Faster Stochastic Algorithms for Minimax Optimization under Polyak-Łojasiewicz Conditions

Lesi Chen
School of Data Science
Fudan University
lschen19@fudan.edu.cn

Boyuan Yao
School of Data Science
Fudan University
byyao19@fudan.edu.cn

Luo Luo*
School of Data Science
Fudan University
luoluo@fudan.edu.cn

Abstract

This paper considers stochastic first-order algorithms for minimax optimization under Polyak-Łojasiewicz (PL) conditions. We propose SPIDER-GDA for solving the finite-sum problem of the form $\min_x \max_y f(x,y) \triangleq \frac{1}{n}\sum_{i=1}^n f_i(x,y)$, where the objective function $f(x,y)$ is μ_x-PL in x and μ_y-PL in y; and each $f_i(x,y)$ is L-smooth. We prove SPIDER-GDA could find an ϵ-approximate solution within $\mathcal{O}\left((n+\sqrt{n}\,\kappa_x\kappa_y^2)\log(1/\epsilon)\right)$ stochastic first-order oracle (SFO) complexity, which is better than the state-of-the-art method whose SFO upper bound is $\mathcal{O}\big((n+n^{2/3}\kappa_x\kappa_y^2)\log(1/\epsilon)\big)$, where $\kappa_x \triangleq L/\mu_x$ and $\kappa_y \triangleq L/\mu_y$. For the ill-conditioned case, we provide an accelerated algorithm to reduce the computational cost further. It achieves $\tilde{\mathcal{O}}\big((n+\sqrt{n}\,\kappa_x\kappa_y)\log^2(1/\epsilon)\big)$ SFO upper bound when $\kappa_y \gtrsim \sqrt{n}$. Our ideas also can be applied to the more general setting that the objective function only satisfies PL condition for one variable. Numerical experiments validate the superiority of proposed methods.

1 Introduction

This paper focuses on smooth minimax optimization problem of the form

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} f(x,y) \triangleq \frac{1}{n}\sum_{i=1}^n f_i(x,y), \tag{1}$$

which covers a lot of important applications in machine learning such as reinforcement learning [10, 42], AUC maximization [13, 24, 48], imitation learning [5, 32], robust optimization [11], causal inference [28], game theory [6, 29] and so on.

We are interested in the minimax problems under PL conditions [9, 32, 45]. The PL condition [35] was originally proposed to relax the strong convexity in minimization problem that is sufficient for achieving the global linear convergence rate for first-order methods. In machine learning community, it has been successfully used to analyze the convergence behavior for overparameterized neural networks [23], robust phase retrieval [40] and a plenty of fundamental models [18]. There are many popular minimax formulations only satisfy PL condition, but lack strong convexity (or strong concavity). The examples include PL-game [32], robust least square [45], deep AUC maximization [24] and generative adversarial imitation learning of LQR [5, 32].

Yang et al. [45] showed that the alternating gradient descent ascent (AGDA) algorithm linearly converges to the saddle point when the objective function satisfies two-sided PL condition. They also proposed the SVRG-AGDA method for the finite-sum problem (1), which could find ϵ-approximate

*The corresponding author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: We present the comparison of SFO complexities under two-sided PL condition. Note that Yang et al. [45] named their stochastic algorithm as variance-reduced-AGDA (VR-AGDA). Here we call it SVRG-AGDA to distinguish with other variance reduced algorithms.

Algorithm	Complexity	Reference
GDA/AGDA	$\mathcal{O}\left(n\kappa_x\kappa_y^2\log\left(1/\epsilon\right)\right)$	Theorem B.1, [45]
SVRG-AGDA	$\mathcal{O}\left((n+n^{2/3}\kappa_x\kappa_y^2)\log\left(1/\epsilon\right)\right)$	[45]
SVRG-GDA	$\mathcal{O}\left((n+n^{2/3}\kappa_x\kappa_y^2)\log\left(1/\epsilon\right)\right)$	Theorem C.1
SPIDER-GDA	$\mathcal{O}\left((n+\sqrt{n}\kappa_x\kappa_y^2)\log\left(1/\epsilon\right)\right)$	Theorem 4.1
AccSPIDER-GDA	$\begin{cases}\tilde{\mathcal{O}}\left(\sqrt{n}\kappa_x\kappa_y\log^2\left(1/\epsilon\right)\right), & \sqrt{n}\lesssim\kappa_y;\\ \tilde{\mathcal{O}}\left(n\kappa_x\log^2\left(1/\epsilon\right)\right), & \kappa_y\lesssim\sqrt{n}\lesssim\kappa_x\kappa_y;\\ \mathcal{O}\left((n+\sqrt{n}\kappa_x\kappa_y^2)\log\left(1/\epsilon\right)\right), & \kappa_x\kappa_y\lesssim\sqrt{n}.\end{cases}$	Theorem 5.1

solution within $\mathcal{O}\big((n+n^{2/3}\kappa_x\kappa_y^2)\log(1/\epsilon)\big)$ stochastic first-order oracle (SFO) calls,[2] where κ_x and κ_y are the condition numbers with respect to PL condition for x and y respectively. The variance reduced technique in the SVRG-AGDA leads to better a convergence rate than full batch AGDA whose SFO complexity is $\mathcal{O}\big(n\kappa_x\kappa_y^2\log(1/\epsilon)\big)$. However, there are still some open questions left. Firstly, Yang et al. [45]'s theoretical analysis heavily relies on the alternating update rules. It remains interesting whether a simultaneous version of GDA (or its stochastic variants) also has similar convergence results. Secondly, it is unclear whether the SFO upper bound obtain by SVRG-AGDA can be improved by designing more efficient algorithms.

For one-sided PL condition, we desire to find the stationary point of $g(x)\triangleq\max_{y\in\mathbb{R}^{d_y}}f(x,y)$, since the saddle point may not exist. Nouiehed et al. [32] proposed the multi-step GDA method that achieves the ϵ-stationary point within $\mathcal{O}(\kappa_y^2L\epsilon^{-2}\log(\kappa_y/\epsilon))$ numbers of full gradient iterations. The similar complexity also can be obtained by AGDA [45]. Recently, Yang et al. [47] proposed the smoothed-AGDA that improves the upper bound into $\mathcal{O}(\kappa_yL\epsilon^{-2})$. Both multi-step GDA Nouiehed et al. [32] and smoothed-AGDA [46] can be extended to online setting [14], but the formulation (1) with finite-sum structure has not been explored.

In this paper, we introduce a variance reduced first-order method, called SPIDER-GDA, which constructs the gradient estimator by stochastic recursive gradient and the iterations are based on simultaneous gradient descent ascent. We prove that SPIDER-GDA could achieve ϵ-approximate solution of the two-sided PL problem of the form (1) within $\mathcal{O}\big((n+\sqrt{n}\,\kappa_x\kappa_y^2)\log(1/\epsilon)\big)$ SFO calls, which has better dependency on n than SVRG-AGDA [45]. We also provide an acceleration framework to improve first-order methods for solving ill-conditioned minimax problems under PL conditions. The accelerated SPIDER-GDA (AccSPIDER-GDA) could achieve ϵ-approximate solution within $\tilde{\mathcal{O}}\big((n+\sqrt{n}\,\kappa_x\kappa_y)\log^2(1/\epsilon)\big)$ SFO calls[3] when $\kappa_y\gtrsim\sqrt{n}$, which is the best known SFO upper bound for this problem. We summarize our main results and compare them with related work in Table 1. Without loss of generality, we always suppose $\kappa_x\gtrsim\kappa_y$. Furthermore, the proposed algorithms also work for minimax problem with one-sided PL condition. We present the results for this case in Table 2.

2 Related Work

The minimax optimization problem (1) can be viewed as the following minimization problem

$$\min_{x\in\mathbb{R}^{d_x}}\left\{g(x)\triangleq\max_{y\in\mathbb{R}^{d_y}}f(x,y)\right\}.$$

A natural way to solve such problem is the multi-step GDA algorithm [21, 25, 32, 36] that contains double-loop iterations in which the outer loop can be regarded as running inexact gradient descent on

[2]The original analysis [45] provided an SFO upper bound $\mathcal{O}\big((n+n^{2/3}\max\{\kappa_x^3,\kappa_y^3\})\log(1/\epsilon)\big)$, which can be refined to $\mathcal{O}\big((n+n^{2/3}\kappa_x\kappa_y^2)\log(1/\epsilon)\big)$ by some little modification in the proof.

[3]In this paper, we ues the notation $\tilde{\mathcal{O}}(\cdot)$ to hide the logarithmic factors of κ_x,κ_y but not $1/\epsilon$.

Table 2: We present the comparison of SFO complexities under one-sided PL condition.

Algorithm	Complexity	Reference
Multi-Step GDA	$\mathcal{O}(n\kappa_y^2 L\epsilon^{-2}\log(\kappa_y/\epsilon))$	[32]
GDA/AGDA	$\mathcal{O}\left(n\kappa_y^2 L\epsilon^{-2}\right)$	Theorem B.2, [45]
Smooothed-AGDA	$\mathcal{O}\left(n\kappa_y L\epsilon^{-2}\right)$	[47]
SVRG-GDA	$\mathcal{O}\left(n + n^{2/3}\kappa_y^2 L\epsilon^{-2}\right)$	Theorem F.1
SPIDER-GDA	$\mathcal{O}\left(n + \sqrt{n}\kappa_y^2 L\epsilon^{-2}\right)$	Theorem 6.1
AccSPIDER-GDA	$\begin{cases} \mathcal{O}\left(\sqrt{n}\kappa_y L\epsilon^{-2}\log(\kappa_y/\epsilon)\right), & \sqrt{n} \lesssim \kappa_y; \\ \mathcal{O}\left(nL\epsilon^{-2}\log(\kappa_y/\epsilon)\right), & \kappa_y \lesssim \sqrt{n} \lesssim \kappa_y^2; \\ \mathcal{O}\left(n + \sqrt{n}\kappa_y^2 L\epsilon^{-2}\right), & \kappa_y^2 \lesssim \sqrt{n}. \end{cases}$	Theorem 6.2

$g(x)$ and the inner loop finds the approximate solution to $\max_{y\in\mathbb{R}^{d_y}} f(x,y)$ for a given x. Another class of methods is the two-timescale (alternating) GDA algorithm [9, 21, 44, 45] that only has single-loop iterations which update two variables with different stepsizes. The two-timescale GDA method can be implemented more easily and typically performs better than multi-step GDA empirically. Its convergence rate also can be established by analyzing function $g(x)$ but the analysis is more challenging than the multi-step GDA.

The variance reduction is a popular technique to improve the efficiency of stochastic optimization algorithms [2–4, 7, 8, 12, 16, 17, 19, 27, 31, 33, 34, 37–39, 43, 49, 50]. It is shown that solving nonconvex minimization problems with stochastic recursive gradient estimator [12, 16, 34, 43, 51] has the optimal SFO complexity. In the context of minimax optimization, the variance reduced algorithms also obtain the best-known SFO complexities in several settings [1, 15, 25, 26, 41, 45]. Specifically, the (near) optimal SFO algorithm for several convex-concave minimax problem has been proposed [15, 26], but the optimality for the more general case is still unclear [25, 45].

The Catalyst acceleration [20] is a useful approach to reduce the computational cost of ill-conditioned optimization problems, which is based on a sequence of inexact proximal point iterations. Lin et al. [22] first introduced Catalyst into minimax optimization. Later, Luo et al. [26], Tominin et al. [41], Yang et al. [46] designed the accelerated stochastic algorithms for convex-concave and nonconvex-concave problem. Recently, Yang et al. [47] also applied this technique to one-sided PL setting.

3 Notation and Preliminaries

First of all, we present the definition of saddle point.

Definition 3.1. *We say $(x^*, y^*) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ is a saddle point of function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ if it holds that $f(x^*, y) \le f(x^*, y^*) \le f(x, y^*)$ for any $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$.*

Then we formally define the Polyak-Łojasiewicz (PL) condition [35] as follows.

Definition 3.2. *We say a differentiable function $h : \mathbb{R}^d \to \mathbb{R}$ satisfies μ-PL for some $\mu > 0$ if $\|\nabla h(z)\|^2 \ge 2\mu\big(h(z) - \min_{z'\in\mathbb{R}^d} h(z')\big)$ holds for any $z \in \mathbb{R}^d$.*

Note that the PL condition does not require the strongly convexity and it can be satisfied even if the function is nonconvex [18].

We are interested in the finite-sum minimax optimization problem (1) under following assumptions.

Assumption 3.1. *We suppose each component $f_i : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is L-smooth, i.e., there exists a constant $L > 0$ such that $\|\nabla f_i(x, y) - \nabla f_i(x', y')\|^2 \le L^2\left(\|x - x'\|^2 + \|y - y'\|^2\right)$ holds for any $x, x' \in \mathbb{R}^{d_x}$ and $y, y' \in \mathbb{R}^{d_y}$.*

Assumption 3.2. *We suppose the differentiable function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ satisfies two-sided PL condition, i.e., there exist constants $\mu_x > 0$ and $\mu_y > 0$ such that $f(\cdot, y)$ is μ_x-PL for any $y \in \mathbb{R}^{d_y}$ and $-f(x, \cdot)$ is μ_y-PL for any $x \in \mathbb{R}^{d_x}$.*

Under Assumption 3.1 and 3.2, we define the condition numbers of problem (1) with respect to PL conditions for x and y as $\kappa_x \triangleq L/\mu_x$ and $\kappa_y \triangleq L/\mu_y$ respectively.

We also introduce the following assumption for the existence of saddle point.

Assumption 3.3 (Yang et al. [45])**.** *We suppose the function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ has at least one saddle point (x^*, y^*). We also suppose that for any fixed $y \in \mathbb{R}^{d_y}$, the problem $\min_{x\in\mathbb{R}^{d_x}} f(x, y)$ has a nonempty solution set and a finite optimal value; and for any fixed $x \in \mathbb{R}^{d_x}$, the problem $\max_{y\in\mathbb{R}^{d_y}} f(x, y)$ has a nonempty solution set and a finite optimal value.*

The goal of solving minimax problem under two-sided PL condition is finding an ϵ-approximate solution or ϵ-saddle point that is defined as follows.

Definition 3.3. *We say x is an ϵ-approximate solution of problem (1) if it holds that $g(x) - g(x^*) \le \epsilon$, where $g(x) = \max_{y\in\mathbb{R}^{d_y}} f(x, y)$.*

Definition 3.4. *Under Assumption 3.3, we say (x, y) is an ϵ-saddle point of problem (1) if it holds that $\|x - x^*\|^2 + \|y - y^*\|^2 \le \epsilon$ for some saddle point (x^*, y^*).*

We allow the saddle point does not exist for the problem with one-sided PL condition. In such case, it is guaranteed that $g(x) \triangleq \max_{y\in\mathbb{R}^{d_y}} f(x, y)$ is differentiable [32, Lemma A.5] and we target to find an ϵ-stationary point of $g(x)$.

Definition 3.5. *If the function $g : \mathbb{R}^{d_x} \to \mathbb{R}$ is differentiable, we say x is an ϵ-stationary point of g if it holds that $\|\nabla g(x)\| \le \epsilon$.*

4 A Faster Algorithm for the Two-Sided PL Condition

We first consider the two-sided PL conditioned minimax problem of the finite-sum form (1) under Assumption 3.1, 3.2 and 3.3. We propose a novel stochastic algorithm, which we refer to as SPIDER-GDA. The detailed procedure of our method is presented in Algorithm 1. SPIDER-GDA constructs the stochastic recursive gradient estimators [12, 31] as follows:

$$\begin{aligned}G_x(x_{t,k}, y_{t,k}) =& \frac{1}{B}\sum_{i\in S_x}\big(\nabla_x f_i(x_{t,k}, y_{t,k}) - \nabla_x f_i(x_{t,k-1}, y_{t,k-1}) + G_x(x_{t,k-1}, y_{t,k-1})\big),\\ G_y(x_{t,k}, y_{t,k}) =& \frac{1}{B}\sum_{i\in S_y}\big(\nabla_y f_i(x_{t,k}, y_{t,k}) - \nabla_y f_i(x_{t,k-1}, y_{t,k-1}) + G_y(x_{t,k-1}, y_{t,k-1})\big).\end{aligned}$$

It simultaneously updates two variables $\mathbf{x}$ and $\mathbf{y}$ by estimators G_x and G_y with different stepsizes $\tau_x = \Theta(1/(\kappa_y^2 L))$ and $\tau_y = \Theta(1/L)$ respectively. Huang et al. [16], Luo et al. [25], Xian et al. [44] have studied the SPIDER-type algorithm for nonconvex-strongly-concave problem and showed it converges to the stationary point of $g(x) \triangleq \max_{y\in\mathbb{R}^{d_y}} f(x, y)$ sublinearly. However, solving the problem minimax problems with two-sided PL condition desires stronger linear convergence rate, which leads to our theoretical analysis be different from previous work.

We measure the convergence of SPIDER-GDA by the following Lyapunov function

$$\mathcal{V}_{t,k} \triangleq g(x_{t,k}) - g(x^*) + \frac{\lambda\tau_x}{\tau_y}\big(g(x_{t,k}) - f(x_{t,k}, y_{t,k})\big),$$

where $x^* \in \arg\min_{x\in\mathbb{R}^{d_x}} g(x)$ and $\lambda = \Theta(\kappa_y^2)$. We can establish recursion for $\mathcal{V}_{t,k}$ as follows

$$\mathbb{E}[\mathcal{V}_{t,K}] \le \mathbb{E}\left[\mathcal{V}_{t,0} - \frac{\tau_x}{16}\left(2 - \frac{M}{B}\right)\sum_{k=0}^{K-1}\|G_x(x_{t,k}, y_{t,k})\|^2 - \frac{\lambda\tau_x}{16}\left(2 - \frac{M}{B}\right)\sum_{k=0}^{K-1}\|G_y(x_{t,k}, y_{t,k})\|^2\right].$$

Using the above inequality by setting $M = B = \sqrt{n}$ leads to the estimators $G_x(\tilde{x}_t, \tilde{y}_t)$ and $G_y(\tilde{x}_t, \tilde{y}_t)$ be sufficiently close to the exact gradient and converge to zero linearly, which indicates $g(\tilde{x}_t)$ also converges to $g(x^*)$ linearly. We formally provide the convergence result for SPIDER-GDA in the following theorem and its detailed proof is shown in appendix.

