Faster Single-loop Algorithms for Minimax Optimization without Strong Concavity

Junchi Yang ETH Zürich

Antonio Orvieto ETH Zürich

Aurelien Lucchi University of Basel

Niao He ETH Zürich

Abstract

Gradient descent ascent (GDA), the simplest single-loop algorithm for nonconvex minimax optimization, is widely used in practical applications such as generative adversarial networks (GANs) and adversarial training. Albeit its desirable simplicity, recent work shows inferior convergence rates of GDA in theory, even when assuming strong concavity of the objective in terms of one variable. This paper establishes new convergence results for two alternative single-loop algorithms – *alternating GDA* and *smoothed GDA* – under the mild assumption that the objective satisfies the Polyak-Łojasiewicz (PL) condition about one variable. We prove that, to find an ϵ-stationary point, (i) alternating GDA and its stochastic variant (without mini batch) respectively require $O(\kappa^2\epsilon^{-2})$ and $O(\kappa^4\epsilon^{-4})$ iterations, while (ii) smoothed GDA and its stochastic variant (without mini batch) respectively require $O(\kappa\epsilon^{-2})$ and $O(\kappa^2\epsilon^{-4})$ iterations. The latter greatly improves over the vanilla GDA and gives the hitherto best known complexity results among single-loop algorithms under similar settings. We further showcase the empirical efficiency of these algorithms in training GANs and robust nonlinear regression.

1 INTRODUCTION

Minimax optimization plays an important role in classical game theory and a wide spectrum of emerging machine learning applications, including but not limited to, generative adversarial networks (GANs) (Goodfellow et al., 2014a), multi-agent reinforcement learning (Zhang et al., 2021b), and adversarial training (Goodfellow et al., 2014b). Many of the aforementioned problems lie outside of the canonical convex-concave setting and can be intractable (Hsieh et al., 2021; Daskalakis et al., 2021). Notably, Daskalakis et al. (2021) showed that, in the worst-case, first-order algorithms need an exponential number of queries to find approximate local solutions for some smooth minimax objectives.

In this paper, we consider finding stationary points for the general nonconvex smooth minimax optimization problems:

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} f(x,y) \triangleq \mathbb{E}[F(x,y;\xi)], \tag{1}$$

where ξ is a random vector with support Ξ and $f(x,y)$ is nonconvex in x for any fixed y and possibly nonconcave in y.

Due to its simplicity and single-loop nature, gradient descent ascent (GDA) and its stochastic variants, have become the *de facto* algorithms for training GANs and many other applications in practice. Their theoretical properties have also been extensively studied in recent literature (Lei et al., 2020; Nagarajan and Kolter, 2017; Heusel et al., 2017; Mescheder et al., 2017, 2018).

Lin et al. (2020a) derived a complexity analysis for simultaneous GDA (with simultaneous updates for x and y) and for stochastic GDA (hereafter Stoc-GDA) for finding stationary points when the objective is concave in y. In particular, they show that GDA requires $O(\epsilon^{-6})$ iterations and Stoc-GDA without mini-batch requires $O(\epsilon^{-8})$ samples to achieve an ϵ-approximate stationary point. When the objective is strongly concave in y, the iteration complexity of GDA can be significantly improved to $O(\kappa^2\epsilon^{-2})$ while the sample complexity for Stoc-GDA reduces to $O(\kappa^3\epsilon^{-4})$ with a large batch of size $O(\epsilon^{-2})$ or $O(\kappa^3\epsilon^{-5})$ without the batch, i.e., using a single sample to construct the gradient estimator. Here κ is the underlying condition number defined as l/μ with l being Lipschitz smoothness parameter and μ strong concavity parameter. However,

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

Table 1: Oracle complexities for deterministic NC-PL problems. Here $\tilde{O}(\cdot)$ hides poly-logarithmic factors. l: Lipschitz smoothness parameter; μ: PL parameter, κ: condition number $\frac{l}{\mu}$; Δ: initial gap of the primal function. We measure the stationarity by $\|\nabla\Phi(x)\|$ with $\Phi(x) = \max_y f(x, y)$ and $\|\nabla f(x, y)\|$. Here * means the complexity is derived by translating from one stationary measure to the other (see Proposition 2.1). $\diamond$ it recovers the same complexity for AGDA as Appendix D in (Yang et al., 2020a)

Algorithms	**Complexity** $\|\nabla\Phi(x)\| \leq \epsilon$	**Complexity** $\|\nabla f(x, y)\| \leq \epsilon$	**Loops**	**Additional assumptions**
GDA (Lin et al., 2020a)	$O(\kappa^2\Delta l\epsilon^{-2})$	$O(\kappa^2\Delta l\epsilon^{-2})^\star$	1	strong concavity in y
Catalyst-EG (Zhang et al., 2021c)	$O(\sqrt{\kappa}\Delta l\epsilon^{-2})$	$O(\sqrt{\kappa}\Delta l\epsilon^{-2})^\star$	3	strong concavity in y
Multi-GDA (Nouiehed et al., 2019)	$\tilde{O}(\kappa^3\Delta l\epsilon^{-2})^\star$	$\tilde{O}(\kappa^2\Delta l\epsilon^{-2})$	2	
Catalyst-AGDA [Appendix D]	$O(\kappa\Delta l\epsilon^{-2})$	$O(\kappa\Delta l\epsilon^{-2})$	2	
AGDA	$O(\kappa^2\Delta l\epsilon^{-2})^\diamond$	$O(\kappa^2\Delta l\epsilon^{-2})$	1	
Smoothed-AGDA	$O(\kappa\Delta l\epsilon^{-2})$	$O(\kappa\Delta l\epsilon^{-2})$	1	

the following question is still unsettled: **can stochastic GDA-type algorithms achieve the better sample complexity of $O(\epsilon^{-4})$ without a large batch size?**

Besides the dependence on ϵ, the condition number also plays a crucial role in the convergence rate. There is a long line of research aiming to reduce such a dependency, see e.g. (Lin et al., 2020b; Zhang et al., 2021c) for some recent results for minimax optimization. These algorithms are typically more complicated as they rely on multiple loops, and are equipped with several acceleration mechanisms. Single-loop algorithms are far more favorable in practice because of their simplicity in implementation. Recently, several single-loop variants of GDA have been proposed, including Alternating Gradient Projection (AGP) (Xu et al., 2020b) and Smoothed-AGDA (Zhang et al., 2020). Unfortunately, most of them fail to provide faster convergence in terms of the condition number and they only consider the deterministic setting. The following question is therefore still unanswered: **is it possible to improve the dependence on the condition number without resorting to multi-loop procedures?**

In short, there is an urgent need to obtain *faster convergence in terms of both the target accuracy ϵ and the condition number κ with single-loop algorithms*. This is even more challenging when the objective is not strongly-concave about y. In this paper, we investigate two viable single-loop algorithms to address this question: (i) *alternating GDA* (hereafter AGDA and Stoc-AGDA for their stochastic variance) and (ii) *Smoothed-AGDA*. Importantly, AGDA, with sequential updates between x and y, is one of the most popular algorithms used in practice and has an edge over GDA in several settings (Zhang et al., 2021a). Smoothed-AGDA, first introduced by (Zhang et al., 2020), utilizes a regularization term to stabilize the performance of GDA when the objective is convex in y. We show that these two algorithms can satisfy our need to achieve faster convergence under milder assumptions.

We are interested in analyzing their theoretical behaviors under the general *NC-PL setting*, namely, the objective is nonconvex in x and satisfies the Polyak-Łojasiewicz (PL) condition in y (Polyak, 1963). This is a milder assumption than strong concavity and does not even require the objective to be concave in y. Such an assumption has been shown to hold in linear quadratic regulators (Fazel et al., 2018), as well as overparametrized neural networks (Liu et al., 2020a). This setting has driven a lot of the recent progress in the quest for understanding deep neural networks (Lee et al., 2017; Jacot et al., 2018), and it therefore appears as an ideal candidate to deepen our understanding of the convergence properties of minimax optimization.

1.1 Contributions

In this work, we study the convergence of AGDA and Smoothed-AGDA in the NC-PL setting. Our goal is to find an approximate stationary point for the objective function $f(\cdot,\cdot)$ and its primal function $\Phi(\cdot) \triangleq \max_y f(\cdot, y)$. For each algorithm, we present a *unified* analysis for the deterministic setting, when we have access to exact gradients of (1), and the stochastic setting, when we have access to noisy gradients. We denote the smoothness parameter by l, PL parameter by μ, condition number by $\kappa \triangleq \frac{l}{\mu}$ and initial primal function gap $\Phi(x) - \inf_x \Phi(x)$ by Δ.

Table 2: Sample complexities for stochastic NC-PL problems when the target accuracy ϵ is small, i.e. $\epsilon \leq \tilde{O}(\sqrt{\Delta l/\kappa^3})$. We measure the stationarity by $\|\nabla\Phi(x)\|$ with $\Phi(x) = \max_y f(x,y)$ and $\|\nabla f(x,y)\|$. Here * means the complexity is derived by translating from one stationary measure to the other (see Proposition 2.1). $^\triangledown$ It assumes the function f is Lipschitz continuous about x and its Hessian is Lipschitz continuous.

Algorithms	Complexity $\|\nabla\Phi(x)\| \leq \epsilon$	Complexity $\|\nabla f(x,y)\| \leq \epsilon$	Batch size	Additional assumptions
Stoc-GDA(Lin et al., 2020a)	$O(\kappa^3\Delta l\epsilon^{-4})$	$O(\kappa^3\Delta l\epsilon^{-4})^\star$	$O(\epsilon^{-2})$	strong concavity in y
Stoc-GDA(Lin et al., 2020a)	$O(\kappa^3\Delta l\epsilon^{-5})$	$O(\kappa^3\Delta l\epsilon^{-5})^\star$	$O(1)$	strong concavity in y
PDSM(Guo et al., 2021)	$O(\kappa^3\Delta l\epsilon^{-4})$	$O(\kappa^3\Delta l\epsilon^{-4})^\star$	$O(1)$	strong concavity in y
ALSET(Chen et al., 2021a)	$O(\kappa^3\Delta l\epsilon^{-4})$	$O(\kappa^3\Delta l\epsilon^{-4})^\star$	$O(1)$	strong concavity in y, Lipschitz$^\triangledown$
Stoc-AGDA	$O(\kappa^4\Delta l\epsilon^{-4})$	$O(\kappa^4\Delta l\epsilon^{-4})$	$O(1)$	
Stoc-Smoothed-AGDA	$O(\kappa^2\Delta l\epsilon^{-4})$	$O(\kappa^2\Delta l\epsilon^{-4})$	$O(1)$	

Deterministic setting. We first show that the output from AGDA is an ϵ-stationary point for both the objective function f and primal function Φ after $O(\kappa^2\Delta l\epsilon^{-2})$ iterations, which recovers the result of primal function stationary convergence in (Yang et al., 2020a) based on a different analysis. The complexity is optimal in ϵ, since $\Omega(\epsilon^{-2})$ is the lower bound for smooth optimization problems (Carmon et al., 2020). We further show that Smoothed-AGDA has $O(\kappa\Delta l\epsilon^{-2})$ complexity in finding an ϵ-stationary point of f. We can translate this point to an ϵ-stationary point of Φ after an additional negligible $\tilde{O}(\kappa)$ oracle complexity. This result improves the complexities of existing single-loop algorithms that require the more restrictive assumption of strong-concavity in y (we refer to this class of function as NC-SC). A comparison of our results to existing complexity bounds is summarized in Table 1.

