s=1}^{t_2} 2^s + \frac{6c\sqrt{M}}{2^{t-t_0}\theta} \sum_{s=t_2+1}^{t-t_0} 2^s \Big(\frac{t-t_0}{2}\Big)^{-\frac{\theta}{1-2\theta}} \\
\overset{(ii)}{\leq}& \frac{1}{2^{t-t_0}} S_{t_0} + \frac{6c\sqrt{M}}{2^{t-t_0}\theta} 2^{t_2+1} + \frac{6c\sqrt{M}}{2^{t-t_0}\theta} \Big(\frac{t-t_0}{2}\Big)^{-\frac{\theta}{1-2\theta}} 2^{t-t_0+1} \\
=& \mathcal{O}\Big[\frac{1}{2^{t-t_0}} + \frac{1}{2^{(t-t_0)/2}} + (t-t_0)^{-\frac{\theta}{1-2\theta}}\Big] = \mathcal{O}\Big[(t-t_0)^{-\frac{\theta}{1-2\theta}}\Big],
\end{aligned} \tag{38}$$

where (i) denotes $t_2 = \lfloor (t-t_0)/2 \rfloor$, (ii) uses the inequality that $\sum_{s=t_2+1}^{t-t_0} 2^s < \sum_{s=0}^{t-t_0} 2^s < 2^{t-t_0+1}$. Therefore,

$$\|x_t - x^*\| \leq S_t \leq \mathcal{O}\Big[(t-t_0)^{-\frac{\theta}{1-2\theta}}\Big],$$

and

$$\begin{aligned}
\|y_t - y^*(x^*)\| \leq& \|y_t - y^*(x_t)\| + \|y^*(x_t) - y^*(x^*)\| \leq 2\sqrt{\kappa} A_t + \kappa \|x_t - x^*\| \\
\leq& 2\sqrt{\kappa} S_t + \kappa \|x_t - x^*\| \leq \mathcal{O}\Big[(t-t_0)^{-\frac{\theta}{1-2\theta}}\Big].
\end{aligned}$$

The two above inequalities imply that $d_t := \max\{\|x_t - x^*\|, \|y_t - y^*(x^*)\|\} \leq \mathcal{O}\Big[(t-t_0)^{-\frac{\theta}{1-2\theta}}\Big]$.

□