Theorem 4.1. *Under Assumption 3.1, 3.2 and 3.3, we run Algorithm 1 with $M = B = \sqrt{n}$, $\tau_y = 1/(5L)$, $\lambda = 32L^2/\mu_y^2$, $\tau_x = \tau_y/(24\lambda)$, $K = \lceil 4224/(\mu_x\tau_x)\rceil$ and $T = \lceil\log(1/\epsilon)\rceil$. Then the output $(\tilde{x}_T, \tilde{y}_T)$ satisfies $g(\tilde{x}_T) - g(x^*) \le \epsilon$ and $g(\tilde{x}_T) - f(\tilde{x}_T, \tilde{y}_T) \le 24\epsilon$ in expectation; and it takes no more than $\mathcal{O}\big((n + \sqrt{n}\kappa_x\kappa_y^2)\log(1/\epsilon)\big)$ SFO calls.*

Algorithm 1 SPIDER-GDA $(f, (x_0, y_0), T, K, M, B, \tau_x, \tau_y)$

$\tilde{x}_0 = x_0, \tilde{y}_t = y_0$
for $t = 0, 1, \ldots, T-1$ **do**
  $x_{t,0} = \tilde{x}_t, y_{t,0} = \tilde{y}_t$
  for $k = 0, 1, \ldots, K-1$ **do**
    if $\mod(k, M) = 0$ **then**
      $G_x(x_{t,k}, y_{t,k}) = \nabla_x f(x_{t,k}, y_{t,k})$
      $G_y(x_{t,k}, y_{t,k}) = \nabla_y f(x_{t,k}, y_{t,k})$
    else
      draw mini-batches S_x and S_y independently with both sizes of B.
      $G_x(x_{t,k}, y_{t,k}) = \frac{1}{B}\sum_{i \in S_x}[\nabla_x f_i(x_{t,k}, y_{t,k}) - \nabla_x f_i(x_{t,k-1}, y_{t,k-1}) + G_x(x_{t,k-1}, y_{t,k-1})]$
      $G_y(x_{t,k}, y_{t,k}) = \frac{1}{B}\sum_{i \in S_y}[\nabla_y f_i(x_{t,k}, y_{t,k}) - \nabla_y f_i(x_{t,k-1}, y_{t,k-1}) + G_y(x_{t,k-1}, y_{t,k-1})]$
    end if
    $x_{t,k+1} = x_{t,k} - \tau_x G_x(x_{t,k}, y_{t,k})$
    $y_{t,k+1} = x_{y,k} + \tau_y G_y(x_{t,k}, y_{t,k})$
  end for
  choose $(\tilde{x}_{t+1}, \tilde{y}_{t+1})$ from $\{(x_{t,k}, y_{t,k})\}_{k=0}^{K-1}$ uniformly at random.
end for
return $(\tilde{x}_T, \tilde{y}_T)$

Algorithm 2 AccSPIDER-GDA

$u_0 = x_0$
for $k = 0, 1, \ldots, K-1$ **do**
  $(x_{k+1}, y_{k+1}) = \text{SPIDER-GDA}\big(f(x,y) + \frac{\beta}{2}\|x - u_k\|^2, (x_k, y_k), T_k, K, M, B, \tau_x, \tau_y\big)$
  $u_{k+1} = x_{k+1} + \gamma(x_{k+1} - x_k)$
end for
option I (two-sided PL): **return** (x_K, y_K)
option II (one-sided PL): **return** $(\hat{x}, \hat{y})$ chosen uniformly at random from $\{(x_k, y_k)\}_{k=0}^{K-1}$

Our results provide an SFO upper bound of $\mathcal{O}((n + \sqrt{n}\kappa_x\kappa_y^2)\log(1/\epsilon))$ for finding an ε-approximate solution that is better than the complexity $\mathcal{O}((n + n^{2/3}\kappa_x\kappa_y^2)\log(1/\epsilon))$ derived from SVRG-AGDA [45]. It is possible to use SVRG-type [17, 49] estimators to replace the stochastic recursive estimators in Algorithm 1, which results the algorithm SVRG-GDA. We can prove that SVRG-GDA also has $\mathcal{O}((n + n^{2/3}\kappa_x\kappa_y^2)\log(1/\epsilon))$ SFO upper bound that matches the theoretical result of SVRG-AGDA. We provide the details in Appendix C.

5 Further Acceleration with Catalyst

Both the proposed SPIDER-GDA (Algorithm 1) and existing SVRG-AGDA [45] have the complexities more heavily depend on the condition number of y than the condition number of x. It is natural to ask can we make the dependency of two condition numbers balanced like the results in the strongly-convex-strongly-concave case [22, 25, 41]. In this section, we show it is possible by introducing the Catalyst acceleration.

To make acceleration possible, we need to assume the uniqueness of the optimal set for inner problem.

Assumption 5.1. *We assume the inner problem* $\max_{y \in \mathbb{R}^{d_y}} f(x, y)$ *has an unique solution.*

We proposed the accelerated SPIDER-GDA (AccSPIDER-GDA) in Algorithm 2 for reducing the computational cost further. Each iteration of the algorithm solve the following sub-problem

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} F_k(x,y) \triangleq \min_{x\in\mathbb{R}^{d_x}} \left\{ g(x) + \frac{\beta}{2}\|x-u_k\|_2^2 \right\}. \tag{2}$$

by SPIDER-GDA (Algorithm 1). AccSPIDER-GDA has the following convergence result if the sub-problem attain the required accuracy.

Lemma 5.1. *Under Assumption 3.1, 3.2 and 3.3, we run Algorithm 2 by $\beta = 2L$, $\gamma = 0$ and the appropriate setting for the sub-problem solver such that $\mathbb{E}[\|x_k - \tilde{x}_k\|^2 + \|y_k - \tilde{y}_k\|^2] \le \delta$, where $(\tilde{x}_k, \tilde{y}_k)$ is a saddle point of F_{k-1} $(k \ge 1)$ and we set the precision*

$$\delta = \frac{\mu_x \epsilon}{11(\mu_x + 4L)L} \tag{3}$$

Then it holds that

$$\mathbb{E}[g(x_k) - g(x^*)] \le \left(1 - \frac{\mu_x}{2\beta + \mu_x}\right)^k \left(g(x_0) - g(x^*)\right) + \frac{\epsilon}{2}.$$

The setting $\beta = \Theta(L)$ in Lemma 5.1 guarantees the sub-problem (2) has condition number of the order $\mathcal{O}(1)$ for x. It is more well-conditioned on x, we prefer to address the following equivalent problem

$$\max_{y\in\mathbb{R}^{d_y}} \min_{x\in\mathbb{R}^{d_x}} F_k(x,y) = -\min_{y\in\mathbb{R}^{d_y}} \max_{x\in\mathbb{R}^{d_x}} \left\{-F_k(x,y)\right\}. \tag{4}$$

Since (4) is a minimax problem satisfying two sided PL condition, we can apply SPIDER-GDA to solve it. And we can show that under Assumption 5.1, the saddle point $(\tilde{x}_k, \tilde{y}_k)$ of each $F_{k-1}(k \ge 1)$ is unique (see Lemma E.2 in appendix) and we are able to obtain a good approximation to it.

Lemma 5.2. *Under Assumption 3.1, 3.2 and 3.3, if we use Algorithm 1 to solve each sub-problem $\max_{y\in\mathbb{R}^{d_y}} \min_{x\in\mathbb{R}^{d_x}} F_k(x,y)$ (2) with $\beta = 2L$, $M = B = \sqrt{n}$, $\tau_x = 1/(15L)$, $\lambda = 288$, $\tau_y = \tau_x/(24\lambda)$, $K = \lceil 4224/(\mu_y \tau_y)\rceil$, $T_k = \lceil \log(1/\delta_k)\rceil$, then it holds that*

$$\mathbb{E}[\|x_{k+1} - \tilde{x}_{k+1}\|^2 + \|y_{k+1} - \tilde{y}_{k+1}\|^2] \le 7236\kappa_y^2 \delta_k \mathbb{E}[\|x_k - \tilde{x}_k\|^2 + \|y_k - \tilde{y}_k\|^2],$$

where $(\tilde{x}_k, \tilde{y}_k)$ is the unique saddle point of $F_{k-1}(k \ge 1)$.

For a short summary, Lemma 5.1 means Algorithm 2 requires $\mathcal{O}(\kappa_x \log(1/\epsilon))$ numbers of inexact proximal point iterations to find an ϵ-approximate solution of the problem. And Lemma E.1 tells us that each sub-problem can be solved within a SFO complexity of $\mathcal{O}\left(n + \sqrt{n}\kappa_y\right)\log(1/\delta_k))$. Thus, the total complexity for AccSPIDER-GDA becomes $\mathcal{O}((n\kappa_x + \sqrt{n}\kappa_x\kappa_y)\log(1/\epsilon)\log(1/\delta_k))$. Our next step is to specify δ_k which would lead to the total SFO complexity of the algorithm.

Theorem 5.1. *Under Assumption 3.1, 3.2, 3.3 and 5.1 if we let $\gamma = 0, \beta = 2L$ and use Algorithm 1 to solve each sub-problem $\max_{y\in\mathbb{R}^{d_y}} \min_{x\in\mathbb{R}^{d_x}} F_k(x,y)$ (2) with M, B, τ_x, τ_y, K defined as Lemma 5.2 and $T_k = \lceil \log(1/\delta_k)\rceil$, where*

$$\delta_k = \begin{cases} \frac{1}{7236\kappa_y^2} \min\left\{\frac{1}{4}, \frac{(\beta-L)\mu_y\delta}{16\beta^2\|x_k - x_{k-1}\|^2}\right\}, & k \ge 1; \\ \frac{\delta\mu_y}{14472\kappa_y^2(g(x_0)-g(x^*))}, & k = 0, \end{cases} \tag{5}$$

and δ is followed by the definition in (3). Then Algorithm 2 can return x_K such that $g(x_K) - g(x^*) \le \epsilon$ in expectation with no more than $\mathcal{O}((n\kappa_x + \sqrt{n}\kappa_x\kappa_y)\log(1/\epsilon)\log(\kappa_x\kappa_y/\epsilon))$ SFO calls.

Lemma 5.1 does not rely on the choice of sub-problem solver, we can apply the acceleration framework in Algorithm 2 by replacing SPIDER-GDA with other algorithms. We summarize the SFO complexities for the acceleration of different algorithms in Table 3.

6 Extension to One-Sided PL Condition

In this section, we show the idea that SPIDER-GDA and its Catalyst acceleration also work for one-sided PL condition. We relax Assumption 3.2 and 3.3 to the following one.

Table 3: Accelerated results for different methods under two-sided PL condition.

Method	Before Acceleration	After Acceleration
GDA	$\mathcal{O}(n\kappa_x\kappa_y^2\log(1/\epsilon))$	$\tilde{\mathcal{O}}\left(n\kappa_x\kappa_y\log^2(1/\epsilon)\right)$
SVRG-GDA	$\mathcal{O}((n+n^{2/3}\kappa_x\kappa_y^2)\log(1/\epsilon))$	$\begin{cases}\tilde{\mathcal{O}}\left(n^{2/3}\kappa_x\kappa_y\log^2(1/\epsilon)\right), & n^{1/3}\lesssim\kappa_y;\\ \tilde{\mathcal{O}}\left(n\kappa_x\log^2(1/\epsilon)\right), & \kappa_y\lesssim n^{1/3}\lesssim\kappa_x\kappa_y;\\ \text{no acceleration}, & \kappa_x\kappa_y\lesssim n^{1/3}.\end{cases}$
SPIDER-GDA	$\mathcal{O}\left((n+\sqrt{n}\kappa_x\kappa_y^2)\log(1/\epsilon)\right)$	$\begin{cases}\tilde{\mathcal{O}}\left(\sqrt{n}\kappa_x\kappa_y\log^2(1/\epsilon)\right), & \sqrt{n}\lesssim\kappa_y;\\ \tilde{\mathcal{O}}\left(n\kappa_x\log^2(1/\epsilon)\right), & \kappa_y\lesssim\sqrt{n}\lesssim\kappa_x\kappa_y;\\ \text{no acceleration}, & \kappa_x\kappa_y\lesssim\sqrt{n}.\end{cases}$

Assumption 6.1. *We suppose that $-f(x,\cdot)$ is μ_y-PL for any $x\in\mathbb{R}^{d_x}$; the problem $\max_{y\in\mathbb{R}^{d_y}} f(x,y)$ has a nonempty solution set and an optimal value ; $g(x)\triangleq\max_{y\in\mathbb{R}^{d_y}} f(x,y)$ is lower bounded, i.e., we have $g^*=\inf_{x\in\mathbb{R}^{d_x}} g(x)>-\infty$.*

We first show that the SFO complexity of SPIDER-GDA outperforms SVRG-GDA [4] by a factor of $\mathcal{O}(n^{1/6})$ in Theorem 6.1.

Theorem 6.1. *Under Assumption 3.1 and 6.1 , Let $T=1$ and $M,B,\tau_x,\tau_y,\lambda$ as defined in Theorem 4.1 and $K=\lceil 64/(\tau_x\epsilon^2)\rceil$, then Algorithm 1 can guarantee the output $\hat{x}$ to satisfy $\|\nabla g(\hat{x})\|\le\epsilon$ in expectation with no more than $\mathcal{O}(n+\sqrt{n}\kappa_y^2 L\epsilon^{-2})$ SFO calls.*

The AccSPIDER-GDA also performs better than SPIDER-GDA in one-sided PL condition for ill conditioned case. In the following lemma, we show that AccSPIDER-GDA could find an approximate stationary point if we solve the sub-problem sufficiently accurate.

Lemma 6.1. *Under Assumption 3.1 and 6.1, if it holds true that $\mathbb{E}[\|x_k-\tilde{x}_k\|^2+\|y_k-\tilde{y}_k\|^2]\le\delta$ for some saddle point $(\tilde{x}_k,\tilde{y}_k)$ of F_{k-1} $(k\ge 1)$, where*

$$\delta=\frac{\epsilon^2}{8L\kappa_y(22\mu_y+1)}. \tag{6}$$

Let $\beta=2L$, then for the output $(\hat{x},\hat{y})$ of Algorithm 2, it holds true that

$$\mathbb{E}\|\nabla g(\hat{x})\|^2\le\frac{8\beta(g(x_0)-g^*)}{K}+\frac{\epsilon^2}{2}.$$

Compared with two-sided PL condition, the analysis of AccSPIDER-GDA is more complicated since the precision δ_k at each round are different. By choosing the parameters of the algorithm carefully, we obtain the following results.

Theorem 6.2. *Under Assumption 3.1, 6.1 and 5.1, if we run Algorithm 2 by $\gamma=0,\beta=2L$ and use Algorithm 1 to solve each sub-problem $\max_{y\in\mathbb{R}^{d_y}}\min_{x\in\mathbb{R}^{d_x}} F_k(x,y)$ (2) with $M,B,\tau_x,\tau_y,\lambda,K$ and T_k (dependent on δ) as in Theorem 5.1 and δ is followed by the definition in Lemma 6.1, then Algorithm 2 can find $\hat{x}$ such that $\|\nabla g(\hat{x})\|\le\epsilon$ in expectation within $\mathcal{O}((n+\sqrt{n}\kappa_y)L\epsilon^{-2}\log(\kappa_y/\epsilon))$ SFO calls.*

We can directly set $\beta=0$ for Algorithm 2 in the case of very large n and in this case AccSPIDER-GDA reduces to SPIDER-GDA. The summary and comparison for the complexities for the one-sided PL condition is shown in Table 2. Besides, the algorithms of GDA and SVRG-GDA also can be accelerated with Catalyst framework and we present the corresponding results in Table 4.

[4]The complexity for finding an ϵ-stationary point of SVRG-GDA in presented in Appendix F.

Table 4: Acceleration for different methods under one-sided PL condition.

Method	Before Acceleration	After Acceleration
GDA	$\mathcal{O}\left(n\kappa_y^2 L\epsilon^{-2}\right)$	$\mathcal{O}\left(n\kappa_y L\epsilon^{-2}\log(\kappa_y/\epsilon)\right)$
SVRG-GDA	$\mathcal{O}\left(n+n^{2/3}\kappa_y^2 L\epsilon^{-2}\right)$	$\begin{cases}\mathcal{O}\left(n^{2/3}\kappa_y L\epsilon^{-2}\log(\kappa_y/\epsilon)\right), & n^{1/3}\lesssim\kappa_y;\\ \mathcal{O}\left(nL\epsilon^{-2}\log(\kappa_y/\epsilon)\right), & \kappa_y\lesssim n^{1/3}\lesssim\kappa_y^2;\\ \text{no acceleration}, & \kappa_y^2\lesssim n^{1/3}.\end{cases}$
SPIDER-GDA	$\mathcal{O}\left(n+\sqrt{n}\kappa_y^2 L\epsilon^{-2}\right)$	$\begin{cases}\mathcal{O}\left(\sqrt{n}\kappa_y L\epsilon^{-2}\log(\kappa_y/\epsilon)\right), & \sqrt{n}\lesssim\kappa_y;\\ \mathcal{O}\left(nL\epsilon^{-2}\log(\kappa_y/\epsilon)\right), & \kappa_y\lesssim\sqrt{n}\lesssim\kappa_y^2;\\ \text{no acceleration}, & \kappa_y^2\lesssim\sqrt{n}.\end{cases}$

7 Experiments

In this section, we conduct the numerical experiments to show the advantage of proposed algorithms and the source code is available[5]. We consider the following two player Polyak-Łojasiewicz game:

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^d} f(x,y)\triangleq\frac{1}{2}x^\top Px-\frac{1}{2}y^\top Qy+x^\top Ry,$$

where

$$P=\frac{1}{n}\sum_{i=1}^n p_ip_i^\top,\quad Q=\frac{1}{n}\sum_{i=1}^n q_iq_i^\top\quad\text{and}\quad R=\frac{1}{n}\sum_{i=1}^n r_ir_i^\top.$$

We independently sample p_i, q_i and r_i from $\mathcal{N}(0,\Sigma_P)$, $\mathcal{N}(0,\Sigma_Q)$ and $\mathcal{N}(0,\Sigma_R)$ respectively. We set the covariance matrix Σ_P as the form of $UDU^\top$ such that $U\in\mathbb{R}^{d\times r}$ is column orthogonal matrix and $D\in\mathbb{R}^{r\times r}$ is diagonal with $r<d$. The diagonal elements of D are distributed uniformly in the interval $[\mu,L]$ with $0<\mu<L$. The matrix Σ_Q is set by the similar way to Σ_P. We also let $\Sigma_R=0.1VV^\top$, where each element of $V\in\mathbb{R}^{d\times d}$ is sampled from $\mathcal{N}(0,1)$ independently. Since the covariance matrices Σ_P and Σ_Q are rank-deficient, it is guaranteed that both P and Q are singular. Hence, the objective function is not strongly-convex and not strongly-concave, but it satisfies the two-sided PL-condition [18]. We set $n=6000$, $d=10$, $r=5$, $L=1$ for all experiments; and let μ be 10^{-5} and 10^{-9} for two different settings.

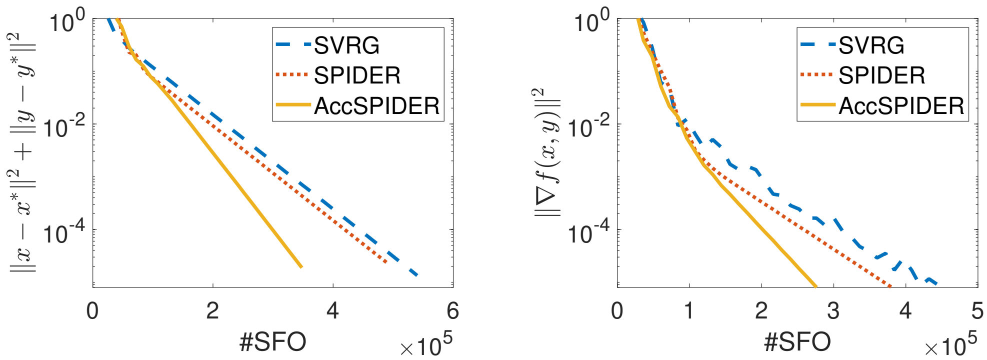
We compare the proposed SPIDER-GDA (Algorithm 1) and AccSPIDER-GDA (Algorithm 2) with the baseline algorithm SVRG-AGDA [45]. We let $B=1$ and $M=n$ for all of these algorithms and both of the stepsizes for x and y are tuned from $\{10^{-1},10^{-2},10^{-3},10^{-4},10^{-5}\}$. For AccSPIDER, we set $\beta=L/(20n)$ and $\gamma=0.999$. We present the results of the number of SFO calls against the norm of gradient and the distance to the saddle point in Figure 1 and Figure 2. It is clear that our algorithms outperform than baselines.

8 Conclusion and Future Work

In this paper, we have investigated stochastic optimization for PL conditioned minimax problem with the finite-sum objective. We have proposed the SPIDER-GDA algorithm, which reduces the dependency of the sample numbers in SFO complexity. Moreover, we have introduced a Catalyst scheme to accelerate our algorithm for solving the ill-conditioned problems. We improve the SFO upper bound of the state-of-the-art algorithms for both two-sided and one-sided PL conditions.

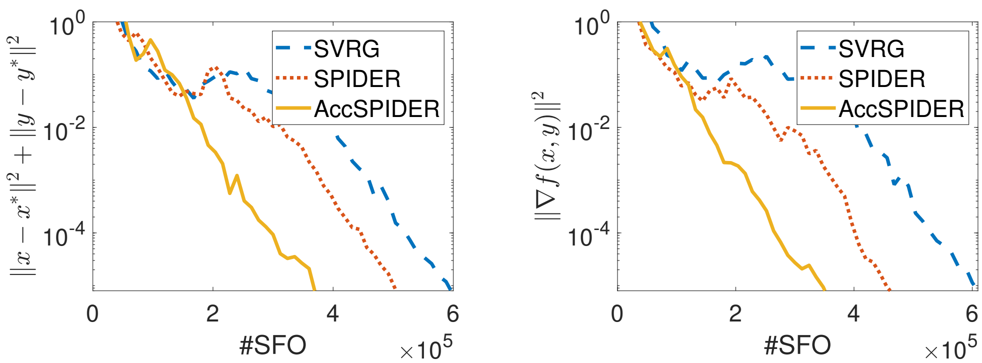
However, the optimality of SFO algorithms for the PL conditioned minimax problem is still unclear. It is interesting to construct the lower bound for verifying the tightness of our results. It is also possible to extend our algorithm to online setting.