Stochastic setting. We show that Stoc-AGDA achieves a sample complexity of $O(\kappa^4\Delta l\epsilon^{-4})$ for both notions of stationary measures, without having to rely on the $O(\epsilon^{-2})$ batch size and Hessian Lipschitz assumption used in prior work. This is the first convergence result for stochastic NC-PL minimax optimization and is also optimal in terms of the dependency to ϵ. We further show that the stochastic Smoothed-AGDA (Stoc-Smoothed-AGDA) algorithm achieves the $O(\kappa^2\Delta l\epsilon^{-4})$ sample complexity in finding an ϵ stationary point of f or Φ for small ϵ. This result improves upon the state-of-the-art complexity $O(\kappa^3\Delta l\epsilon^{-4})$ for NC-SC problems, which is a subclass of the NC-PL family. We refer the reader to Table 2 for a comparison.

1.2 Related Work

PL conditions in minimax optimization. In the deterministic NC-PL setting, Yang et al. (2020a) and Nouiehed et al. (2019) show that AGDA and its multistep variant, which applies multiple updates in y after one update of x, can find an approximate stationary point within $O(\kappa^2\epsilon^{-2})$ and $\tilde{O}(\kappa^2\epsilon^{-2})$ iterations, respectively. Recently, Fiez et al. (2021) showed that GDA converges asymptotically to a differential Stackelberg equilibrium and establish a local convergence rate of $O(\epsilon^{-2})$ for deterministic problems. In comparison, our work establishes non-asymptotic convergence to an ϵ-stationary point regardless of the starting point in both deterministic and stochastic settings, and we also focus on reducing the dependence to the condition number. Xie et al. (2021) consider NC-PL problems in the federated learning setting, showing $O(\epsilon^{-3})$ communication complexity when each client's objective is Lipschitz smooth. Moreover, there are a few work that aim to find global solutions by further imposing PL condition in x (Yang et al., 2020a; Guo et al., 2020a,b).

NC-SC minimax optimization. NC-SC problems are a subclass of NC-PL family. In the deterministic setting, GDA-type algorithms has been shown to have $O(\kappa^2\epsilon^{-2})$ iteration complexity (Lin et al., 2020a; Xu et al., 2020b; Boţ and Böhm, 2020; Lu et al., 2020). Later, Lin et al. (2020b) and Zhang et al. (2021c) improve this to $\tilde{O}(\sqrt{\kappa}\epsilon^{-2})$ by utilizing a proximal point method and Nesterov acceleration, and Zhang et al. (2021c) and Han et al. (2021) develop a tight lower complexity bound of $\Omega(\sqrt{\kappa}\epsilon^{-2})$. Yan et al. (2020) introduce Epoch-GDA for weakly-convex-strongly-concave problems. Comparatively, there are less studies in the stochastic setting. Recently, Chen et al. (2021a) extend their analysis from bilevel optimization to minimax optimization and show $O(\kappa^3\epsilon^{-4})$ sample complexity for an algorithm called ALSET without the $O(\epsilon^{-2})$ batch size required in (Lin et al., 2020a). ALSET reduces to AGDA in minimax optimization when it only does one step of y update in the inner loop. Guo et al. (2021) utilize stochastic moving-average estimator to nonconvex optimization and their algorithm PDSM achieves the same complexity for NC-SC minimax problems. We also refer the reader to the increasing body of bilevel optimization literature; e.g. (Guo and Yang, 2021; Ji et al., 2020; Hong et al., 2020; Chen et al., 2021b; Zhang, 2021). Also, Luo et al. (2020), Huang and Huang (2021) and Tran-Dinh et al. (2020) explore variance-reduced algorithms in this setting under the averaged smoothness assumption. Concurrently, Fiez et al. (2021) prove perturbed GDA converges to ϵ–local minimax equilibria with complexities of $\tilde{O}(\epsilon^{-4})$ and $\tilde{O}(\epsilon^{-2})$ in stochastic and deterministic problems, respectively, under additional second-order conditions. Notably, Li et al. (2021) develop the lower complexity bound of $\Omega\left(\sqrt{\kappa}\epsilon^{-2}+\kappa^{1/3}\epsilon^{-4}\right)$ for the stochastic setting. Other than first-order algorithms, there are a few explorations of zero-order methods (Xu et al., 2021; Huang et al., 2020; Xu et al., 2020a; Wang et al., 2020; Liu et al., 2020b; Anagnostidis et al., 2021) and second-order methods (Luo and Chen, 2021; Chen and Zhou, 2021). All the results above hold in the NC-SC regime, while the PL condition is significantly weaker than strong-concavity as it lies in the nonconvex regime.

Other nonconvex minimax optimization. There is a line of work focusing on the setting where the objective is (non-strongly) concave about y, but achieves slower convergence than NC-SC minimax optimization for both general deterministic and stochastic problems (Zhao, 2020; Thekumparampil et al., 2019; Ostrovskii et al., 2021b; Rafique et al., 2021). For nonconvex-nocnoncave (NC-NC) problems, different notions of local optimal solutions as well as their properties have been investigated in (Mangoubi and Vishnoi, 2021; Jin et al., 2020; Fiez and Ratliff, 2020; Ratliff et al., 2013, 2016). At the same time, many works have studied the relations between the stable limit points of the algorithms and local solutions (Daskalakis and Panageas, 2018; Mazumdar et al., 2020; Fiez and Ratliff, 2020). Realizing the hardness in finding an approximate stationary point, see e.g. (Daskalakis et al., 2021; Hsieh et al., 2021; Letcher, 2020; Wang et al., 2019), some research works then turned to identifying the conditions required for convergence (Grimmer et al., 2020; Lu, 2021; Abernethy et al., 2021). One of the widely explored conditions among them is the existence of solution to Minty variational inequality (MVI), or its approximate condition (Diakonikolas et al., 2021; Liu et al., 2021, 2019; Malitsky, 2020; Mertikopoulos et al., 2018; Song et al., 2020; Zhou et al., 2017). Recently, Ostrovskii et al. (2021a) study the nonconvex-nonconcave minimax optimization when the domain of y is small. Loizou et al. (2021) study the sub-linear convergence of Stoc-GDA under expected co-coercivity.

2 PRELIMINARIES

Notations. Throughout the paper, we let $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$ denote the ℓ_2 (Euclidean) norm and $\langle\cdot,\cdot\rangle$ denote the inner product. For non-negative functions $f(x)$ and $g(x)$, we write $f = O(g)$ if $f(x) \le cg(x)$ for some $c > 0$, and $f = \tilde{O}(g)$ to omit poly-logarithmic terms. We define the primal-dual gap of a function $f(\cdot,\cdot)$ at a point $(\hat{x},\hat{y})$ as $\text{gap}_f(\hat{x},\hat{y}) \triangleq \max_{y\in\mathbb{R}^{d_2}} f(\hat{x},y) - \min_{x\in\mathbb{R}^{d_1}} f(x,\hat{y})$.

We are interested in minimax problems of the form:

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} f(x,y) \triangleq \mathbb{E}[F(x,y;\xi)], \tag{2}$$

where ξ is a random vector with support Ξ, and f is possibly nonconvex-nonconcave. We now present the main setting considered in this paper.

Assumption 2.1 (Lipschitz Smooth) *The function f is differentiable and there exists a positive constant l such that*

$$\|\nabla_x f(x_1,y_1) - \nabla_x f(x_2,y_2)\| \le l[\|x_1-x_2\| + \|y_1-y_2\|],$$
$$\|\nabla_y f(x_1,y_1) - \nabla_y f(x_2,y_2)\| \le l[\|x_1-x_2\| + \|y_1-y_2\|],$$

holds for all $x_1,x_2\in\mathbb{R}^{d_1}, y_1,\ y_2\in\mathbb{R}^{d_2}$.

Assumption 2.2 (PL Condition in y) *For any fixed x, $\max_{y\in\mathbb{R}^{d_2}} f(x,y)$ has a nonempty solution set and a finite optimal value. There exists $\mu > 0$ such that: $\|\nabla_y f(x,y)\|^2 \ge 2\mu[\max_y f(x,y) - f(x,y)], \forall x,y$.*

The PL condition was originally introduced in (Polyak, 1963) who showed that it guarantees global convergence of gradient descent at a linear rate. This condition is shown in (Karimi et al., 2016) to be weaker than strong convexity as well as other conditions under which gradient descent converges linearly. The PL condition has also drawn much attention recently as it was shown to hold for various non-convex applications of interest in machine learning (Fazel et al., 2018; Cai et al., 2019), including problems related to deep neural networks (Du et al., 2019; Liu et al., 2020a). In this work, we assume that the objective function f in (2) is Lipschitz smooth and satisfies the PL condition about the dual variable y, i.e. Assumption 2.1 and 2.2, which is the same setting as in (Nouiehed et al., 2019) and

(Yang et al., 2020b) (Appendix D). However, to the best of our knowledge, stochastic algorithms have not yet been studied under such a setting.

From now on, we will define $\Phi(x) \triangleq \max_y f(x,y)$ as the primal function and $\kappa \triangleq \frac{l}{\mu}$ as the condition number. We will assume that $\Phi(\cdot)$ is lower bounded by a finite Φ^*. According to (Nouiehed et al., 2019), $\Phi(\cdot)$ is $2\kappa l$-lipschitz smooth with Assumption 2.1 and 2.2. There are two popular and natural notions of stationarity for minimax optimization in the form of (2): one is measured with ∇f and the other is measured with $\nabla\Phi$. We give the formal definitions below.

Definition 2.1 (Stationarity Measures)

a). *$(\hat{x},\hat{y})$ is an (ϵ_1,ϵ_2)-stationary point of a differentiable function $f(\cdot,\cdot)$ if $\|\nabla_x f(\hat{x},\hat{y})\| \le \epsilon_1$ and $\|\nabla_y f(\hat{x},\hat{y})\| \le \epsilon_2$. If $(\hat{x},\hat{y})$ is an (ϵ,ϵ)-stationary point, we call it ϵ-stationary point for simplicity.*

b). *$\hat{x}$ is an ϵ-stationary point of a differentiable Φ if $\|\nabla\Phi(\hat{x})\| \le \epsilon$.*

These two notions can be translated to each other by the following proposition.

Proposition 2.1 (Translations)

a). *Under Assumptions 2.1 and 2.2, if $\hat{x}$ is an ϵ-stationary point of Φ and $\|\nabla_y f(\hat{x},\tilde{y})\| \le \epsilon'$, then we can find another $\hat{y}$ by maximizing $f(\hat{x},\cdot)$ from the initial point $\tilde{y}$ with (stochastic) gradient ascent such that $(\hat{x},\hat{y})$ is an $O(\epsilon)$-stationary point of f, which requires $O\left(\kappa\log\left(\frac{\kappa\epsilon'}{\epsilon}\right)\right)$ gradients or $\tilde{O}\left(\kappa+\kappa^3\sigma^2\epsilon^{-2}\right)$ stochastic gradients.*

b). *Under Assumptions 2.1 and 2.2, if $(\tilde{x},\tilde{y})$ is an $(\epsilon,\epsilon/\sqrt{\kappa})$-stationary point of f, then we can find an $O(\epsilon)$-stationary point of Φ by approximately solving $\min_x\max_y f(x,y) + l\|x-\tilde{x}\|^2$ from the initial point $(\tilde{x},\tilde{y})$ with (stochastic) AGDA, which requires $O(\kappa\log(\kappa))$ gradients or $\tilde{O}\left(\kappa+\kappa^5\sigma^2\epsilon^{-2}\right)$ stochastic gradients* [1].