[5] `https://github.com/TrueNobility303/SPIDER-GDA`



(a) Distance to saddle point (b) Norm of gradient

Figure 1: The comparison for the case of $\mu = 10^{-5}$



(a) Distance to saddle point (b) Norm of gradient

Figure 2: The comparison for the case of $\mu = 10^{-9}$

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62206058) and Shanghai Sailing Program (22YF1402900).

References

[1] Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. *arXiv preprint arXiv:2102.08352*, 2021.

[2] Zeyuan Allen-Zhu. Katyusha x: Practical momentum method for stochastic sum-of-nonconvex optimization. *arXiv preprint arXiv:1802.03866*, 2018.

[3] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *ICML*, 2016.

[4] Zeyuan Allen-Zhu and Yang Yuan. Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In *ICML*, 2016.

[5] Qi Cai, Mingyi Hong, Yongxin Chen, and Zhaoran Wang. On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv preprint arXiv:1901.03674*, 2019.

[6] Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Variance reduction for matrix games. In *NeurIPS*, 2019.

[7] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. In *NeurIPS*, 2019.

[8] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, 2014.

[9] Thinh T. Doan. Convergence rates of two-time-scale gradient descent-ascent dynamics for solving nonconvex min-max problems. *arXiv preprint arXiv:2112.09579*, 2021.

[10] Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *ICML*, 2017.

[11] John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *The Journal of Machine Learning Research*, 20(1):2450–2504, 2019.

[12] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *NeurIPS*, 2018.

[13] Zhishuai Guo, Mingrui Liu, Zhuoning Yuan, Li Shen, Wei Liu, and Tianbao Yang. Communication-efficient distributed stochastic auc maximization with deep neural networks. In *ICML*, 2020.

[14] Zhishuai Guo, Zhuoning Yuan, Yan Yan, and Tianbao Yang. Fast objective & duality gap convergence for nonconvex-strongly-concave min-max problems. *arXiv preprint arXiv:2006.06889*, 2020.

[15] Yuze Han, Guangzeng Xie, and Zhihua Zhang. Lower complexity bounds of finite-sum optimization problems: The results and construction. *arXiv preprint arXiv:2103.08280*, 2021.

[16] Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23(36):1–70, 2022.

[17] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.

[18] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016.

[19] Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *ICML*, 2021.

[20] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *NIPS*, 2015.

[21] Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, 2020.

[22] Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. In *COLT*, 2020.

[23] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 2022.

[24] Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic auc maximization with deep neural networks. *arXiv preprint arXiv:1908.10831*, 2019.

[25] Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. In *NeurIPS*, 2020.

[26] Luo Luo, Guangzeng Xie, Tong Zhang, and Zhihua Zhang. Near optimal stochastic algorithms for finite-sum unbalanced convex-concave minimax optimization. *arXiv preprint arXiv:2106.01761*, 2021.

[27] Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

[28] Nicolai Meinshausen. Causality from a distributional robustness point of view. In *DSW*. IEEE, 2018.

[29] John Nash. Two-person cooperative games. *Econometrica: Journal of the Econometric Society*, pages 128–140, 1953.

[30] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[31] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.

[32] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D. Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *NeurIPS*, 2019.

[33] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *NIPS*, 2016.

[34] Nhan H. Pham, Lam M. Nguyen, Dzung T. Phan, and Quoc Tran-Dinh. Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

[35] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[36] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint:1810.02060*, 2018.

[37] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *ICML*, 2016.

[38] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.

[39] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(2), 2013.

[40] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[41] Vladislav Tominin, Yaroslav Tominin, Ekaterina Borodich, Dmitry Kovalev, Alexander Gasnikov, and Pavel Dvurechensky. On accelerated methods for saddle-point problems with composite structure. *arXiv preprint arXiv:2103.09344*, 2021.

[42] Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. In *NeurIPS*, 2018.

[43] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. In *NeurIPS*, 2019.

[44] Wenhan Xian, Feihu Huang, Yanfu Zhang, and Heng Huang. A faster decentralized algorithm for nonconvex minimax problems. In *NeurIPS*, 2021.

[45] Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In *NeurIPS*, 2020.

[46] Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. In *NeurIPS*, 2020.

[47] Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *AISTATS*, 2022.

[48] Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online AUC maximization. In *NIPS*, 2016.

[49] Lijun Zhang, Mehrdad Mahdavi, and Rong Jin. Linear convergence with condition number independent access of full gradients. In *NIPS*, 2013.

[50] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. In *NeurIPS*, 2018.

[51] Pan Zhou, Xiao-Tong Yuan, and Jiashi Feng. Faster first-order methods for stochastic non-convex optimization on riemannian manifolds. In *AISTATS*, 2019.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 8.
   (c) Did you discuss any potential negative societal impacts of your work? [No] The purpose of this work is for to provide a better understanding of GDA on a class of nonconvex-nonconcave minimax optimization.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix for details.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include the codes In the supplemental materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 7.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We wants to compare the training dynamic, and different trials may cause different numerical results, which can not be observed clearly in one graph.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The experiments is certainly simple and easy to run under CPUs.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] These datasets are common.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] These datasets are common.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Some Useful Lemmas

In this section, we provide some lemmas which are useful in the following proofs.

First of all, we define three notations of optimality.

Definition A.1. *We say (x^*, y^*) is a saddle point of function f, if for all (x, y), it holds that*

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*).$$

We say (x^*, y^*) is a global minimax point, if for all $x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}$, it holds that

$$f(x^*, y) \leq f(x^*, y^*) \leq \max_{y' \in \mathbb{R}^{d_y}} f(x, y').$$

And we say (x^*, y^*) is a stationary point, if it holds that

$$\nabla_x f(x^*, y^*) = \nabla_y f(x^*, y^*) = 0.$$

For general nonconvex-nonconcave minimax problem, a stationary point or a global minimax point is weaker than a saddle point, i.e. a stationary point or a global minimax point may not be a saddle point. However, under two-sided PL condition, the above three notations are equivalent.

Lemma A.1 (Yang et al. [45, Lemma 2.1])**.** *Under Assumption 3.2, it holds that*

$$(\textit{saddle point}) \Leftrightarrow (\textit{global minimax point}) \Leftrightarrow (\textit{stationary point}).$$

Further, if (x^*, y^*) is a saddle point of f, then

$$\max_{y \in \mathbb{R}^{d_y}} f(x^*, y) = f(x^*, y^*) = \min_{x \in \mathbb{R}^{d_x}} f(x, y^*).$$

and vice versa.

It is well known that weak duality always holds.

Lemma A.2 (Nesterov [30, Theorem 1.3.1])**.** *Given a function f, we have*

$$\max_{y \in \mathbb{R}^{d_x}} \min_{x \in \mathbb{R}^{d_x}} f(x, y) \leq \min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} f(x, y).$$

It is a standard conclusion that the existence of saddle points implies strong duality. Since strong duality is important for the convergence of Catalyst scheme under PL condition, we present this lemma as follows.

Lemma A.3. *If (x^*, y^*) is a saddle point of function f, then (x^*, y^*) is also a global minimax point and stationary point of f, and it holds that*

$$\max_{y \in \mathbb{R}^{d_y}} \min_{x \in \mathbb{R}^{d_x}} f(x, y) = f(x^*, y^*) = \min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} f(x, y).$$

Lemma A.4 (Yang et al. [45, Lemma A.1])**.** *Under Assumption 3.2, then $f(x, y)$ also satisfies the following quadratic growth condition, i.e. for all $x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}$, it holds that*

$$\begin{aligned} f(x, y) - \min_{x \in \mathbb{R}^{d_x}} f(x, y) &\geq \frac{\mu_x}{2} \|x^*(y) - x\|^2, \\ \max_{y \in \mathbb{R}^{d_y}} f(x, y) - f(x, y) &\geq \frac{\mu_y}{2} \|y^*(x) - y\|^2, \end{aligned}$$

where $x^*(y)$ is the projection of y on the set $\arg\min_{x \in \mathbb{R}^{d_x}} f(x, y)$ and $y^*(x)$ is the projection of x on the set of $\arg\max_{y \in \mathbb{R}^{d_y}} f(x, y)$.

Also, we analyze the properties of function $g(x)$.

Lemma A.5 (Yang et al. [45, Lemma 2.1])**.** *Under Assumption 3.2, then $g(x)$ satisfies μ_x-PL, i.e. for all x we have*

$$\|\nabla g(x)\|^2 \geq 2\mu_x (g(x) - g(x^*)).$$

Lemma A.6 (Yang et al. [45, in the proof of Theorem 3.1])**.** *Under Assumption 3.2 and 3.1, then for all x, y it holds true that*

$$\|\nabla_x f(x,y) - \nabla g(x)\|^2 \leq \frac{2L^2}{\mu_y}(g(x) - f(x,y)).$$

The above lemma is a direct result from quadratic growth property implied by PL condition and L-smooth property of function $f(x,y)$. Using the definition of μ_y-PL in y, we can also show the relationship between $\|\nabla_x f(x,y) - \nabla g(x)\|^2$ and $\|\nabla_y f(x,y)\|^2$ as follows.

Lemma A.7. *Under Assumption 3.2 and 3.1, then for all x, y it holds true that*

$$\|\nabla_x f(x,y) - \nabla g(x)\|^2 \leq \frac{L^2}{\mu_y^2}\|\nabla_y f(x,y)\|^2$$

Lemma A.8 (Nouiehed et al. [32, Lemma A.5])**.** *Under Assumption 6.1 and 3.1, then $g(x)$ satisfies $(L + L^2/\mu_y)$-smooth,that is, it holds for all x, x' that*

$$\|\nabla g(x) - \nabla g(x')\|^2 \leq \left(L + \frac{L^2}{\mu_y}\right)\|x - x'\|^2.$$

Further, noting that $L/\mu_y \geq 1$, it implies that $g(x)$ is $(2L^2/\mu_y)$-smooth.

Lemma A.9 (Nouiehed et al. [32, Modified from Lemma A.3])**.** *Under Assumption 3.2 and 3.1, suppose (x^*, y^*) is a saddle point of f. Denote the operator $y^*(\cdot)$ as the projection onto the optimal set of $\arg\max_{y\in\mathbb{R}^{d_y}} f(x,y)$, then it holds true that*

$$\|y^*(x) - y^*\|^2 \leq \frac{L^2}{\mu_y^2}\|x - x^*\|^2.$$

Proof. The proof is similar to the proof under strongly-convex-strongly-concave setting.

$$\begin{aligned}
L^2\|x - x^*\|^2 &\geq \|\nabla_y f(x,y^*) - \nabla_y f(x^*,y^*)\|^2 \\
&= \|\nabla_y f(x,y^*)\|^2 \\
&\geq 2\mu_y\left(\max_y f(x,y) - f(x,y^*)\right) \\
&\geq \mu_y^2\|y^*(x) - y^*\|^2,
\end{aligned}$$

where the first inequality is due to L-smooth of f, and the second line relies on $\nabla_y f(x^*, y^*) = 0$. The second inequality relies on PL condition in y and in the last inequality we use the quadratic growth property by Lemma A.4. ☐

B Two-timescale GDA matches AGDA

As a warm-up, we study GDA as well as AGDA with full gradient calculation in this section. After that, it is easy to extend the analysis when we are using a gradient estimator constructed by some variance reduction framework instead of the full gradient.

Algorithm 3 AGDA $(f, (x_0, y_0), K, \tau_x, \tau_y)$

for $k = 0, 1, \ldots, K-1$ **do**

$x_{k+1} = x_k - \tau_x \nabla_x f(x_k, y_k)$

$y_{k+1} = y_k + \tau_y \nabla_y f(x_{k+1}, y_k)$

end for

option I (two-sided PL): **return** (x_K, y_K)

option II (one-sided PL): **return** $(\hat{x}, \hat{y})$ chosen uniformly at random from $\{(x_k, y_k)\}_{k=0}^{K-1}$

Algorithm 4 GDA $(f, (x_0, y_0), K, \tau_x, \tau_y)$

for $k = 0, 1, \ldots, K-1$ **do**

$x_{k+1} = x_k - \tau_x \nabla_x f(x_k, y_k)$

$y_{k+1} = y_k + \tau_y \nabla_y f(x_k, y_k)$

end for

option I (two-sided PL): **return** (x_K, y_K)

option II (one-sided PL): **return** $(\hat{x}, \hat{y})$ chosen uniformly at random from $\{(x_k, y_k)\}_{k=0}^{K-1}$

B.1 Convergence under Two-Sided PL condition

Under two-sided PL condition, it is known that AGDA [45] can find an ϵ-approximate solution to a saddle point with a complexity of $\tilde{\mathcal{O}}(n\kappa_x\kappa_y^2 \log(1/\epsilon))$ when $\kappa_x \gtrsim \kappa_y$. However, the authors left us a question that whether GDA can converge under the same setting. We answer this question affirmatively in this section. We show that the same convergence rate can be achieved by GDA algorithm with simultaneous updates.

We define the following Lyapunov function as suggested by Doan [9]:

$$\mathcal{V}_k = \mathcal{A}_k + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_k,$$

where $\mathcal{A}_k = g(x_k) - g(x^*)$, $\mathcal{B}_k = g(x_k) - f(x_k, y_k)$. Then we can obtain the following statement.

Theorem B.1. *Suppose function $f(x, y)$ satisfies L-smooth, μ_x-PL in x, μ_y-PL in y. Let $\tau_y = 1/L$, $\lambda = 6L^2/\mu_y^2$ and $\tau_x = \tau_y/(22\lambda)$, then the sequence $\{(x_k, y_k)\}_{k=1}^K$ generated by Algorithm 4 satisfies:*

$$\mathcal{V}_{k+1} \le \left(1 - \frac{\mu_x\tau_x}{2}\right)^k \mathcal{V}_k.$$

Proof. Since we know that g is $(2L^2/\mu_y)$- smooth by Lemma A.8, let $\tau_x \le \mu_y/(2L^2)$, we have

$$\begin{aligned} g(x_{k+1}) &\le g(x_k) - g(x^*) + \nabla g(x_k)^\top (x_{k+1} - x_k) + \frac{L^2}{\mu_y}\|x_{k+1} - x_k\|^2 \\ &\le g(x_k) - \tau_x \nabla g(x_k)^\top \nabla_x f(x_k, y_k) + \frac{\tau_x}{2}\|\nabla_x f(x_k, y_k)\|^2 \\ &= g(x_k) - \frac{\tau_x}{2}\|\nabla g(x_k)\|^2 + \frac{\tau_x}{2}\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2, \end{aligned} \tag{7}$$

which implies

$$\mathcal{A}_{k+1} \le \mathcal{A}_k - \frac{\tau_x}{2}\|\nabla g(x_k)\|^2 + \frac{\tau_x}{2}\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2. \tag{8}$$

Using the property of L-smooth, we know that the difference between $f(x_k, y_k)$ and $f(x_{k+1}, y_{k+1})$ can be bounded. Noting that $\tau_x \le 1/L$, we can obtain

$$\begin{aligned} f(x_k, y_k) - f(x_{k+1}, y_k) &\le -\nabla_x f(x_k, y_k)^\top (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= \tau_x\|\nabla_x f(x_k, y_k)\|^2 + \frac{\tau_x^2 L}{2}\|\nabla_x f(x_k, y_k)\|^2 \\ &\le \frac{3\tau_x}{2}\|\nabla_x f(x_k, y_k)\|^2. \end{aligned} \tag{9}$$

Let $\tau_y < 1/L$, then we have

$$
\begin{aligned}
& f(x_{k+1}, y_k) - f(x_{k+1}, y_{k+1}) \\
\leq & -\nabla_y f(x_{k+1}, y_k)^\top (y_{k+1} - y_k) + \frac{L}{2}\|y_{k+1} - y_k\|^2 \\
\leq & -\tau_y \nabla_y f(x_{k+1}, y_k)^\top \nabla_y f(x_k, y_k) + \frac{\tau_y}{2}\|\nabla_y f(x_k, y_k)\|^2 \\
= & -\frac{\tau_y}{2}\|\nabla_y f(x_{k+1}, y_k)\|^2 + \frac{\tau_y}{2}\|\nabla_y f(x_k, y_k) - \nabla_y f(x_{k+1}, y_k)\|^2 \\
\leq & -\frac{\tau_y}{4}\|\nabla_y f(x_k, y_k)\|^2 + \tau_y\|\nabla_y f(x_k, y_k) - \nabla_y f(x_{k+1}, y_k)\|^2 \\
\leq & -\frac{\tau_y}{4}\|\nabla_y f(x_k, y_k)\|^2 + \tau_y \tau_x^2 L^2 \|\nabla_x f(x_k, y_k)\|^2 \\
\leq & -\frac{\tau_y}{4}\|\nabla_y f(x_k, y_k)\|^2 + \tau_x \|\nabla_x f(x_k, y_k)\|^2,
\end{aligned}
\tag{10}
$$

where in the first inequality we use f is L-smooth, and we use $\tau_y \leq 1/L$ in the second one and Young's inequality of $-\|a-b\|^2 \leq \frac{1}{2}\|a\|^2 + \|b\|^2$ in the third one.

Combing (9) and (11), we can see that

$$
f(x_k, y_k) - f(x_{k+1}, y_{k+1}) \leq -\frac{\tau_y}{4}\|\nabla_y f(x_k, y_k)\|^2 + \frac{5\tau_x}{2}\|\nabla_x f(x_k, y_k)\|^2. \tag{11}
$$

Now we can describe how $\mathcal{B}_{k+1}$ declines compared with $\mathcal{B}_k$, using (7) and (11), we have

$$
\begin{aligned}
\mathcal{B}_{k+1} &= g(x_{k+1}) - g(x_k) + g(x_k) - f(x_k, y_k) + f(x_k, y_k) - f(x_{k+1}, y_{k+1}) \\
&\leq \mathcal{B}_k - \frac{\tau_x}{2}\|\nabla g(x_k)\|^2 + \frac{\tau_x}{2}\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2 \\
&\quad - \frac{\tau_y}{4}\|\nabla_y f(x_k, y_k)\|^2 + \frac{5\tau_x}{2}\|\nabla_x f(x_k, y_k)\|^2.
\end{aligned}
\tag{12}
$$

Using the inequality $\|\nabla_x f(x_k, y_k)\|^2 \leq 2\|g(x_k)\|^2 + 2\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2$, we have

$$
\mathcal{B}_{k+1} \leq \mathcal{B}_k + \frac{9\tau_x}{2}\|\nabla g(x_k)\|^2 + \frac{11\tau_x}{2}\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2 - \frac{\tau_y}{4}\|\nabla_y f(x_k, y_k)\|^2.
$$

By Lemma A.5, Lemma A.6 and Assumption 3.2, we have

$$
\begin{aligned}
\|\nabla g(x_k)\|^2 &\geq 2\mu_x (g(x_k) - g(x^*)), \\
\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2 &\leq \frac{2L^2}{\mu_y}(g(x_k) - f(x_k, y_k)), \\
\|\nabla_y f(x_k, y_k)\|^2 &\geq 2\mu_y (g(x_k) - f(x_k, y_k)).
\end{aligned}
\tag{13}
$$

Since we let $\tau_y = 1/L$, $\lambda = 6L^2/\mu_y^2$ and $\tau_x = \tau_y/(22\lambda)$, we can obtain

$$
\begin{aligned}
\mathcal{V}_{k+1} &= \mathcal{A}_{k+1} + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_{k+1} \\
&\leq \mathcal{A}_k + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_k - \left(1 - \frac{9\lambda\tau_x}{\tau_y}\right)\frac{\tau_x}{2}\|\nabla g(x_k)\|^2 \\
&\quad + \left(1 + \frac{11\lambda\tau_x}{\tau_y}\right)\frac{\tau_x}{2}\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2 - \frac{\lambda\tau_x}{4}\|\nabla_y f(x_k, y_k)\|^2 \\
&\leq \mathcal{A}_k - \left(1 - \frac{9\lambda\tau_x}{\tau_y}\right)\tau_x\mu_x\mathcal{A}_k + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_k + \left(1 + \frac{11\lambda\tau_x}{\tau_y}\right)\frac{\tau_x L^2}{\mu_y}\mathcal{B}_k - \frac{\lambda\tau_x\mu_y}{2}\mathcal{B}_k \\
&\leq \left(1 - \frac{\mu_x\tau_x}{2}\right)\mathcal{A}_k + \left(1 - \frac{\mu_y\tau_y}{4}\right)\frac{\lambda\tau_x}{\tau_y}\mathcal{B}_k \\
&\leq \left(1 - \frac{\mu_x\tau_x}{2}\right)\mathcal{V}_k,
\end{aligned}
\tag{14}
$$

where in the second inequality we use $11\lambda\tau_x/\tau_y \leq 1/2$ by the choices of τ_x, τ_y and λ, while we use the fact that $3\tau_x L^2/\mu_y \leq \lambda\tau_x\mu_y/2$ in the third one and $\mu_x\tau_x \leq \mu_y\tau_y/2$ in the last one.

$\square$

Now we show that the convergence of $\mathcal{V}_k$ is sufficient to guarantee the convergence to a saddle point.

Corollary B.1. *Suppose function $f(x,y)$ satisfies L-smooth, μ_x-PL in x, μ_y-PL in y and $\kappa_x \gtrsim \kappa_y$. Define τ_x, τ_y as in Lemma B.1 ,then then the sequence $\{(x_k, y_k)\}_{k=1}^K$ generated by Algorithm 4 satisfies:*

$$\|x_k - x^*\|^2 + \|y_k - y^*\|^2 \leq \frac{2c^k}{(1-\sqrt{c})^2} \max\left\{\frac{4}{\mu_x}, \frac{88}{\mu_y}\right\} \mathcal{V}_0. \tag{15}$$

where $c = 1 - \mu_x\tau_x/2$. Further, Algorithm 4 can find an ϵ-saddle point with no more than $\mathcal{O}(n\kappa_x\kappa_y^2 \log(\kappa_x\kappa_y/\epsilon))$ stochastic first-order oracle calls.

Proof. The proof is similar to the proof of Theorem 3.2 in [45].

By Lemma A.4 and the fact that $2\tau_x^2 L^2 \leq 1$, $\tau_x \leq \mu_y/(2L^2)$ and $\tau_y \leq 1/L$ by the choices of τ_x, τ_y, we can see that

$$\begin{aligned}
&\|x_{k+1} - x_k\|^2 + \|y_{k+1} - y_k\|^2 \\
&= \tau_x^2 \|\nabla_x f(x_k, y_k)\|^2 + \tau_y^2 \|\nabla_y f(x_k, y_k)\|^2 \\
&= \tau_x^2 \|\nabla_x f(x_k, y_k)\|^2 + \tau_y^2 \|\nabla_y f(x_k, y_k) - \nabla_y f(x_k, y^*(x_k)\|^2 \\
&\leq \tau_x^2 \|\nabla_x f(x_k, y_k)\|^2 + \|y_k - y^*(x_k)\|^2 \\
&\leq 2\tau_x^2 \|\nabla g(x_k)\|^2 + 2\tau_x^2 \|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2 + \|y_k - y^*(x_k)\|^2 \\
&\leq 2\|x_k - x^*\|^2 + 2\|y_k - y^*(x_k) - y_k\|^2 \\
&\leq \frac{4}{\mu_x}\mathcal{A}_k + \frac{4}{\mu_y}\mathcal{B}_k \\
&\leq \max\left\{\frac{4}{\mu_x}, \frac{88}{\mu_y}\right\} \mathcal{V}_k \\
&\leq \max\left\{\frac{4}{\mu_x}, \frac{88}{\mu_y}\right\} \left(1 - \frac{\mu_x \tau_x}{2}\right)^k \mathcal{V}_0,
\end{aligned} \tag{16}$$

where in the last inequality we use $\lambda\tau_x/\tau_y = 1/22$. Then we have

$$\|x_{k+1} - x_k\| + \|y_{k+1} - y_k\| \leq \left(1 - \frac{\mu_x\tau_x}{2}\right)^{k/2} \sqrt{2\max\left\{\frac{4}{\mu_x}, \frac{88}{\mu_y}\right\} \mathcal{V}_0}.$$

For $n \geq k$, we obtain

$$\begin{aligned}
\|x_n - x_k\| + \|y_n - y_k\| &\leq \sum_{i=k}^{n-1} \|x_{i+1} - x_i\|^2 + \|y_{i+1} - y_i\|^2 \\
&\leq \sqrt{2\max\left\{\frac{4}{\mu_x}, \frac{88}{\mu_y}\right\} \mathcal{V}_0} \sum_{i=k}^{\infty} \left(1 - \frac{\mu_x\tau_x}{2}\right)^{i/2} \\
&\leq \frac{c^{k/2}}{1-\sqrt{c}} \sqrt{2\max\left\{\frac{4}{\mu_x}, \frac{88}{\mu_y}\right\} \mathcal{V}_0},
\end{aligned}$$

where $c = 1 - \mu_x\tau_x/2$. We know that when $n \to \infty$, we have $(x_n, y_n) \to (x^*, y^*)$ where (x^*, y^*) is a saddle point, Taking square on both sides completes our proof.

□

B.2 Convergence under One-sided PL condition

When f is nonconvex in x, we have the following theorem for GDA.

Theorem B.2. *Suppose function $f(x,y)$ satisfies L-smooth, μ_y-PL in y. Let $\tau_y = 1/L$, $\lambda = 4L^2/\mu_y^2$ and $\tau_x = \tau_y/(18\lambda)$, then the sequence $\{(x_k, y_k)\}_{k=0}^{K-1}$ generated by Algorithm 4 satisfies,*

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla g(x_k)\|^2 \leq \frac{288L^3}{K\mu_y^2}\mathcal{V}_0.$$

Furthermore, if we choose the output $(\hat{x}, \hat{y})$ uniformly from $\{(x_k, y_k)\}_{k=0}^{K-1}$, then we can get $\|\nabla g(\hat{x})\| \leq \epsilon$ with no more than $\mathcal{O}(n\kappa_y^2 L/\epsilon^2)$ first-order oracle calls.

Proof. Using equation (8) and Lemma A.6 that $\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2 \leq 2L^2\mathcal{B}_k/\mu_y$,we have

$$\mathcal{A}_{k+1} \leq \mathcal{A}_k - \frac{\tau_x}{2}\|\nabla g(x_k)\|^2 + \frac{\tau_x L^2}{\mu_y}\mathcal{B}_k. \tag{17}$$

Further, using equation (12), we have

$$\begin{aligned}
\mathcal{B}_{k+1} &\leq \mathcal{B}_k + \frac{9\tau_x}{2}\|\nabla g(x_k)\|^2 + \frac{11\tau_x}{2}\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2 - \frac{\tau_y}{4}\|\nabla_y f(x_k, y_k)\|^2 \\
&\leq \mathcal{B}_k + \frac{9\tau_x}{2}\|\nabla g(x_k)\|^2 + \frac{11\tau_x L^2}{\mu_y}\mathcal{B}_k - \frac{\mu_y\tau_y}{2}\mathcal{B}_k \\
&\leq (1 - \frac{\mu_y\tau_y}{4})\mathcal{B}_k + \frac{9\tau_x}{2}\|\nabla g(x_k)\|^2,
\end{aligned} \tag{18}$$

where we use Lemma A.6 and PL condition in y in the first inequality and $11\tau_x L^2/\mu_y \leq \mu_y\tau_y/4$ by the choices of τ_x, τ_y. Thus,

$$\mathcal{B}_k \leq \left(1 - \frac{\mu_y\tau_y}{4}\right)^k \mathcal{B}_0 + \frac{9\tau_x}{2}\sum_{i=0}^{k-1}\left(1 - \frac{\mu_y\tau_y}{4}\right)^{k-i-1}\|\nabla g(x_i)\|^2.$$

Plugging into (17),

$$\mathcal{A}_{k+1} \leq \mathcal{A}_k - \frac{\tau_x}{2}\|\nabla g(x_k)\|^2 + \frac{\tau_x L^2}{\mu_y}\left(1 - \frac{\mu_y\tau_y}{4}\right)^k \mathcal{B}_0 + \frac{9\tau_x^2 L^2}{2\mu_y}\sum_{i=0}^{k-1}\left(1 - \frac{\mu_y\tau_y}{4}\right)^{k-i-1}\|\nabla g(x_i)\|^2.$$

Telescoping and noticing that $18\tau_x^2 L^2/\tau_y\mu_y^2 \leq \tau_x/4$ and $\lambda = 4L^2/\mu_y^2$, we have

$$\begin{aligned}
\mathcal{A}_{K+1} &\leq \mathcal{A}_0 - \frac{\tau_x}{2}\sum_{k=0}^{K}\|\nabla g(x_k)\|^2 + \frac{\tau_x L^2}{\mu_y}\sum_{k=0}^{K-1}\left(1 - \frac{\mu_y\tau_y}{4}\right)^k \mathcal{B}_0 \\
&\qquad + \frac{9\tau_x^2 L^2}{2\mu_y}\sum_{k=1}^{K}\sum_{i=0}^{k-1}\left(1 - \frac{\mu_y\tau_y}{4}\right)^{k-i-1}\|\nabla g(x_i)\|^2 \\
&= \mathcal{A}_0 - \frac{\tau_x}{2}\sum_{k=0}^{K}\|\nabla g(x_k)\|^2 + \frac{\tau_x L^2}{\mu_y}\sum_{k=0}^{K-1}\left(1 - \frac{\mu_y\tau_y}{4}\right)^k \mathcal{B}_0 \\
&\qquad + \frac{9\tau_x^2 L^2}{2\mu_y}\sum_{i=0}^{K-1}\sum_{k=i+1}^{K}\left(1 - \frac{\mu_y\tau_y}{4}\right)^{k-i-1}\|\nabla g(x_i)\|^2 \\
&\leq \mathcal{A}_0 - \frac{\tau_x}{2}\sum_{k=0}^{K}\|\nabla g(x_k)\|^2 + \frac{\tau_x L^2}{\mu_y}\sum_{k=0}^{K-1}\left(1 - \frac{\mu_y\tau_y}{4}\right)^k \mathcal{B}_0 + \frac{18\tau_x^2 L^2}{\tau_y\mu_y^2}\sum_{i=0}^{K-1}\|\nabla g(x_i)\|^2 \\
&\leq \mathcal{A}_0 - \frac{\tau_x}{2}\sum_{k=0}^{K}\|\nabla g(x_k)\|^2 + \frac{\tau_x L^2}{\mu_y}\sum_{k=0}^{K-1}\left(1 - \frac{\mu_y\tau_y}{4}\right)^k \mathcal{B}_0 + \frac{18\tau_x^2 L^2}{\tau_y\mu_y^2}\sum_{k=0}^{K}\|\nabla g(x_k)\|^2 \\
&\leq \mathcal{A}_0 - \frac{\tau_x}{4}\sum_{k=0}^{K}\|\nabla g(x_k)\|^2 + \frac{4\tau_x L^2}{\tau_y\mu_y^2}\mathcal{B}_0 \\
&= \mathcal{V}_0 - \frac{\tau_x}{4}\sum_{k=0}^{K}\|\nabla g(x_k)\|^2.
\end{aligned}$$

Rearranging and noticing that $\mathcal{A}_{K+1} \geq 0$, we can see that

$$\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla g(x_k)\|^2 \leq \frac{4\mathcal{V}_0}{(K+1)\tau_x},$$

which is equivalent to the desired inequality. $\square$

Algorithm 5 SVRG-GDA $(f, (x_0, y_0), T, S, M, B, \tau_x, \tau_y)$

```
x̄_0 = x_0, ȳ_0 = y_0
for t = 0, 1, ..., T − 1 do
    for s = 0, 1, ..., S − 1 do
        x_{s,0} = x̄_s, y_{s,0} = ȳ_s
        compute ∇_x f(x̄_s, ȳ_s) = (1/n) Σ_{i=1}^n ∇_x f_i(x̄_s, ȳ_s)
        compute ∇_y f(x̄_s, ȳ_s) = (1/n) Σ_{i=1}^n ∇_y f_i(x̄_s, ȳ_s)
        for k = 0, 1, ..., M − 1
            draw samples S_x, S_y independently with both size B.
            G_x(x_{s,k}, y_{s,k}) = (1/B) Σ_{i∈S_x} [∇_x f_i(x_{s,k}, y_{s,k}) − ∇_x f_i(x̄_s, ȳ_s) + ∇_x f(x̄_s, ȳ_s)]
            G_y(x_{s,k}, y_{s,k}) = (1/B) Σ_{i∈S_y} [∇_y f_i(x_{s,k}, y_{s,k}) − ∇_x f_i(x̄_s, ȳ_s) + ∇_x f(x̄_s, ȳ_s)]
            x_{s,k+1} = x_{s,k} − τ_x G_x(x_{s,k}, y_{s,k})
            y_{s,k+1} = y_{s,k} + τ_y G_y(x_{s,k}, y_{s,k})
        end for
        x̄_{s+1} = x_{s,M}, ȳ_{s+1} = y_{s,M}
    end for
    choose (x_t, y_t) from {{(x_{s,k}, y_{s,k})}_{k=0}^{M−1}}_{s=0}^{S−1} uniformly at random.
    x̄_0 = x_t, ȳ_0 = y_t
end for
return (x_T, y_T)
```

C Convergence of GDA with SVRG Gradient Estimators

In this section, we will show the convergence rate of GDA with SVRG gradient estimators is $\mathcal{O}((n + n^{2/3}\kappa_x\kappa_y^2)\log(1/\epsilon))$, which is faster than the result of $\mathcal{O}((n + n^{2/3}\max\{\kappa_x^3, \kappa_y^3\})\log(1/\epsilon))$ given by SVRG-AGDA Yang et al. [45]. Plus, we can prove that the same convergence rate can be achieved by AGDA with similar techniques since in our algorithm we set $\tau_x \ll \tau_y$, therefore x_k changes much slower than y_k. The reason for unbalanced step sizes lies in that $g(x)$ is $(L + L^2/\mu_y)$-smooth by Lemma A.8. Thus, we can regard that the condition number of solving problem $\max_{y\in\mathbb{R}^{d_y}} f(x, y)$ is $\mathcal{O}(\kappa_y)$ while that of solving $\min_{x\in\mathbb{R}^{d_x}} g(x)$ is $\mathcal{O}(\kappa_x\kappa_y)$. Thus, it is reasonable that the total complexity has a factor of $\mathcal{O}(\kappa_x\kappa_y^2)$. The algorithm is described in 5.

For the innermost loop about subscript k when t and s are both fixed, we define the Lyapunov function as follows:

$$\mathcal{V}_{s,k} = \mathcal{A}_{s,k} + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_{s,k} + c_{s,k}\|x_{s,k} - \bar{x}_s\|^2 + d_{s,k}\|y_k - \bar{y}_s\|^2,$$

where $\mathcal{A}_{s,k} = g(x_{s,k}) - g(x^*)$ and $\mathcal{B}_{s,k} = g(x_{s,k}) - f(x_{s,k}, y_{s,k})$ and $c_{s,k}, d_{s,k}$ will be defined recursively with $c_{s,M} = d_{s,M} = 0$ in our proof. Then we can have the following lemma.

Lemma C.1. *Under Assumption 6.1 and 3.1, if we let $\tau_y = \nu/(Ln^\alpha)$, $\lambda = 14L^2/\mu_y^2$ and $\tau_x = \tau_y/(22\lambda)$, where $\nu = 1/(176(\mathrm{e} - 1)), 0 < \alpha \le 1$; let $B = 1, M = \lfloor n^{3\alpha/2}/(2\nu)\rfloor$. Then for Algorithm 5, the following statement holds true:*

$$\mathbb{E}[\mathcal{V}_{s,k+1}] \le \mathcal{V}_{s,k} - \frac{\tau_x}{8}\|\nabla g(x_{s,k})\|^2 - \frac{\lambda\tau_x}{16}\|\nabla_y f(x_{s,k}, y_{s,k})\|^2,$$

where $\mathcal{A}_{s,k} = g(x_{s,k}) - g(x^*)$ and $\mathcal{B}_{s,k} = g(x_{s,k}) - f(x_{s,k}, y_{s,k})$. Above, the definitions of $c_{s,k}, d_{s,k}$ is given recursively with $c_{s,M} = d_{s,M} = 0$ as:

$$c_{s,k} = c_{s,k+1}(1 + \tau_x\gamma_1) + \left(c_{s,k+1}\tau_x^2 + \frac{3\tau_x^2L^2}{\mu_y}\right)L^2 + \left(d_{s,k+1}\tau_y^2 + \frac{\lambda\tau_x\tau_yL}{2}\right)L^2,$$

$$d_{s,k} = d_{s,k+1}(1+\tau_y\gamma_2) + \left(c_{s,k+1}\tau_x^2 + \frac{3\tau_x^2 L^2}{\mu_y}\right) L^2 + \left(d_{s,k+1}\tau_y^2 + \frac{\lambda\tau_x\tau_y L}{2}\right) L^2.$$

Proof. Since s is fixed in thie lemma, we omit subscripts of t in the following proofs, then the Lyapunov function can be written as:

$$\mathcal{V}_k = \mathcal{A}_k + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_k + c_k\|x_k - \bar{x}\|^2 + d_k\|y_k - \bar{y}\|^2.$$

Before the formal proof, we present some standard properties of variance reduction. We denote the stochastic gradients as:

$$G_x(x_k, y_k) = \frac{1}{B}\sum_{i\in S_x}\big(\nabla_x f_i(x_k, y_k) - \nabla_x f_i(\bar{x}, \bar{y}) + \nabla_x f(\bar{x}, \bar{y})\big)$$

and

$$G_y(x_k, y_k) = \frac{1}{B}\sum_{i\in S_y}\big(\nabla_y f_i(x_k, y_k) - \nabla_y f_i(\bar{x}, \bar{y}) + \nabla_y f(\bar{x}, \bar{y})\big).$$

Then we know that the stochastic gradients satisfy unbiasedness that

$$\mathbb{E}[G_x(x_k, y_k)] = \nabla_x f(x_k, y_k) \quad \text{and} \quad \mathbb{E}[G_y(x_k, y_k)] = \nabla_y f(x_k, y_k).$$

And we can bound the variance of the stochastic gradients as follows:

$$\begin{aligned}
&\mathbb{E}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2 \\
&= \mathbb{E}\|\nabla_x f_i(x_k, y_k) - \nabla_x f_i(\bar{x}, \bar{y}) + \nabla_x f(\bar{x}, \bar{y}) - \nabla_x f(x_k, y_k)\|^2 \\
&\le \mathbb{E}\|\nabla_x f_i(x_k, y_k) - \nabla_x f_i(\bar{x}, \bar{y})\|^2 \\
&\le L^2\mathbb{E}\|x_k - \bar{x}\|^2] + L^2\mathbb{E}[\|y_k - \bar{y}\|^2.
\end{aligned} \tag{19}$$

Similarly, we have

$$\mathbb{E}\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2 \le L^2\mathbb{E}\|x_k - \bar{x}\|^2 + L^2\mathbb{E}\|y_k - \bar{y}\|^2. \tag{20}$$

Equipped with the above properties of SVRG, now we can begin our proof of Lemma C.1.