Remark 2.1 *The proposition implies that we can convert an ϵ-stationary point of Φ to an ϵ-stationary point of f and an $(\epsilon,\epsilon/\sqrt{\kappa})$-stationary point of f to an ϵ-stationary point of Φ, at a low cost in $1/\epsilon$ dependency compared to the complexity of finding the stationary point of either notion. Therefore, we consider the stationarity of Φ which is a slightly stronger notion than the stationarity of f. Lin et al. (2020a) establish the similar conversion under the NC-SC setting, but it requires an (ϵ/κ)-stationary point of f to find an ϵ-stationary point of Φ. Later we will use this proposition to establish the stationary convergence for some algorithms.*

Finally, we assume to have access to unbiased stochastic gradients of f with bounded variance.

Assumption 2.3 (Stochastic Gradients) *$G_x(x,y,\xi)$ and $G_y(x,y,\xi)$ are unbiased stochastic estimators of $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$ and have variances bounded by $\sigma^2 > 0$.*

3 STOCHASTIC AGDA

Algorithm 1 Stoc-AGDA

1: Input: (x_0,y_0), step sizes $\tau_1 > 0, \tau_2 > 0$
2: **for all** $t = 0,1,2,...,T-1$ **do**
3: Draw two i.i.d. samples ξ_1^t, ξ_2^t
4: $x_{t+1} \leftarrow x_t - \tau_1 G_x(x_t,y_t,\xi_1^t)$
5: $y_{t+1} \leftarrow y_t + \tau_2 G_y(x_{t+1},y_t,\xi_2^t)$
6: **end for**
7: Output: choose $(\hat{x},\hat{y})$ uniformly from $\{(x_t,y_t)\}_{t=0}^{T-1}$

Stochastic alternating gradient descent ascent (Stoc-AGDA) presented in Algorithm 1 sequentially updates primal and dual variables with simple stochastic gradient descent/ascent. In each iteration, only two samples are drawn to evaluate stochastic gradients. Here τ_1 and τ_2 denote the stepsize of x and y, respectively, and they can be very different.

Theorem 3.1 *Under Assumptions 2.1, 2.2 and 2.3, if we apply Stoc-AGDA with stepsizes $\tau_1 = \min\left\{\frac{\sqrt{\Delta}}{4\sigma\kappa^2\sqrt{Tl}}, \frac{1}{68l\kappa^2}\right\}$ and $\tau_2 = \min\left\{\frac{17\sqrt{\Delta}}{\sigma\sqrt{Tl}}, \frac{1}{l}\right\}$, then we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\Phi(x_t)\|^2 \le \frac{1088l\kappa^2}{T}\Delta + \frac{136l\kappa^2}{T}a_0 + \frac{8\kappa^2\sqrt{l}a_0}{\sqrt{\Delta T}}\sigma + \frac{1232\kappa^2\sqrt{l\Delta}}{\sqrt{T}}\sigma,$$

where $\Delta = \Phi(x_0) - \Phi^*$ and $a_0 := \Phi(x_0) - f(x_0,y_0)$. This implies a sample complexity of $O\left(\frac{l\kappa^2\Delta}{\epsilon^2} + \frac{l\kappa^4\Delta\sigma^2}{\epsilon^4}\right)$ to find an ϵ-stationary point of Φ.

We can either use Proposition 2.1 to translate to the other notion with extra computations or show that Stoc-AGDA directly outputs an ϵ-stationary point of f with the same sample complexity.

[1] In the proof of Proposition 2.1 in Supplement A, we also show that an an $(\epsilon,\epsilon/\sqrt{\kappa})$-stationary point of f, $(\tilde{x},\tilde{y})$, satisfies $\|\nabla\Phi_{1/2l}(\tilde{x})\|^2 \le 2\sqrt{2}\epsilon$, where $\Phi_{1/2l}$ is Moreau envelope with parameter $1/2l$.

Corollary 3.1 *Under the same setting as Theorem 3.1, the output $(\hat{x}, \hat{y})$ from Stoc-AGDA satisfies $\mathbb{E}\|\nabla_x f(\hat{x}, \hat{y})\| \leq \epsilon$ and $\mathbb{E}\|\nabla_y f(\hat{x}, \hat{y})\| \leq \epsilon$ after $O\left(\frac{l\kappa^2\Delta}{\epsilon^2} + \frac{l\kappa^4\Delta\sigma^2}{\epsilon^4}\right)$ iterations, which implies the same sample complexity as Theorem 3.1.*

Remark 3.1 *The dependency on $a_0 = \Phi(x_0) - f(x_0, y_0)$ can be improved by initializing y_0 with gradient ascent or stochastic gradient ascent to maximize the function $f(x_0, \cdot)$ satisfying the PL condition, which has exponential convergence in the deterministic setting and $O(\frac{1}{T})$ sublinear rate in the stochastic setting (Karimi et al., 2016).*

Remark 3.2 *The complexity above has different dependency as a function of ϵ and κ for the terms with and without the variance term σ. When $\sigma = 0$, the output from AGDA after $O\left(l\kappa^2\Delta\epsilon^{-2}\right)$ iterations will be an ϵ-stationary point of both f and Φ. It recovers the same complexity result in (Yang et al., 2020b) for the primal function stationary convergence. Nouiehed et al. (2019) show the same complexity for multi-GDA based on the stationary measure of f, which implies $O(l\kappa^3\Delta\epsilon^{-2})$ complexity for the stationary convergence of Φ by Proposition 2.1. See Table 1 for more comparisons.*

Remark 3.3 *When $\sigma > 0$, we establish the new sample complexity of $O(l\kappa^4\Delta\epsilon^{-4})$ for Stoc-AGDA. It is the first analysis of stochastic algorithms for NC-PL minimax problems. The dependency on ϵ is optimal, because the lower complexity bound of $\Omega(\epsilon^{-4})$ for stochastic nonconvex optimization (Arjevani et al., 2019) still holds when considering $f(x, y) = F(x)$ for some nonconvex function $F(x)$. Even under the strictly stronger assumption of imposing strong-concavity in y, to the best of our knowledge, it is the first time that a vanilla stochastic GDA-type algorithm is showed to achieve $O(\epsilon^{-4})$ sample complexity without either increasing batch size as in (Lin et al., 2020a) or Lipschitz continuity of $f(\cdot, y)$ and its Hessian as in (Chen et al., 2021a). In (Lin et al., 2020a), they show a worse complexity of $O(\epsilon^{-5})$ for GDA with $O(1)$ batch size. We refer the reader to Table 2.*

Remark 3.4 *We point out that under our weaker assumption, the dependency on the condition number κ is slightly worse than that in (Lin et al., 2020a; Chen et al., 2021a). If only $O(1)$ samples are available in each iteration, Stoc-GDA only achieves $O(\epsilon^{-5})$ sample complexity (Lin et al., 2020a). On the other hand, the analysis in (Chen et al., 2021b) is not applicable here. It uses a potential function $V_t = \Phi(x_t) + O(\mu)\|y_t - y^*(x_t)\|^2]$, where $y^*(x_t) = \arg\max_y f(x, y)$. To show a descent lemma for $\mathbb{E}[V_t]$, it shows the Lipschitz smoothness of $y^*(\cdot)$, which heavily depends on Lipschtiz continuity of f and its hessian. Under the PL condition, $y^*(x)$ might not be unique and we no longer make additional Lipschitz assumptions. Instead, we present an analysis based on the potential function $V_t = \Phi(x_t) + O(1)[\Phi(x_t) - f(x_t, y_t)]$ (see Supplement B).*

4 STOCHASTIC SMOOTHED-AGDA

Algorithm 2 Stochastic Smoothed-AGDA

1: Input: (x_0, y_0, z_0), step sizes $\tau_1 > 0, \tau_2 > 0$
2: **for all** $t = 0, 1, 2, ..., T-1$ **do**
3: Draw two i.i.d. samples ξ_1^t, ξ_2^t
4: $x_{t+1} = x_t - \tau_1[G_x(x_t, y_t, \xi_1^t) + p(x_t - z_t)]$
5: $y_{t+1} = y_t + \tau_2 G_y(x_{t+1}, y_t, \xi_2^t)$
6: $z_{t+1} = z_t + \beta(x_{t+1} - z_t)$
7: **end for**
8: Output: choose $(\hat{x}, \hat{y})$ uniformly from $\{(x_t, y_t)\}_{t=0}^{T-1}$

Stochastic Smoothed-AGDA presented in Algorithm 2 is closely related to proximal point method (PPM) on the primal function $\Phi(\cdot)$. In each iteration, we consider solving an auxiliary problem: $\min_x \Phi(x) + \frac{p}{2}\|x - z_t\|^2$, which is equivalent to:

$$\min_x \max_y \hat{f}(x, y; z_t) \triangleq f(x, y) + \frac{p}{2}\|x - z_t\|^2,$$

where z_t is called a proximal center to be defined later. Recently, proximal type algorithms including Catalyst have been shown to efficiently accelerate minimax optimization (Lin et al., 2020b; Yang et al., 2020b; Zhang et al., 2021c; Luo et al., 2021). While these algorithms require multiple loops to solve the auxiliary problem to some high accuracy[2], Stoc-Smoothed-AGDA only applies one step of Stoc-AGDA to solve it from the point (x_t, y_t) as in step 4 and 5. Step 6 in Algorithm 2 with some $\beta \in (0, 1)$ guarantees that the proximal point z_t in the auxiliary problem is not too far from the previous one z_{t-1}. Smoothed-AGDA was first introduced by Zhang et al. (2020) in the deterministic nonconvex-concave minimax optimization. To the best of our knowledge, its convergence has not been discussed in either the stochastic or the NC-PL setting.

Stoc-Smoothed-AGDA still maintains the single-loop structure and uses only $O(1)$ samples in each iteration. If we choose $\beta = 1$ or $p = 0$, it reduces to Stoc-AGDA. Later in the analysis, we choose $p = 2l$ so that the

[2]In Supplement D, we present a two-loop Catalyst algorithm combined with AGDA (Catalyst-AGDA) that achieves the same complexity as Algorithm 2 in the deterministic setting.

auxiliary problem is l-strongly convex in x. We will see in the next theorem that this quadratic regularization term enables Smoothed-AGDA to take larger stepsizes for x compared to AGDA. In Smoothed-AGDA, the ratio between stepsize of x and y is $\Theta(1)$[3], while this ratio is $\Theta(1/\kappa^2)$ in AGDA.

Theorem 4.1 *Under Assumptions 2.1, 2.2 and 2.3, if we apply Algorithm 2 with* $\tau_1 = \min\left\{\frac{\sqrt{\Delta}}{2\sigma\sqrt{Tl}}, \frac{1}{3l}\right\}$, $\tau_2 = \min\left\{\frac{\sqrt{\Delta}}{96\sigma\sqrt{Tl}}, \frac{1}{144l}\right\}$, $p = 2l$ *and* $\beta = \frac{\tau_2\mu}{1600}$, *then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\{\left\|\nabla_x f(x_t,y_t)\right\|^2 + \kappa\left\|\nabla_y f(x_t,y_t)\right\|^2\right\} \leq \frac{c_0 l\kappa}{T}[\Delta + b_0] + \frac{c_1\kappa\sqrt{l}b_0}{\sqrt{\Delta T}}\sigma + \frac{c_2\kappa\sqrt{l\Delta}}{\sqrt{T}}\sigma,$$

where $\Delta = \Phi(z_0) - \Phi^*$ *and* $b_0 = 2\,\mathrm{gap}_{\hat{f}(\cdot,\cdot;z_0)}(x_0,y_0)$ *is the primal-dual gap of the first auxiliary function at the initial point, and* c_0, c_1 *and* c_2 *are* $O(1)$ *constants. This implies the sample complexity of* $O\left(\frac{l\kappa\Delta}{\epsilon^2} + \frac{l\kappa^2\Delta\sigma^2}{\epsilon^4}\right)$ *to find an* $(\epsilon, \epsilon/\sqrt{\kappa})$*-stationary point of* f.