Since we know that g is $(2L^2/\mu_y)$-smooth by Lemma A.8 and $\tau_x \le \mu_y/2L^2$, we have

$$\begin{aligned}
\mathbb{E}[g(x_{k+1})] &\le \mathbb{E}\left[g(x_k) + \nabla g(x_k)^\top(x_{k+1} - x_k) + \frac{2L^2}{\mu_y}\|x_{k+1} - x_k\|^2\right] \\
&\le \mathbb{E}\left[g(x_k) - \tau_x\nabla g(x_k)^\top G_x(x_k, y_k) + \frac{\tau_x^2 L^2}{\mu_y}\|G_x(x_k, y_k)\|^2\right] \\
&\le \mathbb{E}\left[g(x_k) - \tau_x\nabla g(x_k)^\top \nabla_x f(x_k, y_k) + \frac{\tau_x}{2}\|\nabla_x f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{\tau_x^2 L^2}{\mu_y}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&= \mathbb{E}\left[g(x_k) - \frac{\tau_x}{2}\|\nabla g(x_k)\|^2 + \frac{\tau_x}{2}\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{\tau_x^2 L^2}{\mu_y}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2\right],
\end{aligned} \tag{21}$$

where we use $\tau_x^2 L^2 \le \mu_y$ in the third inequality. Also, we have

$$\begin{aligned}
\mathbb{E}[f(x_k, y_k)] &\le \mathbb{E}\left[f(x_{k+1}, y_k) - \nabla_x f(x_k, y_k)^\top(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2\right] \\
&= \mathbb{E}\left[f(x_{k+1}, y_k) + \tau_x\nabla_x f(x_k, y_k)^\top G_x(x_k, y_k) + \frac{\tau_x^2 L}{2}\|G_x(x_k, y_k)\|^2\right] \\
&= \mathbb{E}\left[f(x_{k+1}, y_k) + \tau_x\|\nabla_x f(x_k, y_k)\|^2 + \frac{\tau_x^2 L}{2}\|\nabla_x f(x_k, y_k)\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}\left[\frac{\tau_x^2 L}{2}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&\le \mathbb{E}\left[f(x_{k+1}, y_k) + \frac{3\tau_x}{2}\|\nabla_x f(x_k, y_k)\|^2 + \frac{\tau_x^2 L}{2}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2\right],
\end{aligned}
$$

where we use the quadratic upper bound implied by L-smoothness in the first inequality and $\tau_y \le 1/L$ in the second one. Similarly,

$$
\begin{aligned}
\mathbb{E}[f(x_{k+1}, y_k)] &\le \mathbb{E}\left[f(x_{k+1}, y_{k+1}) - \nabla_y f(x_{k+1}, y_k)^\top (y_{k+1} - y_k) + \frac{L}{2}\|y_{k+1} - y_k\|^2\right] \\
&= \mathbb{E}\left[f(x_{k+1}, y_{k+1}) - \tau_y \nabla_y f(x_{k+1}, y_k)^\top G_y(x_k, y_k) + \frac{\tau_y^2 L}{2}\|G_y(x_k, y_k)\|^2\right] \\
&\le \mathbb{E}\left[f(x_{k+1}, y_{k+1}) - \tau_y \nabla_y f(x_{k+1}, y_k)^\top \nabla_y f(x_k, y_k) + \frac{\tau_y}{2}\|\nabla_y f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{\tau_y^2 L}{2}\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2\right] \\
&= \mathbb{E}\left[f(x_{k+1}, y_{k+1}) - \frac{\tau_y}{2}\|\nabla_y f(x_{k+1}, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{\tau_y}{2}\|\nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_k)\|^2 + \frac{\tau_y^2 L}{2}\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2\right] \\
&\le \mathbb{E}\left[f(x_{k+1}, y_{k+1}) - \frac{\tau_y}{4}\|\nabla_y f(x_{k+1}, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\tau_y\|\nabla_y f(x_{k+1}, y_k) - \nabla_y f(x_k, y_k)\|^2 + \frac{\tau_y^2 L}{2}\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2\right] \\
&\le \mathbb{E}\left[f(x_{k+1}, y_{k+1}) - \frac{\tau_y}{4}\|\nabla_y f(x_{k+1}, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\tau_y \tau_x^2 L^2\|G_x(x_k, y_k)\|^2 + \frac{\tau_y^2 L}{2}\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2\right] \\
&\le \mathbb{E}\left[f(x_{k+1}, y_{k+1}) - \frac{\tau_y}{4}\|\nabla_y f(x_{k+1}, y_k)\|^2 + \tau_x\|\nabla_x f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\tau_x^2 L\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2 + \frac{\tau_y^2 L}{2}\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2\right].
\end{aligned}
$$

Above, the first and fourth inequalities are both due to L-smoothness; the second one follows from $\tau_y \le 1/L$; the third one uses the fact that $-\mathbb{E}[\|a-b\|^2] \le -\frac{1}{2}\mathbb{E}[\|a\|^2] + \mathbb{E}[\|b\|^2]$; the last one relies on $\mathbb{E}[\|G_x(x_k, y_k)\|^2] = \mathbb{E}[\|\nabla_x f(x_k, y_k)\|^2 + \|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2]$ and the choices of τ_x, τ_y. Summing up the above two inequalities, we obtain

$$
\begin{aligned}
\mathbb{E}[f(x_k, y_k)] &\le \mathbb{E}\left[f(x_{k+1}, y_{k+1}) + \frac{5\tau_x}{2}\|\nabla_x f(x_k, y_k)\|^2 + \frac{3\tau_x^2 L}{2}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&\quad \mathbb{E}\left[-\frac{\tau_y}{4}\|\nabla_y f(x_k, y_k)\|^2 + \frac{\tau_y^2 L}{2}\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2\right].
\end{aligned}
$$

Combing with inequality (21), we can see that

$$
\begin{aligned}
\mathbb{E}[\mathcal{B}_{k+1}] &\le \mathbb{E}\left[\mathcal{B}_k - \frac{\tau_x}{2}\|\nabla g(x_k)\|^2 + \frac{\tau_x}{2}\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[-\frac{\tau_y}{4}\|\nabla_y f(x_k, y_k)\|^2 + \frac{5\tau_x}{2}\|\nabla_x f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{\tau_y^2 L}{2}\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2 + \frac{5\tau_x^2 L^2}{2\mu_y}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&\le \mathbb{E}\left[\mathcal{B}_k + \frac{9\tau_x}{2}\|\nabla g(x_k)\|^2 + \frac{11\tau_x}{2}\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[-\frac{\tau_y}{4}\|\nabla_y f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{\tau_y^2 L}{2}\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2 + \frac{5\tau_x^2 L^2}{2\mu_y}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2\right],
\end{aligned} \tag{22}
$$

where we use $\mathbb{E}\|\nabla_x f(x_k, y_k)\|^2 \le \mathbb{E}\|\nabla g(x_k)\|^2 + \mathbb{E}\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2$. Using Young's inequality as equation (37) and (38) in [45], we have

$$\begin{aligned}
\mathbb{E}\|x_{k+1} - \bar{x}\|^2 \le{}& \mathbb{E}\left[\tau_x^2 \|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&+ \mathbb{E}\left[(1 + \tau_x\gamma_1)\|x_k - \bar{x}\|^2 + \left(\tau_x^2 + \frac{\tau_x}{\gamma_1}\right)\|\nabla_x f(x_k, y_k)\|^2\right], \\
\mathbb{E}\|y_{k+1} - \bar{y}\|^2 \le{}& \mathbb{E}\left[\tau_y^2 \|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2\right] \\
&+ \mathbb{E}\left[(1 + \tau_y\gamma_2)\|y_k - \bar{y}\|^2 + \left(\tau_y^2 + \frac{\tau_y}{\gamma_2}\right)\|\nabla_y f(x_k, y_k)\|^2\right],
\end{aligned}$$

where γ_1, γ_2 are two positive constant in Young's inequality which will be chosen later.

Then, using equation (19), (20), (21), (22), we have

$$\begin{aligned}
\mathbb{E}[\mathcal{V}_{k+1}] &= \mathbb{E}\left[\mathcal{A}_{k+1} + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_{k+1} + c_{k+1}\|x_{k+1} - \bar{x}\|^2 + d_{k+1}\|y_{k+1} - \bar{y}\|^2\right] \\
&\le \mathcal{A}_k + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_k - \left(1 - \frac{9\lambda\tau_x}{\tau_y}\right)\frac{\tau_x}{2}\|\nabla g(x_k)\|^2 + \\
&\quad + \left(1 + \frac{11\lambda\tau_x}{\tau_y}\right)\frac{\tau_x}{2}\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2 \\
&\quad - \frac{\lambda\tau_x}{4}\|\nabla_y f(x_k, y_k)\|^2 + \frac{2\tau_x^2 L^2}{\mu_y}\left(1 + \frac{5\lambda\tau_x}{4\tau_y}\right)\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2 \\
&\quad + \frac{\lambda\tau_x\tau_y L}{2}\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2 \\
&\quad + \mathbb{E}[c_{k+1}\|x_{k+1} - \bar{x}\|^2 + d_{k+1}\|y_{k+1} - \bar{y}\|^2] \\
&= \mathcal{V}_k - \left(1 - \frac{9\lambda\tau_x}{\tau_y}\right)\frac{\tau_x}{2}\|\nabla g(x_k)\|^2 + \left(1 + \frac{11\lambda\tau_x}{\tau_y}\right)\frac{\tau_x}{2}\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2 \\
&\quad - \frac{\lambda\tau_x}{4}\|\nabla_y f(x_k, y_k)\|^2 + c_{k+1}\left(\tau_x^2 + \frac{\tau_x}{\gamma_1}\right)\|\nabla_x f(x_k, y_k)\|^2 \\
&\quad + d_{k+1}\left(\tau_y^2 + \frac{\tau_y}{\gamma_2}\right)\|\nabla_y f(x_k, y_k)\|^2 \\
&\le \mathcal{V}_k - \left(1 - \frac{9\lambda\tau_x}{\tau_y}\right)\frac{\tau_x}{2}\|\nabla g(x_k)\|^2 + \left(1 + \frac{11\lambda\tau_x}{\tau_y}\right)\frac{\tau_x}{2}\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2 \\
&\quad - \frac{\lambda\tau_x}{4}\|\nabla_y f(x_k, y_k)\|^2 + d_{k+1}\left(\tau_y^2 + \frac{\tau_y}{\gamma_2}\right)\|\nabla_y f(x_k, y_k)\|^2 \\
&\quad + 2c_{k+1}\left(\tau_x^2 + \frac{\tau_x}{\gamma_1}\right)\|\nabla g(x_k)\|^2 \\
&\quad + 2c_{k+1}\left(\tau_x^2 + \frac{\tau_x}{\gamma_1}\right)\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2 \\
&\le \mathcal{V}_k - \frac{\tau_x}{4}\|\nabla g(x_k)\|^2 + \frac{3\tau_x}{4}\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2 - \frac{\lambda\tau_x}{4}\|\nabla_y f(x_k, y_k)\|^2 \\
&\quad + d_{k+1}\left(\tau_y^2 + \frac{\tau_y}{\gamma_2}\right)\|\nabla_y f(x_k, y_k)\|^2 + 2c_{k+1}\left(\tau_x^2 + \frac{\tau_x}{\gamma_1}\right)\|\nabla g(x_k)\|^2 \\
&\quad + 2c_{k+1}\left(\tau_x^2 + \frac{\tau_x}{\gamma_1}\right)\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2,
\end{aligned} \tag{23}$$

where the second last inequality relies on

$$\|\nabla_x f(x_k, y_k)\|^2 \le 2\|\nabla g(x_k)\|^2 + 2\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2;$$

in the last inequality we use $11\lambda\tau_x/\tau_y \le 1/2$ by our choices of λ, τ_x, τ_y; the second equality is due to the definition of c_{k+1}, d_{k+1}.

Now we define $e_k = \max\{c_k, d_k\}$ and we bound e_k by letting $\gamma_1 = \lambda L/n^{\alpha/2}$ and $\gamma_2 = L/n^{\alpha/2}$. Then according to the definition of c_k, d_k given by our definition, we have

$$
\begin{aligned}
e_k &\leq (1 + \tau_y\gamma_2 + \tau_y^2 L^2)e_{k+1} + \frac{3\tau_x^2 L^4}{\mu_y} + \frac{\lambda\tau_x\tau_y L^3}{2} \\
&\leq (1 + \tau_y\gamma_2 + \tau_y^2 L^2)e_{k+1} + \tau_y^2 L^3 \\
&= \left(1 + \frac{\nu}{n^{3\alpha/2}} + \frac{\nu^2}{n^{2\alpha}}\right) + \frac{L\nu^2}{n^{2\alpha}} \\
&\leq \left(1 + \frac{2\nu}{n^{3\alpha/2}}\right) e_{k+1} + \frac{L\nu^2}{n^{2\alpha}},
\end{aligned}
$$

where we use $\tau_x \leq \tau_y$ and $\gamma_1\tau_x \leq \gamma_2\tau_y$ in the first; the second inequality is due to $\tau_x^2 L/\mu_y \leq \tau_y^2/6$ and $\lambda\tau_x \leq \tau_y/4$; we plug in $\tau_y = \nu/Ln^\alpha$ in the third line; we use $\nu \leq 1$ in the last inequality.

Since $M = \lfloor n^{3\alpha/2}/(2\nu)\rfloor$ and $c_M = d_M = 0$, if we define $\theta = 2\nu/n^{3\alpha/2}$, then

$$
e_0 \leq \frac{L\nu^2}{n^{2\alpha}} \frac{(1+\theta)^M - 1}{\theta} \leq \frac{L\nu(\mathrm{e}-1)}{2n^{\alpha/2}}. \tag{24}
$$

Since $e_{k+1} \leq e_k$, we know that $e_k \leq e_0$, then

$$
\begin{aligned}
d_{k+1}\left(\tau_y^2 + \frac{\tau_y}{\gamma_2}\right) &\leq e_0\left(\tau_y + \frac{1}{\gamma_2}\right)\tau_y \\
&\leq \frac{L\nu(\mathrm{e}-1)}{2n^{\alpha/2}}\left(\tau_y + \frac{1}{\gamma_2}\right)\tau_y \\
&= \frac{\nu(\mathrm{e}-1)}{2}\left(\frac{\nu}{n^{3\alpha/2}} + 1\right)\tau_y \\
&\leq \nu(\mathrm{e}-1)\mathrm{\tau_y},
\end{aligned} \tag{25}
$$

where we use $d_{k+1} \leq e_0$ in the first inequality and (24) in the second one, and in the third line we plug in $\tau_y = \nu/(Ln^\alpha)$ and $\gamma_2 = L/n^{\alpha/2}$. The last inequality follows from $\nu/n^{3\alpha/2} \leq 1$.

Similarly, note that $\gamma_1 \geq \gamma_2$ and $\tau_x \leq \tau_y$ by the choices of τ_x, τ_y, then we have

$$
c_{k+1}\left(\tau_x^2 + \frac{\tau_x}{\gamma_1}\right) \leq e_0\left(\tau_x + \frac{1}{\gamma_1}\right)\tau_x \leq e_0\left(\tau_y + \frac{1}{\gamma_2}\right)\tau_x \leq 3\nu(\mathrm{e}-1)\tau_x. \tag{26}
$$

Plugging (25) and (26) into (23),

$$
\begin{aligned}
\mathbb{E}[\mathcal{V}_{k+1}] \leq \mathcal{V}_k &- \frac{\tau_x}{4}\|\nabla g(x_k)\|^2 + \frac{3\tau_x}{4}\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2 - \frac{\lambda\tau_x}{4}\|\nabla_y f(x_k, y_k)\|^2 \\
&+ \nu(\mathrm{e}-1)\tau_y\|\nabla_y f(x_k, y_k)\|^2 + 2\nu(\mathrm{e}-1)\mathrm{\tau_x}\|\mathrm{\nabla g(x_k)}\|^2 \\
&+ 2\nu(\mathrm{e}-1)\tau_x\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2.
\end{aligned}
$$

If we let $\nu \leq 1/(176(\mathrm{e}-1))$, then we can verify that the following statements hold true:

$$
\begin{aligned}
\nu(\mathrm{e}-1)\tau_y &\leq \frac{\lambda\tau_x}{8}, \\
2\nu(\mathrm{e}-1)\tau_x &\leq \frac{\tau_x}{8}.
\end{aligned}
$$

Thus,

$$
\mathbb{E}[\mathcal{V}_{k+1}] \leq \mathcal{V}_k - \frac{\tau_x}{8}\|\nabla g(x_k)\|^2 + \frac{7\tau_x}{8}\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2 - \frac{\lambda\tau_x}{8}\|\nabla_y f(x_k, y_k)\|^2.
$$

Using Lemma A.7 and μ_y-PL condition in y and plugging in $\lambda = 14L^2/\mu_y^2$ yields

$$
\frac{7\tau_x}{8}\|\nabla_x f(x_k, y_k) - \nabla g(x_k)\|^2 \leq \frac{\lambda\tau_x}{16}\|\nabla_y f(x_k, y_k)\|^2.
$$

Thus,

$$
\mathbb{E}[\mathcal{V}_{k+1}] \leq \mathcal{V}_k - \frac{\tau_x}{8}\|\nabla g(x_k)\|^2 - \frac{\lambda\tau_x}{16}\|\nabla_y f(x_k, y_k)\|^2.
$$

□

Now it is sufficient to show the convergence of SVRG-GDA.

Theorem C.1. *Under Assumption 3.2 and 3.1, if we let* $SM = \lceil 8/(\mu_x\tau_x)\rceil$, $T = \lceil \log(1/\epsilon)\rceil$ *and* M, B, τ_x, τ_y *defined in Lemma C.1, then the following statement holds true for Algorithm 5:*

$$\mathbb{E}\left[\tilde{\mathcal{A}}_{t+1} + \frac{\lambda\tau_x}{\tau_y}\tilde{\mathcal{B}}_{t+1}\right] \le \frac{1}{2}\left(\tilde{\mathcal{A}}_t + \frac{\lambda\tau_x}{\tau_y}\tilde{\mathcal{B}}_t\right),$$

where $\tilde{\mathcal{A}}_t = g(x_t) - g(x^*)$ *and* $\tilde{\mathcal{B}}_t = g(x_t) - f(x_t, y_t)$*. Furthermore, let* $\alpha = 2/3$*, then it requires* $\mathcal{O}((n + n^{2/3}\kappa_x\kappa_y^2)\log(1/\epsilon))$ *stochastic first-order calls to achieve* $g(x_T) - g(x^*) \le \epsilon$ *in expectation.*

Proof. By Lemma C.1 and Lemma A.5 that g satisfies μ_x-PL in x and Assumption 3.2 that function f satisfies μ_y-PL in y, we have

$$\mathbb{E}[\mathcal{V}_{s,k+1}] \le \mathcal{V}_{s,k} - \frac{\mu_x\tau_x}{4}\mathcal{A}_{s,k} - \frac{\mu_y\tau_y}{8}\frac{\lambda\tau_x}{\tau_y}\mathcal{B}_{s,k} \le \mathcal{V}_{s,k} - \frac{\mu_x\tau_x}{4}\mathcal{V}_{s,k},$$

where in the last inequality we use $\mu_x\tau_x/4 \le \mu_y\tau_y/8$. Telescoping for $k = 0, 1, \ldots, M-1$ and $s = 0, 1, \ldots, S-1$ and rearranging, we can see that in round t, it holds that

$$\frac{1}{SM}\sum_{s=0}^{S-1}\sum_{k=0}^{M-1}\mathcal{V}_{s,k} \le \frac{4}{\mu_x\tau_x SM}(\mathcal{V}_{0,0} - \mathcal{V}_{S,M}) \le \frac{1}{2}\mathcal{V}_{0,0},$$

where the last inequality is due to the choice of S.

The above inequality is exactly equivalent to what we want to prove:

$$\mathbb{E}\left[\tilde{\mathcal{A}}_{t+1} + \frac{\lambda\tau_x}{\tau_y}\tilde{\mathcal{B}}_{t+1}\right] \le \frac{1}{2}\left(\tilde{\mathcal{A}}_t + \frac{\lambda\tau_x}{\tau_y}\tilde{\mathcal{B}}_t\right).$$

Note that we have $M = \mathcal{O}(n^{3\alpha/2})$ and $S = \mathcal{O}(\kappa_x\kappa_y^2/n^{\alpha/2})$, then the complexity is

$$\mathcal{O}\left((n + SM + Sn)\log\left(\frac{1}{\epsilon}\right)\right) = \mathcal{O}\left((n + (n^{\alpha} + n^{1-\alpha/2})\kappa_x\kappa_y^2)\log\left(\frac{1}{\epsilon}\right)\right).$$

Plugging in $\alpha = 2/3$ yields the desired complexity and it can also be seen that it is also the best choice of α.

□

D Proof of Section 4

In this section, we show that why SPIDER type stochastic gradient estimators outperforms SVRG with complete proofs. Using recursive updates by SPIDER, the variance of gradients are under control.

Lemma D.1 (Fang et al. [12, Modified form Lemma 1])**.** *In Algorithm 1, it holds true that*

$$\mathbb{E}[\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2] \le \frac{L^2}{B}\sum_{j=(n_k-1)M}^{k}\left(\tau_x^2\mathbb{E}[\|G_x(x_j, y_j)\|^2] + \tau_y^2\mathbb{E}[\|G_y(x_j, y_j)\|^2]\right),$$

$$\mathbb{E}[\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2] \le \frac{L^2}{B}\sum_{j=(n_k-1)M}^{k}\left(\tau_x^2\mathbb{E}[\|G_x(x_j, y_j)\|^2] + \tau_y^2\mathbb{E}[\|G_y(x_j, y_j)\|^2]\right),$$

where $n_k = \lceil k/M\rceil$ *and* $(n_k - 1)M \le k \le n_k M - 1$.

The following lemma describes the main convergence property of SPIDER-GDA.

Lemma D.2. *Under Assumption 6.1 and 3.1, setting all the parameters as defined in Theorem 4.1, then it holds true that*

$$\mathbb{E}[\|\nabla_x f(\tilde{x}_{t+1}, \tilde{y}_{t+1})\|^2 + \|\lambda\nabla_y f(\tilde{x}_{t+1}, \tilde{y}_{t+1})\|^2] \le \frac{64}{\tau_x K}\mathbb{E}\left[\tilde{\mathcal{A}}_t + \frac{\lambda\tau_x}{\tau_y}\tilde{\mathcal{B}}_t\right],$$

where $\tilde{\mathcal{A}}_t = g(\tilde{x}_t) - g(x^*)$ *and* $\tilde{\mathcal{B}}_t = g(\tilde{x}_t) - f(\tilde{x}_t, \tilde{y}_t)$.

Proof. First of all, we fix t and analyze the inner loop. We define the Lyapunov function as:

$$\mathcal{V}_{t,k} = \mathcal{A}_{t,k} + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_{t,k},$$

where $\mathcal{A}_{t,k} = g(x_{t,k}) - g(x^*)$ and $\mathcal{B}_{t,k} = g(x_{t,k}) - f(x_{t,k}, y_{t,k})$.

For simplification, we omit the subscripts t when there is no ambiguity. Note that $g(x)$ is $(2L^2/\mu_y)$-smooth, we have

$$\begin{aligned}
&\mathbb{E}[g(x_{k+1}) - g(x^*)] \\
&\leq \mathbb{E}\left[g(x_k) - g(x^*) + \nabla g(x_k)^\top (x_{k+1} - x_k) + \frac{L^2}{\mu_y}\|x_{k+1} - x_k\|^2\right] \\
&= \mathbb{E}\left[g(x_k) - g(x^*) - \tau_x \nabla g(x_k)^\top G_x(x_k, y_k) + \frac{L^2\tau_x^2}{\mu_y}\|G_x(x_k, y_k)\|^2\right] \\
&= \mathbb{E}\left[g(x_k) - g(x^*) - \tau_x(\nabla g(x_k) - G_x(x_k, y_k))^\top G_x(x_k, y_k) + \left(\frac{L^2\tau_x^2}{\mu_y} - \tau_x\right)\|G_x(x_k, y_k)\|^2\right] \\
&\leq \mathbb{E}\left[g(x_k) - g(x^*) + \frac{\tau_x}{2}\|\nabla g(x_k) - G_x(x_k, y_k)\|^2 + \left(\frac{L^2\tau_x^2}{\mu_y} - \frac{\tau_x}{2}\right)\|G_x(x_k, y_k)\|^2\right] \\
&\leq \mathbb{E}\left[g(x_k) - g(x^*) + \tau_x\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\tau_x\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2 - \frac{\tau_x}{4}\|G_x(x_k, y_k)\|^2\right],
\end{aligned} \tag{27}$$

where the second inequality follows from the fact that $\mathbb{E}[a^\top b] \leq \frac{1}{2}\mathbb{E}[\|a\|^2] + \frac{1}{2}\|b\|^2$; the third inequality is because we have $\tau_x \leq \mu_y/(4L^2)$. Similarly, we can show that

$$\begin{aligned}
&\mathbb{E}[f(x_k, y_k) - f(x_{k+1}, y_k)] \\
&\leq \mathbb{E}\left[-\nabla_x f(x_k, y_k)^\top (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2\right] \\
&= \mathbb{E}\left[\tau_x \nabla_x f(x_k, y_k)^\top G_x(x_k, y_k) + \frac{\tau_x^2 L}{2}\|G_x(x_k, y_k)\|^2\right] \\
&= \mathbb{E}\left[\tau_x(\nabla_x f(x_k, y_k) - G_x(x_k, y_k))^\top G_x(x_k, y_k) + \left(\frac{\tau_x^2 L}{2} + \tau_x\right)\|G_x(x_k, y_k)\|^2\right] \\
&\leq \mathbb{E}\left[\frac{\tau_x}{2}\|\nabla_x f(x_k, y_k) - G_x(x_k, y_k)\|^2 + 2\tau_x\|G_x(x_k, y_k)\|^2\right],
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[f(x_{k+1}, y_k) - f(x_{k+1}, y_{k+1})] \\
&\leq \mathbb{E}\left[-\nabla_y f(x_{k+1}, y_k)^\top (y_{k+1} - y_k) + \frac{L}{2}\|y_{k+1} - y_k\|^2\right] \\
&= \mathbb{E}\left[-\tau_y \nabla_y f(x_{k+1}, y_k)^\top G_y(x_k, y_k) + \frac{\tau_y^2 L}{2}\|G_y(x_k, y_k)\|^2\right] \\
&= \mathbb{E}\left[-\tau_y(\nabla_y f(x_{k+1}, y_k) - G_y(x_k, y_k))^\top G_y(x_k, y_k) + \left(\frac{\tau_y^2 L}{2} - \tau_y\right)\|G_y(x_k, y_k)\|^2\right] \\
&\leq \mathbb{E}\left[\frac{\tau_y}{2}\|\nabla_y f(x_{k+1}, y_k) - G_y(x_k, y_k)\|^2 + \left(\frac{\tau_y^2 L}{2} - \frac{\tau_y}{2}\right)\|G_y(x_k, y_k)\|^2\right] \\
&\leq \mathbb{E}\left[\tau_y\|\nabla_y f(x_k, y_k) - G_y(x_k, y_k)\|^2 + \tau_y\|\nabla_y f(x_k, y_k) - \nabla_y f(x_{k+1}, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\left(\frac{\tau_y^2 L}{2} - \frac{\tau_y}{2}\right)\|G_y(x_k, y_k)\|^2\right] \\
&\leq \mathbb{E}\left[\tau_y\|\nabla_y f(x_k, y_k) - G_y(x_k, y_k)\|^2 + \tau_y\tau_x^2 L^2\|G_x(x_k, y_k)\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}\left[\left(\frac{\tau_y^2 L}{2} - \frac{\tau_y}{2}\right)\|G_y(x_k, y_k)\|^2\right] \\
&\leq \mathbb{E}\left[\tau_y\|\nabla_y f(x_k, y_k) - G_y(x_k, y_k)\|^2 + \tau_x\|G_x(x_k, y_k)\|^2 - \frac{\tau_y}{4}\|G_y(x_k, y_k)\|^2\right],
\end{aligned}
$$

where we use $\tau_y \leq 1/(2L)$ and $\tau_x \leq 1/L$ and Young's inequality.

Summing up the above two inequalities, we have

$$
\begin{aligned}
\mathbb{E}[f(x_k, y_k) - f(x_{k+1}, y_{k+1})] \leq\ & \mathbb{E}\left[\frac{\tau_x}{2}\|\nabla_x f(x_k, y_k) - G_x(x_k, y_k)\|^2 + 3\tau_x\|G_x(x_k, y_k)\|^2\right] \\
&+ \mathbb{E}\left[\tau_y\|\nabla_y f(x_k, y_k) - G_y(x_k, y_k)\|^2 - \frac{\tau_y}{4}\|G_y(x_k, y_k)\|^2\right].
\end{aligned}
$$

Combing with inequality (27), it can be seen that

$$
\begin{aligned}
\mathbb{E}[\mathcal{B}_{k+1}] &= \mathbb{E}[g(x_{k+1}) - f(x_{k+1}, y_{k+1})] \\
&\leq \mathbb{E}[g(x_{k+1}) - g(x_k) + g(x_k) - f(x_k, y_k) + f(x_k, y_k) - f(x_{k+1}, y_{k+1})] \\
&\leq \mathbb{E}\left[\mathcal{B}_k + \tau_x\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2 + \frac{3\tau_x}{2}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\tau_y\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2 - \frac{\tau_y}{4}\|G_y(x_k, y_k)\|^2 + 3\tau_x\|G_x(x_k, y_k)\|^2\right].
\end{aligned}
$$

Therefore, using $24\lambda\tau_x \leq \tau_y$ and inequality (27) again, we obtain

$$
\begin{aligned}
\mathbb{E}[\mathcal{V}_{k+1}] &= \mathbb{E}\left[\mathcal{A}_{k+1} + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_{k+1}\right] \\
&\leq \mathbb{E}\left[\mathcal{A}_k + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_k + \left(\tau_x + \frac{\lambda\tau_x}{\tau_y}\right)\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\left(\tau_x + \frac{3\lambda\tau_x^2}{2\tau_y}\right)\|G_x(x_k, y_k - \nabla_x f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\left(\frac{3\lambda\tau_x^2}{\tau_y} - \frac{\tau_x}{4}\right)\|G_x(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\lambda\tau_x\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2 - \frac{\lambda\tau_x}{4}\|G_y(x_k, y_k)\right] \\
&\leq \mathbb{E}\left[\mathcal{A}_k + \frac{\lambda\tau_x}{\tau_y}\mathcal{B}_k + 2\tau_x\|\nabla g(x_k) - \nabla_x f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{5\tau_x}{2}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2 - \frac{\tau_x}{8}\|G_x(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\lambda\tau_x\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2 - \frac{\lambda\tau_x}{4}\|G_y(x_k, y_k)\|^2\right].
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
\mathbb{E}[\mathcal{V}_{k+1}] &\leq \mathbb{E}\left[\mathcal{V}_k + \frac{2L^2\tau_x}{\mu_y^2}\|\nabla_y f(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{5\tau_x}{2}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2 - \frac{\tau_x}{8}\|G_x(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\lambda\tau_x\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2 - \frac{\lambda\tau_x}{4}\|G_y(x_k, y_k)\|^2\right] \\
&\leq \mathbb{E}\left[\mathcal{V}_k + \frac{5\tau_x}{2}\|G_x(x_k, y_k) - \nabla_x f(x_k, y_k)\|^2 - \frac{\tau_x}{8}\|G_x(x_k, y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{9\lambda\tau_x}{8}\|G_y(x_k, y_k) - \nabla_y f(x_k, y_k)\|^2 - \frac{\lambda\tau_x}{8}\|G_y(x_k, y_k)\|^2\right].
\end{aligned}
$$

Above, the first inequality follows from Lemma A.7 and the second inequality uses Young's inequality that $\mathbb{E}[\|a-b\|^2] \le \mathbb{E}[\|a\|^2 + \|b\|^2]$ and $\lambda = 32L^2/\mu_y^2$.

Now, plug in the variance bound of Spider given by Lemma D.1 and $B = M$, we have

$$
\begin{aligned}
\mathbb{E}[\mathcal{V}_{k+1}] &\le \mathbb{E}\left[\mathcal{V}_k + \left(\frac{5}{2} + \frac{9\lambda}{8}\right)\frac{\tau_x^3 L^2}{M} \sum_{j=(n_k-1)M}^{k} \|G_x(x_j,y_j)\|^2 - \frac{\tau_x}{8}\|G_x(x_k,y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\left(\frac{5}{2} + \frac{9\lambda}{8}\right)\frac{\tau_x \tau_y^2 L^2}{M} \sum_{j=(n_k-1)M}^{k} \|G_y(x_j,y_j)\|^2 - \frac{\lambda\tau_x}{8}\|G_y(x_k,y_k)\|^2\right] \\
&\le \mathbb{E}\left[\mathcal{V}_k + \frac{5\lambda\tau_x^3 L^2}{4M} \sum_{j=(n_k-1)M}^{k} \|G_x(x_j,y_j)\|^2 - \frac{\tau_x}{8}\|G_x(x_k,y_k)\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{5\lambda\tau_x \tau_y^2 L^2}{4M} \sum_{j=(n_k-1)M}^{k} \|G_y(x_j,y_j)\|^2 - \frac{\lambda\tau_x}{8}\|G_y(x_k,y_k)\|^2\right].
\end{aligned}
$$

Now we telescope for $i = (n_k-1)M, \cdots, k$.

$$
\begin{aligned}
\mathbb{E}[\mathcal{V}_{k+1}] &\le \mathbb{E}\left[\mathcal{V}_{(n_k-1)M}\right] \\
&\quad + \mathbb{E}\left[\sum_{i=(n_k-1)M}^{k}\left(\sum_{j=(n_k-1)M}^{i} \frac{5\lambda\tau_x^3 L^2}{4M}\|G_x(x_j,y_j)\|^2 - \frac{\tau_x}{8}\|G_x(x_i,y_i)\|^2\right)\right] \\
&\quad + \mathbb{E}\left[\sum_{i=(n_k-1)M}^{k}\left(\sum_{j=(n_k-1)M}^{i} \frac{5\lambda\tau_x\tau_y^2 L^2}{4M}\|G_y(x_j,y_j)\|^2 - \frac{\lambda\tau_x}{8}\|G_y(x_i,y_i)\|^2\right)\right] \\
&\le \mathbb{E}\left[\mathcal{V}_{(n_k-1)M} + \sum_{j=(n_k-1)M}^{k}\left(\frac{5\lambda\tau_x^3 L^2}{4}\|G_x(x_j,y_j)\|^2 - \frac{\tau_x}{8}\|G_x(x_j,y_j)\|^2\right)\right] \\
&\quad + \mathbb{E}\left[\sum_{j=(n_k-1)M}^{k}\left(\frac{5\lambda\tau_x\tau_y^2 L^2}{4}\|G_y(x_j,y_j)\|^2 - \frac{\lambda\tau_x}{8}\|G_y(x_j,y_j)\|^2\right)\right] \\
&\le \mathbb{E}\left[\mathcal{V}_{(n_k-1)M} - \frac{\tau_x}{16}\sum_{j=(n_k-1)M}^{k}\|G_x(x_j,y_j)\|^2 - \frac{\lambda\tau_x}{16}\sum_{j=(n_k-1)M}^{k}\|G_y(x_j,y_j)\|^2\right],
\end{aligned}
$$

where we use $\lambda\tau_x^2 L^2 \le 1/20$ and $\tau_y^2 L^2 \le 1/20$ in the last inequality.

From now on, we need to write down the subscripts with respect to t. Telescope for $k = 0, \cdots, K-1$.

$$
\mathbb{E}[\mathcal{V}_{t,K}] \le \mathbb{E}\left[\mathcal{V}_{t,0} - \frac{\tau_x}{16}\sum_{k=0}^{K-1}\|G_x(x_{t,k},y_{t,k})\|^2 - \frac{\lambda\tau_x}{16}\sum_{k=0}^{K-1}\|G_y(x_{t,k},y_{t,k})\|^2\right].
$$

Noting that we choose $(\tilde{x}_{t+1},\tilde{y}_{t+1})$ from $\{(x_{t,k},y_{t,k})\}_{k=0}^{K-1}$ uniformly at random, we have

$$
\mathbb{E}\left[\frac{\tau_x}{16}\|G_x(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2 + \frac{\lambda\tau_x}{16}\|G_y(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2\right] \le \frac{1}{K}\mathbb{E}\left[\tilde{\mathcal{A}}_t + \frac{\lambda\tau_x}{\tau_y}\tilde{\mathcal{B}}_t\right]. \tag{28}
$$

Additionally, denote random variable ξ_t as the index from $k = 0, 1, \cdots, K-1$ that is chosen as $(\tilde{x}_{t+1},\tilde{y}_{t+1})$, it holds that

$$
\begin{aligned}
&\mathbb{E}[\|G_x(\tilde{x}_{t+1},\tilde{y}_{t+1}) - \nabla_x f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2] \\
&\le \mathbb{E}\left[\frac{L^2\tau_x^2}{B}\sum_{j=(n_{\xi_t}-1)M}^{\xi_t}\|G_x(x_j,y_j)\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{L^2\tau_x^2 M}{TB}\sum_{j=0}^{K-1}\mathbb{E}\|G_x(x_j,y_j)\|^2 \\
&= \frac{L^2\tau_x^2}{K}\sum_{j=0}^{K-1}\mathbb{E}\|G_x(x_j,y_j)\|^2 \\
&= L^2\tau_x^2\mathbb{E}\|G_x(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2,
\end{aligned}
$$

where we use Lemma D.1 again in the first inequality; the second inequality holds because the probability that $n_{\xi_t} = 1, 2, \cdots, n_K$ is less than or equal to M/K. Similarly,

$$
\mathbb{E}\|G_x(\tilde{x}_{t+1},\tilde{y}_{t+1}) - \nabla_x f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2 \leq L^2\tau_x^2\mathbb{E}\|G_x(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2.
$$

Now it is sufficient to show that

$$
\begin{aligned}
&\mathbb{E}[\|\nabla_x f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2 + \lambda\|\nabla_y f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2] \\
&\leq \mathbb{E}[2\|\nabla_x f(\tilde{x}_{t+1},\tilde{y}_{t+1}) - G_x(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2 + 2\|G_x(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2] \\
&\quad + \mathbb{E}[2\|\nabla_y f(\tilde{x}_{t+1},\tilde{y}_{t+1}) - G_y(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2 + 2\|G_y(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2] \\
&\leq 2(1+L^2\tau_x^2)\mathbb{E}[\|G_x(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2] + 2\lambda(1+L^2\tau_y^2)\mathbb{E}[\|G_y(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2] \\
&\leq 4\mathbb{E}\|G_x(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2 + 4\lambda\mathbb{E}\|G_y(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2 \\
&\leq \frac{64}{\tau_x K}\mathbb{E}\left[\tilde{\mathcal{A}}_t + \frac{\lambda\tau_x}{\tau_y}\tilde{\mathcal{B}}_t\right],
\end{aligned}
$$

where we use inequality (28) in the last line.

□

Equipped with the above lemma, we can easily prove Theorem 4.1.

D.1 Proof of Theorem 4.1

Proof. Using the properties of PL condition, it holds true that

$$
\begin{aligned}
&\mathbb{E}\left[\tilde{\mathcal{A}}_{t+1} + \frac{\lambda\tau_x}{\tau_y}\tilde{\mathcal{B}}_{t+1}\right] \\
&\leq \mathbb{E}\left[\frac{1}{2\mu_x}\|\nabla g(\tilde{x}_{t+1})\|^2 + \frac{\lambda\tau_x}{2\mu_y\tau_y}\|\nabla_y f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{2\mu_x}\|\nabla g(\tilde{x}_{t+1})\|^2 + \frac{1}{48\mu_y}\|\nabla_y f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{\mu_x}\|\nabla_x f(\tilde{x}_{t+1},y_{t+1})\|^2 + \frac{1}{\mu_x}\|\nabla_x f(\tilde{x}_{t+1},\tilde{y}_{t+1}) - \nabla g(\tilde{x}_{t+1})\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{1}{48\mu_y}\|\nabla_y f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{\mu_x}\|\nabla_x f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2 + \left(\frac{1}{48\mu_y} + \frac{L^2}{\mu_x\mu_y^2}\right)\|\nabla_y f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{\mu_x}\|\nabla_x f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2 + \frac{33L^2}{32\mu_x\mu_y^2}\|\nabla_y f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2\right] \\
&\leq \frac{33}{\mu_x}\mathbb{E}\left[\|\nabla_x f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2 + \lambda\|\nabla_y f(\tilde{x}_{t+1},\tilde{y}_{t+1})\|^2\right] \\
&\leq \frac{2112}{\mu_x\tau_x K}\mathbb{E}\left[\tilde{\mathcal{A}}_t + \frac{\lambda\tau_x}{\tau_y}\tilde{\mathcal{B}}_t\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[\tilde{\mathcal{A}}_t + \frac{\lambda\tau_x}{\tau_y}\tilde{\mathcal{B}}_t\right].
\end{aligned}
$$

Above, in the first inequality we use the definition of PL condition; in the the second we plug in $\lambda\tau_x/\tau_y = 1/24$; the third one is due to Young's inequality; the fourth one follows from Lemma A.7; the fifth and sixth ones are both trivial; in the second last one we use Lemma D.2 and in the last one we plug in our choice of K. Therefore, to find $\hat{x}$ such that $g(\hat{x}) - g(x^*) \leq \epsilon$ and $g(\hat{x}) - f(\hat{x}, \hat{y}) \leq 24\epsilon$ in expectation, the complexity is

$$\mathcal{O}\left((n + MK)\log\left(\frac{1}{\epsilon}\right)\right) = \mathcal{O}\left((n + K\sqrt{n})\log\left(\frac{1}{\epsilon}\right)\right) = \mathcal{O}\left((n + \sqrt{n}\kappa_x\kappa_y^2)\log\left(\frac{1}{\epsilon}\right)\right)$$

□

E Proof of Section 5

In this section, we present the convergence results of AccSPIDER-GDA when $\gamma = 0$, now F_k can be written as:

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} F_k(x, y) \triangleq f(x, y) + \frac{\beta}{2}\|x - x_k\|^2.$$

First of all, we take a closer look at F_k. The regularization term β transforms the condition number of the problem.