Remark 4.1 *In the theorem above,* b_0 *measures the optimality of* (x_0, y_0) *in the first auxiliary problem:* $\min_x \max_y f(x,y) + l\|x - z_0\|^2$, *which is* l*-strongly convex about* x *and* μ*-PL about* y*. Therefore, the dependency on* b_0 *can be reduced if we initialize* (x_0, y_0) *by approximately solving the first auxiliary problem with (Stochastic) AGDA, which converges exponentially in the deterministic setting and sublinearly at* $O(1/T)$ *rate in the stochastic setting for strongly-convex-PL minimax optimization (Yang et al., 2020a).*

By Proposition 2.1 b), after we find an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f from Stoc-Smoothed-AGDA, we can convert it to an $O(\epsilon)$-stationary point of Φ.

Corollary 4.1 *From the output* $(\hat{x}, \hat{y})$ *of stochastic Smoothed-AGDA, we can apply (stochastic) AGDA to find an* $O(\epsilon)$*-stationary point of* Φ *by approximately solving* $\min_x \max_y f(x,y) + l\|x - \hat{x}\|^2$ *as in Proposition 2.1. Therefore, the total complexity is* $O\left(\frac{l\kappa\Delta}{\epsilon^2}\right)$ *in the deterministic setting and* $\tilde{O}\left(\frac{l\kappa\Delta}{\epsilon^2} + \frac{l\kappa^2\Delta\sigma^2}{\epsilon^4} + \frac{\kappa^5\sigma^2}{\epsilon^2}\right)$ *in the stochastic setting.*

Remark 4.2 *In the deterministic setting, the translation cost is* $\kappa\log(\kappa)$*, which is dominated by the complexity of finding* $(\epsilon, \epsilon/\sqrt{\kappa})$*-stationary point of* f *in Theorem 4.1. In the stochastic setting, the extra translation cost* $\tilde{O}\left(\frac{\kappa^5\sigma^2}{\epsilon^2}\right)$ *is low in the dependency of* $\frac{1}{\epsilon}$ *but larger in terms of the condition number. In practice, the inverse of the target accuracy is usually large. We leave the question of reducing translation cost and whether Stocastic Smoothed-AGDA can directly output an approximate stationary point of* Φ *to future research.*

Remark 4.3 *The term without variance* σ *has better dependency on* ϵ *and* κ *than the term with* σ*. In the deterministic setting, Smoothed-AGDA achieves the complexity of* $O(l\kappa\Delta\epsilon^{-2})$*, which improves over AGDA (Yang et al., 2020a) and Multi-AGDA (Nouiehed et al., 2019) with either notion of stationarity. Notably, this complexity under our weaker assumptions is better than that of other single-loop algorithms under a stronger assumption of strong-concavity in* y *(see Table 2). Recently, Zhang et al. (2021c) provide a tight lower bound of* $O(l\sqrt{\kappa}\Delta\epsilon^{-2})$ *for deterministic NC-SC minimax optimization. However, we do not expect the same complexity can be achieved under weaker assumptions.*

Remark 4.4 *In the stochastic setting, we show that Stoc-Smoothed-AGDA achieves a sample complexity of* $O(l\kappa^2\Delta\epsilon^{-4})$ *in finding an* ϵ*-stationary point of* f*. To find an* ϵ*-stationary point of* Φ*, it bears an additional complexity of* $O(\kappa^5\sigma^2\epsilon^{-2})$*, which is negligible as long as* ϵ *is asymptotically small, i.e. when* $\epsilon \leq \tilde{O}(\sqrt{\Delta/l\kappa^3})$*. This sample complexity improves over* $O(l\kappa^4\Delta\epsilon^{-4})$ *sample complexity of Stoc-AGDA in NC-PL setting, and even* $O(l\kappa^3\Delta\epsilon^{-4})$ *complexity of Stoc-GDA (Lin et al., 2020a), ALSET (Chen et al., 2021a) and PDSM (Guo et al., 2021) in NC-SC setting. Moreover, this sample complexity improvement comes without any large batch size, additional Lipschitz assumptions, or multi-loop structure. Very recently, Li et al. (2021) develop the lower complexity bound of* $\Omega\left(\sqrt{\kappa}\epsilon^{-2} + \kappa^{1/3}\epsilon^{-4}\right)$ *in NC-SC setting, but there is no matching upper bound yet.*

5 EXPERIMENTS

We illustrate the effectiveness of stochastic AGDA (Algorithm 1) and stochastic Smoothed-AGDA (Algorithm 2) for solving NC-PL min-max problems. In particular, we show that the smoothed version of stochastic AGDA can compete with state-of-the-art deep learning optimizers [4].

Toy WGAN with linear generator. We consider the same setting as (Loizou et al., 2020), i.e. using a Wasserstein GAN (Arjovsky et al., 2017) to approximate a one-dimensional Gaussian distribution. In particular, we have a dataset of real data x^{real}

[3] In Supplement D, we show Catalyst-AGDA takes the stepsizes of the same order in the deterministic setting.

[4] Code available at `https://github.com/aorvieto/NCPL.git`

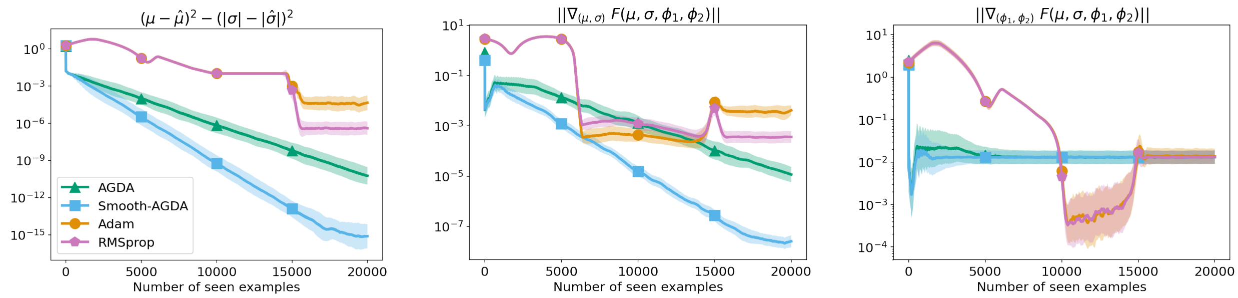

Figure 1: Training of a toy regularized WGAN with linear generator. Shown is the evolution of the *stochastic* gradients norm and the distance to the optimum. All methods are tuned at best for a minibatch size of 100, and each experiment is repeated 5 times (1 std shown). For Adam and RMSprop, we tuned over 4 learning rates ($1e-4, 5e-4, 1e-3, 5e-3$) and 2 momentum parameters $0.5, 0.9$. The optimal configuration is obtained for a stepsize of $5e-4$ and momentum 0.5. For stochastic AGDA we considered each combination of $\tau_1, \tau_2 \in \{1e-2, 5e-2, 1e-1, 5e-1, 1\}$. The optimal configuration was found to be $\tau_1 = 5e-1, \tau_2 = 1$. For stochastic Smoothed-AGDA we use $\beta = 0.9$, $p = 10$ and tuned it to best: $\tau_1 = 5e-1, \tau_2 = 5e-1$.

and latent variable z from a normal distribution with mean 0 and variance 1. The generator is defined as $G_{\mu,\sigma}(z) = \mu + \sigma z$ and the discriminator (a.k.a the critic) as $D_\phi(x) = \phi_1 x + \phi_2 x^2$, where x is either real data or fake data from the generator. The true data is generated from $\hat{\mu} = 0, \hat{\sigma} = 0.1$. The problem can be written in the form of:

$$\min_{\mu,\sigma} \max_{\phi_1,\phi_2} \; f(\mu, \sigma, \phi_1, \phi_2) \triangleq \mathbb{E}_{(x^{real},z)\sim\mathcal{D}} \; D_\phi(x^{real}) - D_\phi(G_{\mu,\sigma}(z)) - \lambda\|\phi\|^2,$$

where $\mathcal{D}$ is the distribution for the real data and latent variable, and the regularization $\lambda\|\phi\|^2$ with $\lambda = 0.001$ makes the problem strongly concave. This problem is non-convex in σ: indeed since z is symmetric around zero, both σ and $-\sigma$ are solutions. We fixed the batch size to 100 and tuned each algorithm at best (see plots in the appendix). Each experiment is repeated for 3 times. In Figure 1 we provide evidence of the superiority of Stoc-Smoothed-AGDA over Stoc-AGDA, Adam (Kingma and Ba, 2014) and RMSprop (Tieleman et al., 2012). As the reader can notice, Stoc-Smoothed-AGDA is competitive with fine-tuned popular adaptive methods, and provides a significant speedup over AGDA with carefully tuned learning rates, which verifies our theoretical results.

Toy WGAN with neural generator. Inspired by (Lei et al., 2020), we consider a regularized WGAN with a neural network as generator. For ease of comparison, we leave all the problem settings identical to last paragraph, and only change the generator $G_{\mu,\sigma}$ to G_θ, where θ are the parameters of a small neural network (one hidden layer with five neurons and ReLU activations). After careful tuning for each algorithm, we observe from Figure 2 that Stoc-Smoothed-AGDA still performs significantly better than vanilla Stoc-AGDA and Adam in this setting. The adaptiveness (without momentum) of RMSprop is able to yield slightly better results. This is not surprising, as adaptive methods are the de facto optimizers of choice in generative adversarial nets. Hence, a clear direction of future research is to combine adaptiveness and Smoothed-AGDA.

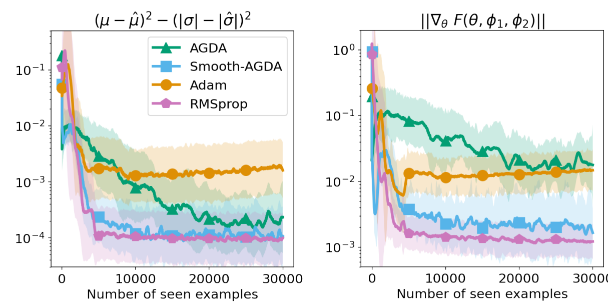

Figure 2: ReLU Network generator for a regularized WGAN (same settings as for Figure 1). Each algorithm is tuned to yield best performance, with a procedure similar to the one in Figure 1. The gradient with respect to the discriminator evolves very similarly to the last example, with fast convergence to a non-zero value.

Robust non-linear regression. The experiments above suggest that Smoothed-AGDA accelerates convergence of AGDA. We found that this holds true also outside the WGAN setting: in this last paragraph, we show how this accelerated behavior in a few robust regression problems. We first consider a synthetic dataset of 1000 datapoints z in 500 dimensions, sampled from a Gaussian distribution with mean zero and variance 1. The target values y_0 are sampled according to a random noisy linear model. We consider fitting this synthetic dataset with a two-hidden-layer ReLU network (256 units in the first layer, 64 in the second): $\text{net}_x(z)$ with x being the parameter. For the robustness part, we proceed in the standard way (see e.g.(Adolphs et al., 2019)) and add the concave objective $-\frac{\lambda}{2}\|y-y_0\|^2$ to the loss:

$$F(x,y)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\|\mathrm{net}_x(z)-y\|^2-\frac{\lambda}{2}\|y-y_0\|^2,$$

where we chose $\lambda=1$. In this experiement, we compare the performance of AGDA and Smoothed-AGDA under the same stepsize τ_1,τ_2. From Figure 3, we observe that Smoothed-AGDA has much faster convergence than AGDA both in the stochastic and deterministic setting (i.e. with full batch).