Lemma E.1. *Given $\beta > L$, the sub-problem $F_k(x, y)$ is $(\beta - L)$-PL in x , μ_y-PL in y and $(\beta + L)$-smooth if f satisfies L-smmoth and μ_x-PL in x.*

Strong duality also holds for sub-problem $F_k(x, y)$ since its saddle point exist.

Lemma E.2. *Under Assumption 6.1 and 5.1, given $\beta > L$, the sub-problem $F_k(x, y)$ has a unique saddle point.*

Proof. Denote $G_k(x) \triangleq \max_{y \in \mathbb{R}^{d_y}} F_k(x, y)$. According to Assumption 5.1, the inner problem $\max_{y \in \mathbb{R}^{d_y}} F_k(x, y)$ has a unique solution $y^*(x)$.

Additionally, it is clear that $F_k(x, y)$ is strongly convex in x for $\beta > L$. Hence, we know that $G_k(x)$ is strongly convex since taking the supremum is an operation that preserve (strong) convexity and . In this case, the outer problem $\min_{x \in \mathbb{R}^{d_x}} G_k(x)$ also has a unique solution x^*.

Above, the point (x^*, y^*) is a unique global minimax point of $F_k(x, y)$. And the global minimax point of $F_k(x, y)$ is equivalent to a saddle point of $F_k(x, y)$ by Lemma A.1. □

From now on, throughout this section, we always $(\tilde{x}_k, \tilde{y}_k)$ be the saddle point of F_{k-1} for all $k \geq 1$.

Next, we study the error brought by the inexact solution to the sub-problem. The idea is that when we can controls the precision of F_k with a global constant δ, then the algorithm will be close to the exact proximal point algorithm. We omit the notation of expectation when no ambiguity arises.

Lemma E.3. *Suppose x_{k+1} satisfies $\mathbb{E}[\|x_{k+1} - \tilde{x}_{k+1}\|^2] \leq \delta$ for any saddle point $(\tilde{x}_{k+1}, \tilde{y}_{k+1})$ of F_k, then it holds that*

$$\mathbb{E}\left[F_k(x_{k+1}, \tilde{y}_{k+1}) - F_k(\tilde{x}_{k+1}, \tilde{y}_{k+1})\right] \leq \frac{(\beta + L)^2\delta}{2(\beta - L)}.$$

Proof. Lemma A.2 tells us that $F_k(\tilde{x}_{k+1}, \tilde{y}_{k+1}) = \min_{x \in \mathbb{R}^{d_x}} F_k(x, \tilde{y}_{k+1})$, thus, we have

$$\begin{aligned}
&\mathbb{E}[F_k(x_{k+1}, \tilde{y}_{k+1}) - F_k(\tilde{x}_{k+1}, \tilde{y}_{k+1})] \\
&= \mathbb{E}[F_k(x_{k+1}, \tilde{y}_{k+1}) - \min_{x \in \mathbb{R}^{d_x}} F_k(x, \tilde{y}_{k+1})] \\
&\leq \frac{1}{2(\beta - L)}\mathbb{E}[\|\nabla_x F_k(x_{k+1}, \tilde{y}_{k+1})\|^2] \\
&= \frac{1}{2(\beta - L)}\mathbb{E}[\|\nabla_x F_k(x_{k+1}, \tilde{y}_{k+1}) - \nabla_x F_k(\tilde{x}_{k+1}, \tilde{y}_{k+1})\|^2]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{(\beta+L)^2}{2(\beta-L)}\mathbb{E}[\|x_{k+1}-\tilde{x}_{k+1}\|^2] \\
&\leq \frac{(\beta+L)^2\delta}{2(\beta-L)},
\end{aligned}
$$

where the first and third inequalities rely on Lemma E.1 that $F_k(x,y)$ is $(\beta-L)$-PL in x and $(\beta+L)$-smooth and the second equality is dependent on the fact that $\nabla_x F_k(\tilde{x}_{k+1},\tilde{y}_{k+1})=0$.

□

We can see that when we can find a δ-saddle point of F_k, then we can approximate $g(x)$ well.

Lemma E.4. *Suppose x_{k+1} satisfies $\mathbb{E}\|x_{k+1}-\tilde{x}_{k+1}\|^2\leq\delta$ for any saddle point $(\tilde{x}_{k+1},\tilde{y}_{k+1})$ of F_k, then it holds that*

$$
\mathbb{E}|g(x_{k+1})-g(\tilde{x}_{k+1})|\leq\left(\frac{(\beta+L)^2}{2(\beta-L)}+\frac{\beta}{2}\right)\delta.
$$

Proof. By definition we know the relationship between $g(x)$ and $F_k(x,y)$ is given by:

$$
g(x)=\max_{y\in\mathbb{R}^{d_y}} f(x,y)=F_k(x,\tilde{y}_{k+1})-\frac{\beta}{2}\|x-x_k\|^2.
$$

Thus,

$$
\begin{aligned}
&\mathbb{E}|g(x_{k+1})-g(\tilde{x}_{k+1})| \\
=\ &\mathbb{E}\left|\left(F_k(x_{k+1},\tilde{y}_{k+1})-\frac{\beta}{2}\|x_{k+1}-x_k\|^2\right)-\left(F_k(\tilde{x}_{k+1},\tilde{y}_{k+1})-\frac{\beta}{2}\|\tilde{x}_{k+1}-x_k\|^2\right)\right| \\
\leq\ &\mathbb{E}\left[F_k(x_{k+1},\tilde{y}_{k+1})-F_k(\tilde{x}_{k+1},\tilde{y}_{k+1})+\frac{\beta}{2}\|x_{k+1}-\tilde{x}_{k+1}\|^2\right] \\
\leq\ &\left(\frac{(\beta+L)^2}{2(\beta-L)}+\frac{\beta}{2}\right)\delta,
\end{aligned}
$$

where the second inequality follows from the triangle inequality of distance and the third inequality follows from Lemma E.3.

□

When the sub-problem is solved precisely enough, we can show that $g(x)$ decreases in each iteration.

Lemma E.5. *Suppose x_{k+1} satisfies $\mathbb{E}\|x_{k+1}-\tilde{x}_{k+1}\|^2\leq\delta$ for any saddle point $(\tilde{x}_{k+1},\tilde{y}_{k+1})$ of F_k, then it holds that*

$$
\mathbb{E}[g(x_{k+1})-g(x^*)]\leq\mathbb{E}\left[g(x_k)-g(x^*)-\frac{\beta}{2}\|x_{k+1}-x_k\|^2+\frac{(\beta+L)^2\delta}{2(\beta-L)}\right],
$$

Proof. Consider the following inequalities:

$$
\begin{aligned}
\mathbb{E}[g(x_{k+1})-g(x^*)]&=\mathbb{E}\left[F_k(x_{k+1},\tilde{y}_{k+1})-g(x^*)-\frac{\beta}{2}\|x_{k+1}-x_k\|^2\right] \\
&\leq\mathbb{E}\left[F_k(\tilde{x}_{k+1},\tilde{y}_{k+1})-g(x^*)-\frac{\beta}{2}\|x_{k+1}-x_k\|^2+\frac{(\beta+L)^2\delta}{2(\beta-L)}\right] \\
&\leq\mathbb{E}\left[g(x_k)-g(x^*)-\frac{\beta}{2}\|x_{k+1}-x_k\|^2+\frac{(\beta+L)^2\delta}{2(\beta-L)}\right],
\end{aligned}
$$

where in the first inequality we use Lemma E.3, the second inequality is because we know it holds that $F_k(\tilde{x}_{k+1},\tilde{y}_{k+1})\leq F_k(x_k,\tilde{y}_{k+1})=g(x_k)$. □

Now, we consider how $g(x_k)$ converge to $g(x^*)$ when precision δ is obtained.

Lemma E.6. *Suppose x_{k+1} satisfies $\mathbb{E}\|x_{k+1} - \tilde{x}_{k+1}\|^2 \leq \delta$ for any saddle point $(\tilde{x}_{k+1}, \tilde{y}_{k+1})$ of F_k, then it holds that*

$$\mathbb{E}[g(x_{k+1}) - g(x^*)] \leq \mathbb{E}\left[g(x_k) - g(x^*) - \frac{1}{4\beta}\|\nabla g(\tilde{x}_{k+1})\|^2 + \left(\frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\right].$$

Proof. The proof is based on Lemma E.5

$$\begin{aligned}
\mathbb{E}[g(x_{k+1}) - g(x^*)] &\leq \mathbb{E}\left[g(x_k) - g(x^*) - \frac{\beta}{2}\|x_{k+1} - x_k\|^2 + \frac{(\beta+L)^2\delta}{2(\beta-L)}\right] \\
&\leq \mathbb{E}\left[g(x_k) - g(x^*) - \frac{\beta}{4}\|\tilde{x}_{k+1} - x_k\|^2 + \frac{\beta}{2}\|x_{k+1} - \tilde{x}_{k+1}\|^2 + \frac{(\beta+L)^2\delta}{2(\beta-L)}\right] \\
&= \mathbb{E}\left[g(x_k) - g(x^*) - \frac{1}{4\beta}\|\nabla g(\tilde{x}_{k+1})\|^2 + \left(\frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\right],
\end{aligned}$$

where the second inequality relies on the fact that $-\|a-b\|^2 \leq \frac{1}{2}\|a\|^2 + \|b\|^2$. In the last equality we use the fact that $(\tilde{x}_{k+1}, \tilde{y}_{k+1})$ is also a stationary point by Lemma A.1, which implies that $\nabla g(\tilde{x}_{k+1}) + \beta(\tilde{x}_{k+1} - x_k) = 0$.

□

E.1 Proof of Lemma 5.1

Proof. Noting that $g(x)$ satisfies μ_x-PL by Lemma A.5 and using the result of Lemma E.6, we can see that

$$\begin{aligned}
\mathbb{E}[g(x_{k+1}) - g(x^*)] &\leq \mathbb{E}\left[g(x_k) - g(x^*) - \frac{1}{4\beta}\|\nabla g(\tilde{x}_{k+1})\|^2 + \left(\frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\right] \\
&\leq \mathbb{E}\left[g(x_k) - g(x^*) - \frac{\mu_x}{2\beta}(g(\tilde{x}_{k+1}) - g(x^*)) + \left(\frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\right].
\end{aligned}$$

Using Lemma E.4, we obtain

$$\mathbb{E}[g(x_{k+1}) - g(x^*)] \leq \mathbb{E}\left[g(x_k) - g(x^*) - \frac{\mu_x}{2\beta}(g(x_{k+1}) - g(x^*)) + \left(1 + \frac{\mu_x}{2\beta}\right)\left(\frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\right].$$

Rearranging,

$$\mathbb{E}[g(x_{k+1}) - g(x^*)] \leq \mathbb{E}\left[\left(1 - \frac{\mu_x}{2\beta + \mu_x}\right)(g(x_k) - g(x^*)) + \left(\frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\right].$$

Let $q \triangleq \mu_x/(2\beta + \mu_x)$ and telescope, then we can obtain that

$$\begin{aligned}
\mathbb{E}[g(x_k) - g(x^*)] &\leq (1-q)^k(g(x_0) - g(x^*)) + \left(\frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\sum_{i=0}^{k-1}(1-q)^i \\
&\leq (1-q)^k(g(x_0) - g(x^*)) + \left(\frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\frac{\delta}{q}.
\end{aligned}$$

Plugging in β, δ yields the desired statement, □

Now we show that how we can control the precision of the sub-problem δ recursively to satisfy the condition of Lemma 5.1 that $\mathbb{E}[\|x_k - \tilde{x}_k\|^2 + \|y_k - \tilde{y}_k\|^2] \leq \delta$ holds for all $k \geq 1$.

Before that, we need the following lemma showing that when $\|x_k - x_{k+1}\|$ is small, then the distance between the saddle points of F_k and F_{k+1} will be also small. Denote $(\tilde{x}_k, \tilde{y}_k)$ be a saddle point of sub-problem F_{k-1} and $(\tilde{x}_{k+1}, \tilde{y}_{k+1})$ be a saddle point of sub-problem F_k.

Lemma E.7. *If we let $\beta > L$, then it holds true that*

$$\|\tilde{x}_{k+1} - \tilde{x}_k\|^2 + \|\tilde{y}_{k+1} - \tilde{y}_k\|^2 \leq \frac{4\beta^2}{(\beta-L)\mu_y}\|x_k - x_{k-1}\|^2.$$

Proof. Noting that F_{k-1} is $(\beta - L)$-PL in x and μ_y-PL in y by Lemma E.1 and using the quadratic growth condition by Lemma A.4, we have

$$\begin{aligned}\frac{\beta-L}{2}\|\tilde{x}_{k+1}-\tilde{x}_k\|^2 &\le F_{k-1}(\tilde{x}_{k+1},\tilde{y}_k) - \min_{x\in\mathbb{R}^{d_x}} F_{k-1}(x,\tilde{y}_k) = F_{k-1}(\tilde{x}_{k+1},\tilde{y}_k) - F_{k-1}(\tilde{x}_k,\tilde{y}_k),\\ \frac{\mu_y}{2}\|\tilde{y}_{k+1}-\tilde{y}_k\|^2 &\le \max_{y\in\mathbb{R}^{d_y}} F_{k-1}(\tilde{x}_k,y) - F_{k-1}(\tilde{x}_k,\tilde{y}_{k+1}) = F_{k-1}(\tilde{x}_k,\tilde{y}_k) - F_{k-1}(\tilde{x}_k,\tilde{y}_{k+1}).\end{aligned}$$

Combining the above two inequalities, we can see that

$$\begin{aligned}&\frac{\beta-L}{2}\|\tilde{x}_{k+1}-\tilde{x}_k\|^2 + \frac{\mu_y}{2}\|\tilde{y}_{k+1}-\tilde{y}_k\|^2\\ &\le F_{k-1}(\tilde{x}_{k+1},\tilde{y}_k) - F_{k-1}(\tilde{x}_k,\tilde{y}_{k+1})\\ &= F_k(\tilde{x}_{k+1},\tilde{y}_k) + \frac{\beta}{2}\|\tilde{x}_{k+1}-x_{k-1}\|^2 - \frac{\beta}{2}\|\tilde{x}_{k+1}-x_k\|^2\\ &\quad - F_k(\tilde{x}_k,\tilde{y}_{k+1}) - \frac{\beta}{2}\|\tilde{x}_k - x_{k-1}\|^2 + \frac{\beta}{2}\|\tilde{x}_k - x_k\|^2\\ &\le \frac{\beta}{2}\|\tilde{x}_{k+1}-x_{k-1}\|^2 - \frac{\beta}{2}\|\tilde{x}_{k+1}-x_k\|^2 - \frac{\beta}{2}\|\tilde{x}_k - x_{k-1}\|^2 + \frac{\beta}{2}\|\tilde{x}_k - x_k\|^2\\ &= \beta(\tilde{x}_{k+1}-\tilde{x}_k)^\top(x_k - x_{k-1})\\ &\le \frac{\beta-L}{4}\|\tilde{x}_{k+1}-\tilde{x}_k\|^2 + \frac{\beta^2}{\beta-L}\|x_k - x_{k-1}\|^2,\end{aligned}$$

where the second inequality is based on $F_k(\tilde{x}_{k+1},\tilde{y}_k) \le F_k(\tilde{x}_{k+1},\tilde{y}_{k+1}) \le F_k(\tilde{x}_k,\tilde{y}_{k+1})$ by $(\tilde{x}_{k+1},\tilde{y}_{k+1})$ is a saddle point of F_k. In the last inequality we use Young's inequality. Rearranging,

$$\frac{\beta-L}{4}\|\tilde{x}_{k+1}-\tilde{x}_k\|^2 + \frac{\mu_y}{2}\|\tilde{y}_{k+1}-\tilde{y}_k\|^2 \le \frac{\beta^2}{\beta-L}\|x_k - x_{k-1}\|^2.$$

Since we have $(\beta - L)/\mu_y \ge L/4 \ge \mu_y/4$, we can obtain that

$$\|\tilde{x}_{k+1}-\tilde{x}_k\|^2 + \|\tilde{y}_{k+1}-\tilde{y}_k\|^2 \le \frac{4\beta^2}{(\beta-L)\mu_y}\|x_k - x_{k-1}\|^2,$$

□

An additional bound is for the use of the base case, i.e. $k = 0$.

Lemma E.8. *If we let $\beta > L$, then it holds true that*

$$\|x_0 - \tilde{x}_1\|^2 + \|y_0 - \tilde{y}_1\|^2 \le \frac{2}{\mu_y}(g(x_0) - g(x^*))$$

Proof. Note that F_1 satisfies $(\beta - L)$-PL in x and μ_y-PL in y. We can bound $\|x_0 - \tilde{x}_1\|^2 + \|y_0 - \tilde{y}_1\|^2$ as follows:

$$\begin{aligned}\frac{\beta-L}{2}\|x_0-\tilde{x}_1\|^2 &\le F_0(x_0,\tilde{y}_1) - \min_{x\in\mathbb{R}^{d_x}} F_0(x,\tilde{y}_1) = F_0(x_0,\tilde{x}_1) - F_0(\tilde{x}_1,\tilde{y}_1),\\ \frac{\mu_y}{2}\|y_0-\tilde{y}_1\|^2 &\le \max_{y\in\mathbb{R}^{d_y}} F_0(\tilde{x}_1,y) - F_0(\tilde{x}_1,y_0) = F_0(\tilde{x}_1,\tilde{y}_1) - F_0(\tilde{x}_1,y_0).\end{aligned}$$

Combining the above two inequalities, we have

$$\begin{aligned}&\frac{\beta-L}{2}\|x_0-\tilde{x}_1\|^2 + \frac{\mu_y}{2}\|y_0-\tilde{y}_1\|^2\\ &\le F_0(x_0,\tilde{y}_1) - F_0(\tilde{x}_1,y_0)\\ &= f(x_0,\tilde{y}_1) - f(\tilde{x}_1,y_0) - \frac{\beta}{2}\|x_0-\tilde{x}_1\|^2\\ &\le f(x_0,\tilde{y}_1) - f(\tilde{x}_1,y_0)\\ &\le \max_{y\in\mathbb{R}^{d_y}} f(x_0,y) - \min_{x\in\mathbb{R}^{d_x}} f(x,y_0)\end{aligned}$$

$$
\begin{aligned}
&= \max_{y\in\mathbb{R}^{d_y}} f(x_0,y) - \min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} f(x,y) + \min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} f(x,y) - \min_{x\in\mathbb{R}^{d_x}} f(x,y_0) \\
&\le g(x_0) - g(x^*),
\end{aligned}
$$

where we use the definition of g and the fact that $\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} f(x,y) \ge \min_{x\in\mathbb{R}^{d_x}} f(x,y_0)$ in the last inequality. Rearranging and noting that $\beta - L \ge L \ge \mu_y$, we finish the proof. $\square$

E.2 The Proof of Lemma 5.2

Proof. Recall we solve the sub-problem:

$$
\max_{y\in\mathbb{R}^{d_y}} \min_{x\in\mathbb{R}^{d_x}} F_k(x,y) = -\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \{-F_k(x,y)\}.
$$

It is μ_y-PL in y and L-strongly-convex in x and thus clearly satisfies L-PL in x.