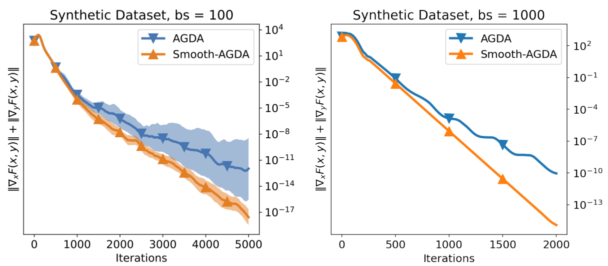

Figure 3: Robust non-linear regression on a synthetic Gaussian Dataset. Using $\tau_1=5e-4,\tau_2=5$ for both AGDA and Smoothed-AGDA, we notice a performance improvement for the latter using $\beta=0.5, p=10$.

6 Conclusion

In this paper, we established faster convergence rate for two *single-loop* algorithms under an assumption than is weaker than strong concavity. In particular, we showed that stochastic AGDA can achieve $O(\epsilon^{-4})$ sample complexity without having to rely on a large batch size. In addition, we established a better complexity in terms of the dependency to the condition number for Smooth AGDA in both stochastic and deterministic settings, which also improves over other single-loop algorithms for nonconvex-strongly-concave minimax optimization. There are several questions that are worth further investigation such as: (a) what is the lower complexity bound for optimization under the PL condition; (b) whether single-loop algorithms can always achieve a rate as fast as multi-loop algorithms; (c) how to design adaptive algorithms for minimax problems without strong concavity.

References

Abernethy, J., Lai, K. A., and Wibisono, A. (2021). Last-iterate convergence rates for min-max optimization: Convergence of hamiltonian gradient descent and consensus optimization. In *Algorithmic Learning Theory*, pages 3–47. PMLR.

Adolphs, L., Daneshmand, H., Lucchi, A., and Hofmann, T. (2019). Local saddle point optimization: A curvature exploitation approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 486–495. PMLR.

Anagnostidis, S.-K., Lucchi, A., and Diouane, Y. (2021). Direct-search for a class of stochastic minmax problems. In *International Conference on Artificial Intelligence and Statistics*, pages 3772–3780. PMLR.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2019). Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR.

Boţ, R. I. and Böhm, A. (2020). Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *arXiv preprint arXiv:2007.13605*.

Cai, Q., Hong, M., Chen, Y., and Wang, Z. (2019). On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv preprint arXiv:1901.03674*.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2020). Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120.

Chen, T., Sun, Y., and Yin, W. (2021a). Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Chen, T., Sun, Y., and Yin, W. (2021b). A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*.

Chen, Z. and Zhou, Y. (2021). Escaping saddle points in nonconvex minimax optimization via cubic-regularized gradient descent-ascent. *arXiv preprint arXiv:2110.07098*.

Daskalakis, C. and Panageas, I. (2018). The limit points of (optimistic) gradient descent in min-max optimization. *arXiv preprint arXiv:1807.03907*.

Daskalakis, C., Skoulakis, S., and Zampetakis, M. (2021). The complexity of constrained min-max optimization. *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*.

Diakonikolas, J., Daskalakis, C., and Jordan, M. (2021). Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR.

Drusvyatskiy, D. and Paquette, C. (2019). Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR.

Fiez, T. and Ratliff, L. J. (2020). Local convergence analysis of gradient descent ascent with finite timescale separation. In *International Conference on Learning Representations*.

Fiez, T., Ratliff, L. J., Mazumdar, E., Faulkner, E., and Narang, A. (2021). Global convergence to local minmax equilibrium in classes of nonconvex zero-sum games. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014b). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Grimmer, B., Lu, H., Worah, P., and Mirrokni, V. (2020). The landscape of the proximal point method for nonconvex-nonconcave minimax optimization. *arXiv preprint arXiv:2006.08667*.

Guo, Z., Liu, M., Yuan, Z., Shen, L., Liu, W., and Yang, T. (2020a). Communication-efficient distributed stochastic auc maximization with deep neural networks. In *International Conference on Machine Learning*, pages 3864–3874. PMLR.

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. (2021). On stochastic moving-average estimators for non-convex optimization. *arXiv preprint arXiv:2104.14840*.

Guo, Z. and Yang, T. (2021). Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*.

Guo, Z., Yuan, Z., Yan, Y., and Yang, T. (2020b). Fast objective and duality gap convergence for non-convex strongly-concave min-max problems. *arXiv preprint arXiv:2006.06889*.

Han, Y., Xie, G., and Zhang, Z. (2021). Lower complexity bounds of finite-sum optimization problems: The results and construction. *arXiv preprint arXiv:2103.08280*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Srocessing Systems*, 30.

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. (2020). A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*.

Hsieh, Y.-P., Mertikopoulos, P., and Cevher, V. (2021). The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *International Conference on Machine Learning*, pages 4337–4348. PMLR.

Huang, F., Gao, S., Pei, J., and Huang, H. (2020). Accelerated zeroth-order momentum methods from mini to minimax optimization. *arXiv e-prints*, pages arXiv–2008.

Huang, F. and Huang, H. (2021). Adagda: Faster adaptive gradient descent ascent methods for minimax optimization. *arXiv preprint arXiv:2106.16101*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*.

Ji, K., Yang, J., and Liang, Y. (2020). Bilevel optimization: Nonasymptotic analysis and faster algorithms. *arXiv preprint arXiv:2010.07962*.

Jin, C., Netrapalli, P., and Jordan, M. (2020). What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889. PMLR.

Karimi, H., Nutini, J., and Schmidt, M. (2016). 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*, pages 795–811. Springer.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. (2017). Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*.

Lei, Q., Lee, J., Dimakis, A., and Daskalakis, C. (2020). Sgd learns one-layer networks in wgans. In *International Conference on Machine Learning*, pages 5799–5808. PMLR.

Letcher, A. (2020). On the impossibility of global convergence in multi-loss optimization. *arXiv preprint arXiv:2005.12649*.

Li, H., Tian, Y., Zhang, J., and Jadbabaie, A. (2021). Complexity lower bounds for nonconvex-strongly-concave min-max optimization. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Lin, T., Jin, C., and Jordan, M. (2020a). On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR.

Lin, T., Jin, C., and Jordan, M. I. (2020b). Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR.

Liu, C., Zhu, L., and Belkin, M. (2020a). Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *arXiv preprint arXiv:2003.00307*.

Liu, M., Mroueh, Y., Ross, J., Zhang, W., Cui, X., Das, P., and Yang, T. (2019). Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *International Conference on Learning Representations*.

Liu, M., Rafique, H., Lin, Q., and Yang, T. (2021). First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22(169):1–34.

Liu, S., Lu, S., Chen, X., Feng, Y., Xu, K., Al-Dujaili, A., Hong, M., and O'Reilly, U.-M. (2020b). Min-max optimization without gradients: Convergence and applications to black-box evasion and poisoning attacks. In *International Conference on Machine Learning*, pages 6282–6293. PMLR.

Loizou, N., Berard, H., Gidel, G., Mitliagkas, I., and Lacoste-Julien, S. (2021). Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34.

Loizou, N., Berard, H., Jolicoeur-Martineau, A., Vincent, P., Lacoste-Julien, S., and Mitliagkas, I. (2020). Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381. PMLR.

Lu, H. (2021). An $o(s^r)$-resolution ode framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems. *Mathematical Programming*, pages 1–52.

Lu, S., Tsaknakis, I., Hong, M., and Chen, Y. (2020). Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691.

Luo, L. and Chen, C. (2021). Finding second-order stationary point for nonconvex-strongly-concave minimax problem. *arXiv preprint arXiv:2110.04814*.

Luo, L., Xie, G., Zhang, T., and Zhang, Z. (2021). Near optimal stochastic algorithms for finite-sum unbalanced convex-concave minimax optimization. *arXiv preprint arXiv:2106.01761*.

Luo, L., Ye, H., Huang, Z., and Zhang, T. (2020). Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33.

Malitsky, Y. (2020). Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1):383–410.

Mangoubi, O. and Vishnoi, N. K. (2021). Greedy adversarial equilibrium: an efficient alternative to nonconvex-nonconcave min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 896–909.

Mazumdar, E., Ratliff, L. J., and Sastry, S. S. (2020). On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131.

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. (2018). Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations*.

Mescheder, L., Geiger, A., and Nowozin, S. (2018). Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR.

Mescheder, L., Nowozin, S., and Geiger, A. (2017). The numerics of gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1823–1833.

Nagarajan, V. and Kolter, J. Z. (2017). Gradient descent gan optimization is locally stable. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5591–5600.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. (2019). Solving a class of nonconvex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32:14934–14942.

Ostrovskii, D. M., Barazandeh, B., and Razaviyayn, M. (2021a). Nonconvex-nonconcave min-max optimization with a small maximization domain. *arXiv preprint arXiv:2110.03950*.

Ostrovskii, D. M., Lowy, A., and Razaviyayn, M. (2021b). Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538.

Polyak, B. T. (1963). Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653.

Rafique, H., Liu, M., Lin, Q., and Yang, T. (2021). Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pages 1–35.

Ratliff, L. J., Burden, S. A., and Sastry, S. S. (2013). Characterization and computation of local nash equilibria in continuous games. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 917–924. IEEE.

Ratliff, L. J., Burden, S. A., and Sastry, S. S. (2016). On the characterization of local nash equilibria in continuous games. *IEEE transactions on automatic control*, 61(8):2301–2307.

Song, C., Zhou, Z., Zhou, Y., Jiang, Y., and Ma, Y. (2020). Optimistic dual extrapolation for coherent non-monotone variational inequalities. *Advances in Neural Information Processing Systems*, 33.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. (2019). Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32:12680–12691.

Tieleman, T., Hinton, G., et al. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

Tran-Dinh, Q., Liu, D., and Nguyen, L. M. (2020). Hybrid variance-reduced sgd algorithms for minimax problems with nonconvex-linear function. In *Advances in Neural Information Processing Systems*.

Wang, Y., Zhang, G., and Ba, J. (2019). On solving minimax optimization locally: A follow-the-ridge approach. In *International Conference on Learning Representations*.

Wang, Z., Balasubramanian, K., Ma, S., and Razaviyayn, M. (2020). Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819*.

Xie, J., Zhang, C., Zhang, Y., Shen, Z., and Qian, H. (2021). A federated learning framework for nonconvex-pl minimax problems. *arXiv preprint arXiv:2105.14216*.

Xu, T., Wang, Z., Liang, Y., and Poor, H. V. (2020a). Gradient free minimax optimization: Variance reduction and faster convergence. *arXiv preprint arXiv:2006.09361*.

Xu, Z., Shen, J., Wang, Z., and Dai, Y. (2021). Zeroth-order alternating randomized gradient projection algorithms for general nonconvex-concave minimax problems. *arXiv preprint arXiv:2108.00473*.

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

Yan, Y., Xu, Y., Lin, Q., Liu, W., and Yang, T. (2020). Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33.

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

Yang, J., Zhang, S., Kiyavash, N., and He, N. (2020b). A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33.

Zhang, G., Wang, Y., Lessard, L., and Grosse, R. (2021a). Don't fix what ain't broke: Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. *arXiv preprint arXiv:2102.09468*.

Zhang, J., Xiao, P., Sun, R., and Luo, Z.-Q. (2020). A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *arXiv preprint arXiv:2010.15768*.

Zhang, K., Yang, Z., and Başar, T. (2021b). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384.

Zhang, L. (2021). Variance reduction for non-convex stochastic optimization: General analysis and new applications. Master's thesis, ETH Zurich.

Zhang, S., Yang, J., Guzmán, C., Kiyavash, N., and He, N. (2021c). The complexity of nonconvex-strongly-concave minimax optimization. *arXiv preprint arXiv:2103.15888*.

Zhao, R. (2020). A primal dual smoothing framework for max-structured nonconvex optimization. *arXiv preprint arXiv:2003.04375*.

Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S., and Glynn, P. W. (2017). Stochastic mirror descent in variationally coherent optimization problems. *Advances in Neural Information Processing Systems*, 30:7040–7049.

Supplementary Material: Faster Single-loop Algorithms for Minimax Optimization without Strong Concavity

A USEFUL LEMMAS

Lemma A.1 (Lemma B.2 (Lin et al., 2020b)) *Assume $f(\cdot, y)$ is μ_x-strongly convex for $\forall y \in \mathbb{R}^{d_2}$ and $f(x, \cdot)$ is μ_y-strongly concave for $\forall x \in \mathbb{R}^{d_1}$ (we will later refer to this as (μ_x, μ_y)-SC-SC)) and f is l-Lipschitz smooth. Then we have*

a) $y^*(x) = \arg\max_{y\in\mathbb{R}^{d_2}} f(x, y)$ is $\frac{l}{\mu_y}$-Lipschitz;

b) $\Phi(x) = \max_{y\in\mathbb{R}^{d_2}} f(x, y)$ is $\frac{2l^2}{\mu_y}$-Lipschitz smooth and μ_x-strongly convex with $\nabla\Phi(x) = \nabla_x f(x, y^*(x))$;

c) $x^*(y) = \arg\min_{x\in\mathbb{R}^{d_1}} f(x, y)$ is $\frac{l}{\mu_x}$-Lipschitz;

d) $\Psi(y) = \min_{x\in\mathbb{R}^{d_1}} f(x, y)$ is $\frac{2l^2}{\mu_x}$-Lipschitz smooth and μ_y-strongly concave with $\nabla\Psi(y) = \nabla_y f(x^*(y), y)$.

Lemma A.2 (Karimi et al. (2016)) *If $f(\cdot)$ is l-smooth and it satisfies PL condition with constant μ, i.e.*

$$\|\nabla f(x)\|^2 \geq 2\mu[f(x) - \min_x f(x)], \forall x,$$

then it also satisfies error bound (EB) condition with μ, i.e.

$$\|\nabla f(x)\| \geq \mu\|x_p - x\|, \forall x,$$

where x_p is the projection of x onto the optimal set, and it satisfies quadratic growth (QG) condition with μ, i.e.

$$f(x) - \min_x f(x) \geq \frac{\mu}{2}\|x_p - x\|^2, \forall x.$$

Lemma A.3 (Nouiehed et al. (2019)) *Under Assumption 2.1 and 2.2, define $\Phi(x) = \max_y f(x, y)$ then*

a) for any x_1, x_2, and $y^*(x_1) \in \operatorname{Arg}\max_y f(x_1, y)$, there exists some $y^*(x_2) \in \operatorname{Arg}\max_y f(x_2, y)$ such that

$$\|y_1^* - y_2^*\| \leq \frac{l}{2\mu}\|x_1 - x_2\|.$$

b) $\Phi(\cdot)$ is L-smooth with $L := l + \frac{l\kappa}{2}$ with $\kappa = \frac{l}{\mu}$ and $\nabla\Phi(x) = \nabla_x f(x, y^*(x))$ for any $y^*(x) \in \operatorname{Arg}\max_y f(x, y)$.

Now we present a Theorem adopted from (Yang et al., 2020a). Under the two-sided PL condition, it captures the convergence of AGDA with dual updated first[5]:

$$\begin{aligned} y^{k+1} &= y^k + \tau_2\nabla_y f(x^k, y^k), \\ x^{k+1} &= x^k - \tau_1\nabla_x f(x^k, y^{k+1}). \end{aligned} \tag{3}$$

[5]The update is equivalent to applying AGDA with primal variable update first to $\min_y \max_x -f(x, y)$, so its convergence is a direct result from (Yang et al., 2020a). We believe similar convergence rate to Theorem A.1 holds for AGDA with x update first. But for simplicity, here we consider update (3) without additional derivation.

Theorem A.1 (Yang et al. (2020a)) *Consider a minimax optimization problem under Assumption 2.3:*

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} f(x,y) \triangleq \mathbb{E}[F(x,y;\xi)].$$

Suppose the function f is l-smooth, $f(\cdot,y)$ satisfies the PL condition with constant μ_1 and $-f(x,\cdot)$ satisfies the PL condition with constant μ_2 for any x and y. Define

$$P_k = \mathbb{E}[\Psi^* - \Psi(y_t)] + \frac{1}{10}\mathbb{E}[f(x^k,y^k) - \Psi(x^k)]$$

with $\Psi(y) = \min_x f(x,y)$ and $\Psi^* = \max_y \Psi(y)$. If we run Stoc-AGDA (with update rule (3)) with stepsizes $\tau_1 \le \frac{1}{l}$ and $\tau_2 \le \frac{\mu_1^2\tau_1}{18l^2}$, then

$$P_k \le \left(1 - \frac{\mu_2\tau_2}{2}\right)^k P_0 + \frac{23l^2\tau_2^2/\mu_1 + l\tau_1^2}{10\mu_2\tau_2}\sigma^2. \tag{4}$$

In the deterministic setting, e.g. $\sigma = 0$, if we run AGDA with stepsizes $\tau_1 = \frac{1}{l}$ and $\tau_2 = \frac{\mu_1^2}{18l^3}$ then

$$P_k \le \left(1 - \frac{\mu_1^2\mu_2}{36l^3}\right)^k P_0. \tag{5}$$

Definition A.1 (Moreau Envelope) *The Moreau envelope of a function Φ with a parameter $\lambda > 0$ is:*

$$\Phi_\lambda(x) = \min_{z\in\mathbb{R}^{d_1}} \Phi(z) + \frac{1}{2\lambda}\|z - x\|^2.$$

The proximal point of x is defined as: $\mathrm{prox}_{\lambda\Phi}(x) = \arg\min_{z\in\mathbb{R}^{d_1}}\left\{\Phi(z) + \frac{1}{2\lambda}\|z-x\|^2\right\}$. The gradients of Φ and and Φ_λ are closely related by the following well-known lemma; see e.g. (Drusvyatskiy and Paquette, 2019).

Lemma A.4 *When F is differentiable and ℓ-Lipschitz smooth, for $\lambda \in (0, 1/\ell)$ we have $\nabla\Phi(\mathrm{prox}_{\lambda F}(x)) = \nabla\Phi_\lambda(x) = \lambda^{-1}(x - \mathrm{prox}_{\lambda\Phi}(x))$.*

Proof of Proposition 2.1

Proof We will prove Part (a) and (b) separately.
Part (a): If we can find $\hat{y}$ such that $\max_y f(\hat{x},y) - f(\hat{x},\hat{y}) \le \frac{\epsilon^2}{l\kappa}$, then as $\|\nabla_y f(\hat{x}, y^*(\hat{x}))\| = 0$,

$$\|\nabla_y f(\hat{x},\hat{y})\| \le \|\nabla_y f(\hat{x},\hat{y}) - \nabla_y f(\hat{x},y^*(\hat{x}))\| \le l\|\hat{y} - y^*(\hat{x})\| \le l\sqrt{\frac{2}{\mu}[\max_y f(\hat{x},y) - f(\hat{x},\hat{y})]} \le \sqrt{2}\epsilon,$$

where in the first inequality we fix $y^*(x)$ to the projection from $\hat{y}$ to $\mathrm{Arg}\max_y f(\hat{x},y)$, in the second inequality we use Lipschitz smoothness, and in the third inequality we use PL condition and Lemma A.2. Also,

$$\begin{aligned}\|\nabla_x f(\hat{x},\hat{y})\| \le& \|\nabla_x f(\hat{x},y^*(\hat{x}))\| + \|\nabla_x f(\hat{x},\hat{y}) - \nabla_x f(\hat{x},y^*(\hat{x}))\| \\ \le& \|\nabla\Phi(\hat{x})\| + l\|\hat{y} - y^*(\hat{x})\| \\ \le& \|\nabla\Phi(\hat{x})\| + l\sqrt{\frac{2}{\mu}[\max_y f(\hat{x},y) - f(\hat{x},\hat{y})]} \le (1+\sqrt{2})\epsilon,\end{aligned}$$

where in the second inequality we use Lemma A.3. Therefore, our goal is to find $\hat{y}$ such that $\max_y f(\hat{x},y) - f(\hat{x},\hat{y}) \le \frac{\epsilon^2}{l\kappa}$ by applying (stochastic) gradient ascent to $f(\hat{x},\cdot)$ from initial point $\tilde{y}$.

Deterministic case: Since $\|\nabla_y f(\hat{x},\tilde{y})\| \le \epsilon'$, we have $\max_y f(\hat{x},y) - f(\hat{x},\tilde{y}) \le \frac{\epsilon'^2}{2\mu}$ by PL condition. Let y^k denote k-th iterates of gradient ascent from initial point $\tilde{y}$ with stepsize $\frac{1}{l}$. Then by (Karimi et al., 2016)

$$\max_y f(\hat{x},y) - f(\hat{x},y^k) \le \left(1 - \frac{1}{\kappa}\right)^k \left[\max_y f(\hat{x},y) - f(\hat{x},\tilde{y})\right].$$

So after $O\left(\kappa \log\left(\frac{\kappa\epsilon'}{\epsilon}\right)\right)$, we can find the point we want.

Stochastic Case: Let y^k denote k-th iterates of stochastic gradient ascent from initial point $\tilde{y}$ with stepsize $\tau \leq \frac{1}{l}$. Then by Lemma A.4 in (Yang et al., 2020b)

$$\mathbb{E}\left[\max_y f(\hat{x}, y)) - f(\hat{x}, y^{k+1})\right] \leq (1-\mu\tau)\mathbb{E}\left[\max_y f(\hat{x}, y)) - f(\hat{x}, y^k)\right] + \frac{l\tau^2}{2}\sigma^2,$$

which implies

$$\mathbb{E}\left[\max_y f(\hat{x}, y)) - f(\hat{x}, y^k)\right] \leq (1-\mu\tau)^k\mathbb{E}\left[\max_y f(\hat{x}, y)) - f(\hat{x}, \tilde{y})\right] + \frac{\kappa\tau}{2}\sigma^2.$$

So with $\tau = \min\left\{\frac{1}{l}, \Theta\left(\frac{\epsilon^2}{l\kappa^2\sigma^2}\right)\right\}$, we can find the point we want with a complexity of $O\left(\kappa\log\left(\frac{\kappa\epsilon'}{\epsilon}\right) + \kappa^3\sigma^2\log\left(\frac{\kappa\epsilon'}{\epsilon}\right)\epsilon^{-2}\right)$.