We define $H_k(y) = \min_{x\in\mathbb{R}_{d_x}} F_k(x,y)$. By Lemma A.8 and A.5, we know that $H_k(y)$ is also μ_y-PL in y and it is $12L$-smooth since F_k is $3L$-smooth.

According to Theorem 4.1, We know that SPIDER-GDA makes sure

$$
\begin{aligned}
&\underbrace{\mathbb{E}\left[H_{k+1}(\tilde{y}_{k+1}) - H_{k+1}(y_{k+1}) + \frac{1}{24}\left(F_{k+1}(x_{k+1},y_{k+1}) - H_{k+1}(y_{k+1})\right)\right]}_{\text{LHS}} \\
&\le \delta_k \underbrace{\mathbb{E}\left[H_{k+1}(\tilde{y}_{k+1}) - H_{k+1}(y_k) + \frac{1}{24}\left(F_{k+1}(x_k,y_k) - H_{k+1}(y_k)\right)\right]}_{\text{RHS}}.
\end{aligned}
$$

For the left hand side (LHS), we bound it according to

$$
\mathbb{E}\|y_{k+1} - \tilde{y}_{k+1}\|^2 \le \frac{2}{\mu_y}\mathbb{E}[H_{k+1}(\tilde{y}_{k+1}) - H_{k+1}(y_{k+1})] \tag{29}
$$

(where we use the μ_y-PL condition in y) and

$$
\begin{aligned}
&\mathbb{E}\|x_{k+1} - \tilde{x}_{k+1}\|^2 \\
&\le 2\mathbb{E}[\|x_{k+1} - x^*(y_{k+1})\|^2 + \|x^*(y_{k+1}) - \tilde{x}_{k+1}\|^2] \\
&= 2\mathbb{E}[\|x_{k+1} - x^*(y_{k+1})\|^2 + \|x^*(y_{k+1}) - x^*(\tilde{y}_{k+1})\|^2] \\
&\le 2\mathbb{E}[\|x_{k+1} - x^*(y_{k+1})\|^2 + 18\mathbb{E}[\|y_{k+1} - \tilde{y}_{k+1}\|^2] \\
&\le \frac{4}{L}\mathbb{E}[F_{k+1}(x_{k+1},y_{k+1}) - H_{k+1}(y_{k+1})] + \frac{36}{\mu_y}\mathbb{E}[H_{k+1}(\tilde{y}_{k+1}) - H_{k+1}(y_{k+1}))],
\end{aligned} \tag{30}
$$

where we use Lemma A.9 in the second last inequality and the PL condition of $F_k(x,y)$ and $H_k(y)$ in the last one. Summing up (29) and (30), we have

$$
\begin{aligned}
&\mathbb{E}[\|y_{k+1} - \tilde{y}_{k+1}\|^2 + \|x_{k+1} - \tilde{x}_{k+1}\|^2] \\
&\le \frac{4}{L}\mathbb{E}[F_{k+1}(x_{k+1},y_{k+1}) - H_{k+1}(y_{k+1})] + \frac{38}{\mu_y}\mathbb{E}[H_{k+1}(\tilde{y}_{k+1}) - H_{k+1}(y_{k+1}))] \\
&\le \frac{96}{\mu_y} \times \text{LHS} \\
&\le \frac{96\delta_k}{\mu_y} \times \text{RHS}.
\end{aligned}
$$

For the right hand side (RHS), we bound it using

$$
\begin{aligned}
\text{RHS} &= H_{k+1}(\tilde{y}_{k+1}) - H_{k+1}(y_k) + \frac{1}{24}\left(F_{k+1}(x_k,y_k) - H_{k+1}(y_k)\right) \\
&\le \frac{1}{2\mu_y}\|\nabla H_{k+1}(y_k)\|^2 + \frac{1}{48L}\|\nabla_x F_{k+1}(x_k,y_k)\|^2
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2\mu_y}\|\nabla H_{k+1}(y_k) - \nabla H_{k+1}(\tilde{y}_{k+1})\|^2 + \frac{1}{48L}\|\nabla_x F_{k+1}(x_k, y_k) - \nabla_x F_{k+1}(x^*(y_k), y_k)\|^2 \\
&\leq \frac{72L^2}{\mu_y}\|y_k - \tilde{y}_{k+1}\|^2 + \frac{3L}{16}\|x_k - x^*(y_k)\|^2 \\
&\leq \frac{72L^2}{\mu_y}\|y_k - \tilde{y}_{k+1}\|^2 + \frac{3L}{8}\|x_k - \tilde{x}_{k+1}\|^2 + \frac{3L}{8}\|x^*(\tilde{y}_{k+1}) - x^*(y_k)\|^2 \\
&\leq \frac{72L^2}{\mu_y}\|y_k - \tilde{y}_{k+1}\|^2 + \frac{L}{24}\|x_k - \tilde{x}_{k+1}\|^2 + \frac{27L}{8}\|y_k - \tilde{y}_{k+1}\|^2.
\end{aligned}
$$

Above, the first inequality is due to the PL condition of $F_k(x,y)$ and $H_k(y)$; the second inequality follows from $H_k(y)$ is $12L$-smooth and $F_k(x,y)$ is $3L$-smooth; the third one directly follows from the Young's inequality; the fourth inequality uses Lemma A.9 and the fact that $F_k(x,y)$ is L-PL in x and $3L$-smooth.

Therefore, we obtain

$$\mathbb{E}[\|y_{k+1} - \tilde{y}_{k+1}\|^2 + \|x_{k+1} - \tilde{x}_{k+1}\|^2] \leq \delta_k'(\|y_k - \tilde{y}_{k+1}\|^2 + \|x_k - \tilde{x}_{k+1}\|^2),$$

where

$$\delta_k' = 7236\kappa_y^2\delta_k. \tag{31}$$

□

Now it is sufficient to control δ recursively.

Lemma E.9. *If we solve each sub-problem F_k with precision δ_k as defined in Theorem 5.1, then for all k it holds true that*

$$\mathbb{E}\|x_k - \tilde{x}_k\|^2 + \|y_k - \tilde{y}_k\|^2 \leq \delta.$$

Proof. We prove by induction. Suppose we the following statement holds true for all $1 \leq k' \leq k$ that we have

$$\mathbb{E}\|x_{k'} - \tilde{x}_{k'}\|^2 + \|y_{k'} - \tilde{y}_{k'}\|^2 \leq \delta,$$

Then, by Lemma 5.2 we have

$$
\begin{aligned}
&\mathbb{E}[\|x_{k+1} - \tilde{x}_{k+1}\|^2 + \|y_{k+1} - \tilde{y}_{k+1}\|^2] \\
&\leq \delta_k'(\|x_k - \tilde{x}_{k+1}\|^2 + \|y_k - \tilde{y}_{k+1}\|^2) \\
&\leq 2\delta_k'(\|x_k - \tilde{x}_k\|^2 + \|y_k - \tilde{y}_k\|^2) + 2\delta_k'(\|\tilde{x}_{k+1} - \tilde{x}_k\|^2 + \|\tilde{y}_{k+1} - \tilde{y}_k\|^2) \\
&\leq 2\delta_k'\delta + \frac{8\beta^2\delta_k'}{(\beta - L)\mu_y}\|x_k - x_{k-1}\|^2,
\end{aligned}
$$

where δ_k' follows from (31) and we use the induction hypothesis and Lemma E.7 in the third inequality. Note that our choice of δ_k and the relationship between δ_k and δ_k' satisfy

$$\max\left\{2\delta_k'\delta, \frac{8\beta^2\delta_k'\|x_k - x_{k-1}\|^2}{(\beta - L)\mu_y}\right\} \leq \frac{\delta}{2}.$$

Therefore, we can see that

$$\mathbb{E}[\|x_{k+1} - \tilde{x}_{k+1}\|^2 + \|y_{k+1} - \tilde{y}_{k+1}\|^2] \leq \delta,$$

which completes the induction from k to $k+1$. For the induction base, using Lemma E.8 we have

$$
\begin{aligned}
\mathbb{E}[\|x_1 - \tilde{x}_1\|^2 + \|y_1 - \tilde{y}_1\|^2] &\leq \delta_0'(\|x_0 - \tilde{x}_1\|^2 + \|y_0 - \tilde{y}_1\|^2) \\
&\leq \frac{2\delta_0'}{\mu_y}(g(x_0) - g^*) \\
&\leq \delta.
\end{aligned}
$$

□

E.3 Proof of Theorem 5.1

Combing Lemma 5.1, Lemma 5.2 and Lemma E.9, we can easily prove Theorem 5.1.

Proof. Note that each sub-problem F_k is $3L$-smooth, L-PL in x and μ_y-PL in y for $\beta = 2L$. Now if we choose

$$K = \lceil ((2\beta + \mu_x)/\mu_x) \log(2/\epsilon) \rceil = \mathcal{O}(\kappa_x \log(1/\epsilon)),$$

then by Lemma 5.1 it is sufficient to guarantee that $\mathbb{E}[g(x_K) - g(x^*)] \leq \epsilon$, while solving each sub-problem F_k requires no more than $T_k \leq a(n + \sqrt{n}\kappa_y) \log(\kappa_y/\delta_k)$ first-order oracle calls in expectation by Lemma 5.2, where a is an independent positive constant.

Now we telescope the inequality in Lemma E.5 and we can obtain

$$\sum_{k=0}^{K-1} \mathbb{E}\left[\|x_{k+1} - x_k\|^2\right] \leq \frac{2}{\beta}(g(x_0) - g^*) + \frac{(\beta + L)^2 \delta}{\beta(\beta - L)}. \tag{32}$$

Note that we have

$$\frac{1}{\delta_k} \leq \omega \times \max\left\{4, \frac{16\kappa_y \|x_k - x_{k-1}\|^2}{\delta}\right\} \leq \omega \times \left(4 + \frac{16\kappa_y \|x_k - x_{k-1}\|^2}{\delta}\right),$$

by the choice of δ_k for all $k \geq 1$ (5), where $\omega = 7236\kappa_y^2$. Denote $C = a(n + \sqrt{n}\kappa)$, then

$$\begin{aligned}
&\sum_{k=0}^{K} T_k \\
&= T_0 + \sum_{k=1}^{K} T_k \\
&\leq C \log\left(\omega \times \frac{2\kappa_y(g(x_0) - g^*)}{\delta\mu_y}\right) + C\sum_{k=1}^{K} \log\left(\frac{\kappa_y}{\delta_k}\right) \\
&\leq C \log\left(\omega \times \frac{2\kappa_y(g(x_0) - g^*)}{\delta\mu_y}\right) + C\sum_{k=1}^{K} \log\left(\omega \times \left(4\kappa_y + \frac{16\kappa_y^2 \|x_k - x_{k-1}\|^2}{\delta}\right)\right) \\
&\leq C \log\left(\omega \times \frac{2\kappa_y(g(x_0) - g^*)}{\delta\mu_y}\right) + CK \log\left(\omega \times \sum_{k=1}^{K}\left(4\kappa_y + \frac{16\kappa_y^2 \|x_k - x_{k-1}\|^2}{\delta}\right)\right) \\
&= C \log\left(\omega \times \frac{2\kappa_y(g(x_0) - g^*)}{\delta\mu_y}\right) + CK \log\left(\omega \times \sum_{k=0}^{K-1}\left(4\kappa_y + \frac{16\kappa_y^2 \|x_k - x_{k-1}\|^2}{\delta}\right)\right).
\end{aligned} \tag{33}$$

Above, the second inequality relies on the choice of δ_k and the third inequality is due to the fact that $(\prod_{i=1}^n x_i)^{1/n} \leq \frac{1}{n}\sum_{i=1}^n x_i$, which implies that $\sum_{i=1}^n \log x_i \leq n \log\left(\frac{1}{n}\sum_{i=1}^n x_i\right)$.

Lastly, we use (32) and notice that δ (3) is dependent on ϵ, κ_x and ω is dependent on κ_y to show the SFO complexity of the order

$$\mathcal{O}((n\kappa_x + \sqrt{n}\kappa_x\kappa_y) \log(1/\epsilon) \log(\kappa_x\kappa_y/\epsilon)).$$

□

F Proof of Section 6

First of all, we show the convergence of SVRG-GDA and SPIDER-GDA under one-sided PL condition as studied in Section 6. We reuse the lemmas under two-sided PL condition. It is worth noticing that we can discard the outermost loop with respect to restart strategy for both SVRG-GDA and SPIDER-GDA, i.e we set $T = 1$ in this setting.

F.1 SVRG-GDA under one-sided PL condition

For SVRG-GDA, we have the following theorem.

Theorem F.1. *Under Assumption 6.1 and 3.1, let $T = 1$ and $M, \tau_x, \tau_y, \lambda$ defined in Lemma C.1; $\alpha = 2/3$, $SM = \lceil 8/(\tau_x \epsilon^2) \rceil$. Algorithm 2 can guarantee the output $\hat{x}$ to satisfy $\|\nabla g(\hat{x})\|^2 \le \epsilon$ in expectation with no more than $\mathcal{O}(n + n^{2/3}\kappa_y^2 L \epsilon^{-2})$ stochastic first-order oracle calls.*

Proof. Telescoping for $k = 0, \dots, M-1$ and $s = 0, \dots, S-1$ for the inequality in Lemma C.1:

$$\frac{1}{SM}\sum_{s=0}^{S-1}\sum_{k=0}^{M-1}\mathbb{E}[\|\nabla g(x_{s,k})\|^2] \le \frac{8\mathcal{V}_{0,0}}{\tau_x SM}.$$

Note that we have $M = \mathcal{O}(n^{3\alpha/2})$ and if we let $SM = \lceil 8/(\tau_x\epsilon^2)\rceil$, then $S = \mathcal{O}(L\kappa_y^2/(n^{\alpha/2}\epsilon^2))$, so the complexity is

$$\mathcal{O}(n + SM + Sn) = \mathcal{O}\left(n + \frac{\kappa_y^2 L(n^{\alpha} + n^{1-\alpha/2})}{\epsilon^2}\right).$$

Plugging in $\alpha = 2/3$ yields the desired complexity. □

F.2 Proof of Theorem 6.1

Similarly to SVRG-GDA, we can also analyze the convergence of SPIDER-GDA.

Proof. It is almost direct result of Lemma D.2. Denote $(\hat{x}, \hat{y})$ the output, then

$$\begin{aligned}
\mathbb{E}\|\nabla g(\hat{x})\|^2 &\le 2\mathbb{E}\left[\|\nabla_x f(\hat{x},\hat{y})\|^2 + \|\nabla_x f(\hat{x},\hat{y}) - \nabla g(\hat{x})\|^2\right] \\
&\le 2\mathbb{E}\left[\|\nabla_x f(\hat{x},\hat{y})\|^2 + \frac{L^2}{\mu_y^2}\|\nabla_y f(\hat{x},\hat{y})\|^2\right] \\
&\le 2\mathbb{E}[\|\nabla_x f(\hat{x},\hat{y})\|^2 + \lambda\|\nabla_y f(\hat{x},\hat{y})\|^2] \\
&\le \frac{32}{\tau_x K}\mathbb{E}\left[\tilde{\mathcal{A}}_0 + \tilde{\mathcal{B}}_0\right],
\end{aligned}$$

where the first inequality follows from Young's inequality; the second one relies on Lemma A.7; the third one uses the definition of λ and the last one uses Lemma D.2.

Since $\tau_x = \mathcal{O}(1/(\kappa_y^2 L))$ and $M = B = \sqrt{n}$, the complexity becomes:

$$\mathcal{O}\left(n + \frac{\sqrt{n}}{\tau_x\epsilon^2}\right) = \mathcal{O}\left(n + \frac{\sqrt{n}\kappa_y^2 L}{\epsilon^2}\right).$$

□

Above, we have show that the complexity of SVRG-GDA is $\mathcal{O}(n + n^{2/3}\kappa_y^2 L\epsilon^{-2})$ and the complexity of SPIDER-GDA is $\mathcal{O}(n + \sqrt{n}\kappa_y^2 L\epsilon^{-2})$ [6]. Thus, we can come to the conclusion that SPIDER-GDA strictly outperforms SVRG-GDA under both two-sided and one-sided PL conditions. In the rest of this section, we mainly focus on the complexity of AccSPIDER-GDA under one-sided PL condition.

In the following lemma, we show that AccSPIDER-GDA converge when we can control the precision of solving each sub-problem with a global constant δ.

[6]To be more precise, our theorem only suits the case when $1/(\kappa_y^2 L\epsilon^2) > \sqrt{n}$ for SPIDER-GDA. If not, we can directly set $K = 2M$ to achieve the same convergence result.

F.3 Proof of Lemma 6.1

Proof. Similar to the proof under two-sided PL condition, we begin our proof with Lemma E.6. We can see that

$$
\begin{aligned}
\mathbb{E}[g(x_{k+1})] &\leq \mathbb{E}\left[g(x_k) - \frac{1}{4\beta}\|\nabla g(\tilde{x}_{k+1})\|^2 + \left(\frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\right] \\
&\leq \mathbb{E}\left[g(x_k) - \frac{1}{8\beta}\|\nabla g(x_{k+1})\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{1}{4\beta}\|\nabla g(\tilde{x}_{k+1}) - \nabla g(x_{k+1})\|^2 + \left(\frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\right] \\
&\leq \mathbb{E}\left[g(x_k) - \frac{1}{8\beta}\|\nabla g(x_{k+1})\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{L^2}{2\mu_y\beta}\|\tilde{x}_{k+1} - x_{k+1}\|^2 + \left(\frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\right] \\
&\leq \mathbb{E}\left[g(x_k) - \frac{1}{8\beta}\|\nabla g(x_{k+1})\|^2 + \left(\frac{L^2}{2\mu_y\beta} + \frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\right],
\end{aligned}
$$

where we use the fact that $-\|a-b\|^2 \leq \frac{1}{2}\|a\|^2 + \|b\|^2$ in the second inequality, $g(x)$ is $(2L^2/\mu_y)$-smooth in the third one and $\|\tilde{x}_{k+1} - x_{k+1}\|^2 \leq \delta$ in the last one. Telescoping for $k = 0, 1, 2, ...K-1$, we can see that

$$
\begin{aligned}
\frac{1}{8\beta}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla g(x_k)\|^2\right] &\leq \mathbb{E}\left[g(x_0) - g(x_K) + \left(\frac{L^2}{2\mu_y\beta} + \frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)K\delta\right] \\
&\leq \mathbb{E}\left[g(x_0) - g^* + \left(\frac{L^2}{2\mu_y\beta} + \frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)K\delta\right].
\end{aligned}
$$

Divide both sides by K, then

$$
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla g(x_k)\|^2\right] \leq \mathbb{E}\left[\frac{8\beta(g(x_0) - g^*)}{K} + 8\beta\left(\frac{L^2}{2\mu_y\beta} + \frac{(\beta+L)^2}{2(\beta-L)} + \frac{\beta}{2}\right)\delta\right].
$$

Plugging the choice of δ yields the desired inequality. □

F.4 Proof of Theorem 6.2

Combing Lemma 6.1, Lemma 5.2 and Lemma E.9, we can easily prove Theorem 6.2. We remark that both the proof Lemma 5.2 and Lemma E.9 only uses the PL property in the direction of y, so they can both be directly applied to the one-sided PL case.

Proof. Note that each sub-problem F_k is $3L$-smooth, L-PL in x and μ_y-PL in y for $\beta = 2L$. Now if we choose

$$
K = \lceil 16\beta(g(x_0) - g^*)/\epsilon^2 \rceil = \mathcal{O}(L\epsilon^{-2}),
$$

then by Lemma 6.1 it is sufficient to guarantee that $\mathbb{E}[\|g(\hat{x})\|] \leq \epsilon$, while solving each sub-problem F_k requires no more than $T_k \leq a(n + \sqrt{n}\kappa_y)\log(\kappa_y/\delta_k)$ first-order oracle calls in expectation by Lemma 5.2, where a is an independent positive constant.

Therefore, using (32), (33) and noticing that δ (6) is dependent on ϵ, κ_y and ω in (33) is dependent on κ_y to show the SFO complexity of the order

$$
\mathcal{O}((n + \sqrt{n}\kappa_y)L\epsilon^{-2}\log(\kappa_y/\epsilon)).
$$

□