Part (b): We first look at $\Phi_{1/2l}(\tilde{x}) = \min_z \Phi(z) + l\|z-\tilde{x}\|^2$. Then by Lemma 4.3 in (Drusvyatskiy and Paquette, 2019),

$$\begin{aligned}
&\|\nabla\Phi_{1/2l}(\tilde{x})\|^2 \\
=&4l^2\|\tilde{x} - \operatorname{prox}_{\Phi/2l}(\tilde{x})\|^2 \\
\leq&8l[\Phi(\tilde{x}) - \Phi(\operatorname{prox}_{\Phi/2l}(\tilde{x})) - l\|\operatorname{prox}_{\Phi/2l}(\tilde{x}) - \tilde{x}\|^2] \\
=&8l\left[\Phi(\tilde{x}) - f(\tilde{x}, \tilde{y}) + f(\tilde{x}, \tilde{y}) - f(\operatorname{prox}_{\Phi/2l}(\tilde{x}), \tilde{y}) + f(\operatorname{prox}_{\Phi/2l}(\tilde{x}), \tilde{y}) - \Phi(\operatorname{prox}_{\Phi/2l}(\tilde{x})) - l\|\operatorname{prox}_{\Phi/2l}(\tilde{x}) - \tilde{x}\|^2\right] \\
\leq&8l\left[\frac{1}{2\mu}\|\nabla_y f(\tilde{x}, \tilde{y})\|^2 + f(\tilde{x}, \tilde{y}) - f(\operatorname{prox}_{\Phi/2l}(\tilde{x}), \tilde{y}) - l\|\operatorname{prox}_{\Phi/2l}(\tilde{x}) - \tilde{x}\|^2\right]
\end{aligned} \tag{6}$$

where in the first inequality we use the l-strong-convexity in x of $\Phi(x) + l\|x - \tilde{x}\|^2$, in the second inequality we use $\Phi(\tilde{x}) - f(\tilde{x}, \tilde{y}) \leq \frac{1}{2\mu}\|\nabla_y f(\tilde{x}, \tilde{y})\|^2$ by PL condition, and $f(\operatorname{prox}_{\Phi/2l}(\tilde{x}), \tilde{y}) - \Phi(\operatorname{prox}_{\Phi/2l}(\tilde{x})) \leq 0$. Note that by defining $\hat{f}(x, y) = f(x, y) + l\|x - \tilde{x}\|^2$, we have

$$\begin{aligned}
f(\tilde{x}, \tilde{y}) - f(\operatorname{prox}_{\Phi/2l}(\tilde{x}), \tilde{y}) - l\|\operatorname{prox}_{\Phi/2l}(\tilde{x}) - \tilde{x}\|^2 =&\hat{f}(\tilde{x}, \tilde{y}) - \hat{f}(\operatorname{prox}_{\Phi/2l}(\tilde{x}), \tilde{y}) \\
\leq&\langle\nabla_x f(\tilde{x}, \tilde{y}), x - \operatorname{prox}_{\Phi/2l}(\tilde{x})\rangle - \frac{l}{2}\|x - \operatorname{prox}_{\Phi/2l}(\tilde{x})\|^2 \\
\leq&\frac{1}{2l}\|\nabla_x \hat{f}(\tilde{x}, \tilde{y})\|^2 + \frac{l}{2}\|x - \operatorname{prox}_{\Phi/2l}(\tilde{x})\|^2 - \frac{l}{2}\|x - \operatorname{prox}_{\Phi/2l}(\tilde{x})\|^2 \\
\leq&\frac{1}{2l}\|\nabla_x \hat{f}(\tilde{x}, \tilde{y})\|^2 = \frac{1}{2l}\|\nabla_x f(\tilde{x}, \tilde{y})\|^2,
\end{aligned}$$

where in the second inequality we use l-strong-convexity in x of $\hat{f}(x, y)$. Plugging into (6),

$$\|\nabla\Phi_{1/2l}(\tilde{x})\|^2 = 4l^2\|\tilde{x} - \operatorname{prox}_{\Phi/2l}(\tilde{x})\|^2 \leq 4\kappa\|\nabla_y f(\tilde{x}, \tilde{y})\|^2 + 4\|\nabla_x f(\tilde{x}, \tilde{y})\|^2 \leq 8\epsilon^2. \tag{7}$$

If we can find $\hat{x}$ such that $\|\operatorname{prox}_{\Phi/2l}(\tilde{x}) - \hat{x}\| \leq \frac{\epsilon}{\kappa l}$, then

$$\|\nabla\Phi(\hat{x})\| \leq \|\nabla\Phi(\operatorname{prox}_{\Phi/2l}(\tilde{x}))\| + \|\nabla\Phi(\hat{x}) - \nabla\Phi(\operatorname{prox}_{\Phi/2l}(\tilde{x}))\| \leq \|\nabla\Phi_{1/2l}(\tilde{x})\| + 2\kappa l\|\operatorname{prox}_{\Phi/2l}(\tilde{x}) - \hat{x}\| \leq (2\sqrt{2}+2)\epsilon,$$

where in the second inequality we use Lemma A.3 and Lemma A.4. Note that $\operatorname{prox}_{\Phi/2l}(\tilde{x})$ is the solution to $\min_x \Phi(x) + l\|x - \tilde{x}\|^2$, which is equivalent to

$$\min_x \max_y f(x, y) + l\|x - \tilde{x}\|^2. \tag{8}$$

This minimax problem is l-strongly convex about x, μ-PL about y and $3l$-smooth. Therefore, we can use (stochastic) alternating gradient descent ascent (AGDA) to find $\hat{x}$ such that $\|\operatorname{prox}_{\Phi/2l}(\tilde{x}) - \hat{x}\| \leq \frac{\epsilon}{\kappa l}$ from initial point $(\tilde{x}, \tilde{y})$.

Deterministic case: Let (x^k, y^k) denote k-th iterates of AGDA with y updated first from initial point $(\tilde{x}, \tilde{y})$ on function (8). Define $\hat{\Phi}(x) = \max_y \hat{f}(x,y) = \max_y f(x,y) + l\|x-\tilde{x}\|^2$, $\hat{\Psi}(y) = \min_x \hat{f}(x,y) = \min_x f(x,y) + l\|x-\tilde{x}\|^2$ and $\hat{\Psi}^* = \max_y \hat{\Psi}(y)$. We also denote $x^* = \arg\min_x \hat{\Phi}(x) = \operatorname{prox}_{\Phi/2l}(\tilde{x})$. Then we define $P_k = \hat{\Psi}^* - \hat{\Psi}(y^k) + \frac{1}{10}\left[\hat{f}(x^k,y^k) - \hat{\Psi}(y^k)\right]$. Note that

$$P_0 = \hat{\Psi}^* - \hat{\Psi}(\tilde{y}) + \frac{1}{10}\left[\hat{f}(\tilde{x},\tilde{y}) - \hat{\Psi}(\tilde{y})\right] \leq \hat{\Psi}^* - \hat{\Psi}(\tilde{y}) + \frac{1}{20l}\|\nabla_x \hat{f}(\tilde{x},\tilde{y})\|^2 \leq \hat{\Psi}^* - \hat{\Psi}(\tilde{y}) + \frac{\epsilon^2}{20l}. \tag{9}$$

Also we note that

$$\begin{aligned}
\hat{\Psi}^* - \hat{\Psi}(\tilde{y}) =& \max_y \min_x \hat{f}(x,y) - \min_x \hat{f}(x,\tilde{y}) \\
=& \max_y \min_x \hat{f}(x,y) - \max_y \hat{f}(\tilde{x},y) + \max_y \hat{f}(\tilde{x},y) - \hat{f}(\tilde{x},\tilde{y}) + \hat{f}(\tilde{x},\tilde{y}) - \min_x \hat{f}(x,\tilde{y}) \\
\leq& \frac{1}{2\mu}\|\nabla_y \hat{f}(\tilde{x},\tilde{y})\|^2 + \frac{1}{2l}\|\nabla_x \hat{f}(\tilde{x},\tilde{y})\|^2 = \frac{1}{2\mu}\|\nabla_y f(\tilde{x},\tilde{y})\|^2 + \frac{1}{2l}\|\nabla_x f(\tilde{x},\tilde{y})\|^2 \leq \frac{1}{l}\epsilon^2,
\end{aligned}$$

where in the first inequality we use $\max_y \min_x \hat{f}(x,y) \leq \max_y \hat{f}(\tilde{x},y)$, l-strong-convexity of $\hat{f}(\cdot,\tilde{y})$ and μ-PL of $\hat{f}(\tilde{x},\cdot)$. Combined with (9) we have

$$P_0 \leq \frac{2\epsilon^2}{l}.$$

Then we note that

$$\begin{aligned}
\|x^k - x^*\|^2 \leq& 2\|x^k - x^*(y^k)\|^2 + 2\|x^*(y^k) - x^*\|^2 \leq \frac{4}{l}[\hat{f}(x^k,y^k) - \hat{\Psi}(y^k)] + 18\|y^k - y^*\|^2 \\
\leq& \frac{4}{l}[\hat{f}(x^k,y^k) - \hat{\Psi}(y^k)] + \frac{18}{\mu}[\hat{\Psi}(y^k) - \hat{\Psi}^*] \leq \frac{40}{\mu}P_k,
\end{aligned}$$

where in the second inequality we use l-strong-convexity of $\hat{f}(\cdot,y^k)$ and Lemma A.1, in the third inequality we use μ-PL of $\hat{\Psi}(\cdot)$ (see e.g. (Yang et al., 2020a)). Because $\hat{f}(x,y)$ is l-strongly convex about x, μ-PL about y and $3l$-smooth, it satifies the two-sided PL condition in (Yang et al., 2020a) and it can be solved by AGDA. By Theorem A.1, if we choose $\tau_1 = \frac{1}{3l}$ and $\tau_2 = \frac{l^2}{18(3l)^3} = \frac{1}{486l}$, we have

$$P_k \leq \left(1 - \frac{1}{972\kappa}\right)^k P_0,$$

Therefore,

$$\|x^k - x^*\|^2 \leq \frac{40}{\mu}P_k \leq \frac{40}{\mu}\left(1 - \frac{1}{972\kappa}\right)^k P_0 \leq \frac{80\epsilon^2}{\mu l}\left(1 - \frac{1}{972\kappa}\right)^k.$$

So after $O(\kappa \log \kappa)$ iterations we have $\|x^k - x^*\|^2 \leq \frac{\epsilon^2}{\kappa^2 l^2}$.

Stochastic case: By Theorem A.1, if we choose $\tau_1 \leq \frac{1}{3l}$ and $\tau_2 = \frac{l^2\tau_1}{18(3l)^2} = \frac{\tau_1}{162}$, we have

$$P_k \leq \left(1 - \frac{\mu\tau_2}{2}\right)^k P_0 + O(\kappa\tau_2\sigma^2).$$

With $\tau_2 = \min\left\{\frac{1}{486l}, \Theta\left(\frac{\epsilon^2}{\kappa^4 l\sigma^2}\right)\right\}$ and $\tau_1 = 162\tau_2$, we have $\|x^k - x^*\|^2 \leq \frac{\epsilon^2}{\kappa^2 l^2}$ after $O\left(\kappa\log(\kappa) + \kappa^5\sigma^2\log(\kappa)\epsilon^{-2}\right)$ iterations.

■

B PROOFS FOR STOCHASTIC AGDA

Proof of Theorem 3.1

Proof

Because Φ is L-smooth with $L = l + \frac{l\kappa}{2}$ by Lemma A.3, we have the following by Lemma A.4 in (Yang et al., 2020a)

$$\begin{aligned}\Phi(x_{t+1}) \leq & \Phi(x_t) + \langle \nabla\Phi(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ = & \Phi(x_t) - \tau_1\langle \nabla\Phi(x_t), G_x(x_t, y_t, \xi_{t1})\rangle + \frac{L}{2}\tau_1^2\|G_x(x_t, y_t, \xi_1^t)\|^2.\end{aligned}$$

Taking expectation of both side and use Assumption 2.3, we get

$$\begin{aligned}\mathbb{E}[\Phi(x_{t+1})] \leq & \mathbb{E}[\Phi(x_t)] - \tau_1\mathbb{E}[\langle \nabla\Phi(x_t), \nabla_x f(x_t, y_t)\rangle] + \frac{L}{2}\tau_1^2\mathbb{E}[\|G_x(x_t, y_t, \xi_1^t)\|^2] \\ \leq & \mathbb{E}[\Phi(x_t)] - \tau_1\mathbb{E}[\langle \nabla\Phi(x_t), \nabla_x f(x_t, y_t)\rangle] + \frac{L}{2}\tau_1^2\mathbb{E}[\|\nabla_x f(x_t, y_t)\|^2] + \frac{L}{2}\tau_1^2\sigma^2 \\ \leq & \mathbb{E}[\Phi(x_t)] - \tau_1\mathbb{E}[\langle \nabla\Phi(x_t), \nabla_x f(x_t, y_t)\rangle] + \frac{\tau_1}{2}\mathbb{E}[\|\nabla_x f(x_t, y_t)\|^2] + \frac{L}{2}\tau_1^2\sigma^2 \\ \leq & \mathbb{E}[\Phi(x_t)] - \frac{\tau_1}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 + \frac{\tau_1}{2}\mathbb{E}\|\nabla_x f(x_t, y_t) - \nabla\Phi(x_t)\|^2 + \frac{L}{2}\tau_1^2\sigma^2, \end{aligned} \tag{10}$$

where in the second inequality we use Assumption 2.3, and in the third inequality we use $\tau_1 \leq 1/L$. By smoothness of $f(x, \cdot)$, we have

$$\begin{aligned}f(x_{t+1}, y_{t+1}) \geq & f(x_{t+1}, y_t) + \langle \nabla_y f(x_{t+1}, y_t), y_{t+1} - y_t\rangle - \frac{l}{2}\|y_{t+1} - y_t\|^2 \\ \geq & f(x_{t+1}, y_t) + \tau_2\langle \nabla_y f(x_{t+1}, y_t), G_y(x_{t+1}, y_t, \xi_2^t)\rangle - \frac{l\tau_2^2}{2}\|G_y(x_{t+1}, y_t, \xi_2^t)\|^2.\end{aligned}$$

Taking expectation, as $\tau_2 \leq \frac{1}{l}$

$$\begin{aligned}\mathbb{E}f(x_{t+1}, y_{t+1}) - \mathbb{E}f(x_{t+1}, y_t) \geq & \tau_2\mathbb{E}\|\nabla_y f(x_{t+1}, y_t)\|^2 - \frac{l\tau_2^2}{2}\mathbb{E}\|\nabla_y f(x_{t+1}, y_t)\|^2 - \frac{l\tau_2^2}{2}\sigma^2 \\ \geq & \frac{\tau_2}{2}\mathbb{E}\|\nabla_y f(x_{t+1}, y_t)\|^2 - \frac{l\tau_2^2}{2}\sigma^2. \end{aligned} \tag{11}$$

By smoothness of $f(\cdot, y)$, we have

$$\begin{aligned}f(x_{t+1}, y_t) \geq & f(x_t, y_t) + \langle \nabla_x f(x_t, y_t), x_{t+1} - x_t\rangle - \frac{l}{2}\|x_{t+1} - x_t\|^2 \\ \geq & f(x_t, y_t) - \tau_1\langle \nabla_x f(x_t, y_t), G_x f(x_t, y_t, \xi_1^t)\rangle - \frac{l\tau_1^2}{2}\|G_x(x_t, y_t, \xi_1^t)\|^2.\end{aligned}$$

Taking expectation, as $\tau_1 \leq \frac{1}{l}$

$$\begin{aligned}\mathbb{E}f(x_{t+1}, y_t) - \mathbb{E}f(x_t, y_t) \geq & -\tau_1\mathbb{E}\|\nabla_x f(x_t, y_t)\| - \frac{l\tau_1^2}{2}\mathbb{E}\|\nabla_x f(x_t, y_t)\| - \frac{l\tau_1^2}{2}\sigma^2 \\ \geq & -\frac{3\tau_1}{2}\mathbb{E}\|\nabla_x f(x_t, y_t)\|^2 - \frac{l\tau_1^2}{2}\sigma^2. \end{aligned} \tag{12}$$

Therefore, summing (12) and (11) together

$$\mathbb{E}f(x_{t+1}, y_{t+1}) - \mathbb{E}f(x_t, y_t) \geq \frac{\tau_2}{2}\mathbb{E}\|\nabla_y f(x_{t+1}, y_t)\|^2 - \frac{3\tau_1}{2}\mathbb{E}\|\nabla_x f(x_t, y_t)\|^2 - \frac{l\tau_1^2}{2}\sigma^2 - \frac{l\tau_2^2}{2}\sigma^2. \tag{13}$$

Now we consider the following potential function, for some $\alpha > 0$ which we will pick later

$$V_t = V(x_t, y_t) = \Phi(x_t) + \alpha[\Phi(x_t) - f(x_t, y_t)] = (1+\alpha)\Phi(x_t) - \alpha f(x_t, y_t). \tag{14}$$

Then by combining (14) and (10) we have

$$
\begin{aligned}
&\mathbb{E}V_t - \mathbb{E}V_{t+1} \\
\geq &\frac{\tau_1}{2}(1+\alpha)\mathbb{E}\left\|\nabla\Phi\left(x_t\right)\right\|^2 - \frac{\tau_1}{2}(1+\alpha)\mathbb{E}\left\|\nabla_x f\left(x_t, y_t\right) - \nabla\Phi\left(x_t\right)\right\|^2 + \frac{\tau_2\alpha}{2}\mathbb{E}\|\nabla_y f(x_{t+1}, y_t)\|^2 - \\
&\frac{3\tau_1\alpha}{2}\mathbb{E}\|\nabla_x f(x_t, y_t)\|^2 - \left[\frac{L(1+\alpha)}{2}\tau_1^2 + \frac{l\tau_2^2\alpha}{2} + \frac{l\tau_1^2\alpha}{2}\right]\sigma^2 \\
\geq &\left[\frac{\tau_1}{2}(1+\alpha) - 3\tau_1\alpha\right]\mathbb{E}\left\|\nabla\Phi\left(x_t\right)\right\|^2 - \left[\frac{\tau_1}{2}(1+\alpha) + 3\tau_1\alpha\right]\mathbb{E}\left\|\nabla_x f\left(x_t, y_t\right) - \nabla\Phi\left(x_t\right)\right\|^2 + \qquad (15) \\
&\frac{\tau_2\alpha}{4}\mathbb{E}\|\nabla_y f(x_t, y_t)\|^2 - \frac{\tau_2\alpha}{2}\mathbb{E}\|\nabla_y f(x_{t+1}, y_t) - \nabla_y f(x_t, y_t)\|^2 - \left[\frac{L(1+\alpha)}{2}\tau_1^2 + \frac{l\tau_2^2\alpha}{2} + \frac{l\tau_1^2\alpha}{2}\right]\sigma^2 \\
\geq &\left[\frac{\tau_1}{2}(1+\alpha) - 3\tau_1\alpha\right]\mathbb{E}\left\|\nabla\Phi\left(x_t\right)\right\|^2 - \left[\frac{\tau_1}{2}(1+\alpha) + 3\tau_1\alpha\right]\mathbb{E}\left\|\nabla_x f\left(x_t, y_t\right) - \nabla\Phi\left(x_t\right)\right\|^2 + \\
&\frac{\tau_2\alpha}{4}\mathbb{E}\|\nabla_y f(x_t, y_t)\|^2 - \frac{\tau_2\alpha}{2}l^2\mathbb{E}\|x_{t+1} - x_t\|^2 - \left[\frac{L(1+\alpha)}{2}\tau_1^2 + \frac{l\tau_2^2\alpha}{2} + \frac{l\tau_1^2\alpha}{2}\right]\sigma^2 \\
\geq &\left[\frac{\tau_1}{2}(1+\alpha) - 3\tau_1\alpha\right]\mathbb{E}\left\|\nabla\Phi\left(x_t\right)\right\|^2 - \left[\frac{\tau_1}{2}(1+\alpha) + 3\tau_1\alpha\right]\mathbb{E}\left\|\nabla_x f\left(x_t, y_t\right) - \nabla\Phi\left(x_t\right)\right\|^2 + \\
&\frac{\tau_2\alpha}{4}\mathbb{E}\|\nabla_y f(x_t, y_t)\|^2 - \frac{\tau_2\alpha}{2}l^2\tau_1^2\mathbb{E}\|\nabla_x f(x_t, y_t)\|^2 - \left[\frac{L(1+\alpha)}{2}\tau_1^2 + \frac{l\tau_2^2\alpha}{2} + \frac{l\tau_1^2\alpha}{2} + \frac{\tau_2}{2}\alpha l^2\tau_1^2\right]\sigma^2 \\
\geq &\left[\frac{\tau_1}{2}(1+\alpha) - 3\tau_1\alpha - \tau_2\alpha l^2\tau_1^2\right]\mathbb{E}\left\|\nabla\Phi\left(x_t\right)\right\|^2 - \left[\frac{\tau_1}{2}(1+\alpha) + 3\tau_1\alpha + \tau_2\alpha l^2\tau_1^2\right]\mathbb{E}\left\|\nabla_x f\left(x_t, y_t\right) - \nabla\Phi\left(x_t\right)\right\|^2 + \\
&\frac{\tau_2\alpha}{4}\mathbb{E}\|\nabla_y f(x_t, y_t)\|^2 - \left[\frac{L(1+\alpha)}{2}\tau_1^2 + \frac{l\tau_2^2\alpha}{2} + \frac{l\tau_1^2\alpha}{2} + \frac{\tau_2}{2}\alpha l^2\tau_1^2\right]\sigma^2, \qquad (16)
\end{aligned}
$$

where in the first inequality we use $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and $\|a\|^2 \geq \|b\|^2/2 - \|a-b\|^2$, in the second inequality we use smoothness, and in the last inequality we use $\|a+b\|^2 \leq \|a\|^2 + \|b\|^2$. Note that by smoothness and PL condition, fixing $y^*(x_t)$ to be the projection of y_t to the set $\underset{y}{\operatorname{Argmin}} f(x_t, y)$,

$$
\left\|\nabla_x f\left(x_t, y_t\right) - \nabla\Phi\left(x_t\right)\right\|^2 \leq l^2\|y_t - y^*(x_t)\|^2 \leq \kappa^2\|\nabla_y f(x_t, y_t)\|^2.
$$

Plugging it into (16), we get

$$
\begin{aligned}
\mathbb{E}V_t - \mathbb{E}V_{t+1} \geq &\left[\frac{\tau_1}{2}(1+\alpha) - 3\tau_1\alpha - \tau_2\alpha l^2\tau_1^2\right]\mathbb{E}\left\|\nabla\Phi\left(x_t\right)\right\|^2 + \\
&\left[\frac{\tau_2\alpha}{4} - \frac{\tau_1}{2}(1+\alpha)\kappa^2 - 3\tau_1\alpha\kappa^2 - \tau_2\alpha l^2\tau_1^2\kappa^2\right]\mathbb{E}\|\nabla_y f(x_t, y_t)\|^2 - \\
&\left[\frac{L(1+\alpha)}{2}\tau_1^2 + \frac{l\tau_2^2\alpha}{2} + \frac{l\tau_1^2\alpha}{2} + \frac{\tau_2}{2}\alpha l^2\tau_1^2\right]\sigma^2. \qquad (17)
\end{aligned}
$$

Then we note that when $\alpha = \frac{1}{8}$, $\tau_1 \leq \frac{1}{l}$ and $\tau_2 \leq \frac{1}{l}$,

$$
\frac{\tau_1}{2}(1+\alpha) - 3\tau_1\alpha - \tau_2\alpha l^2\tau_1^2 \geq \frac{\tau_1}{16}.
$$

Furthermore, when $\tau_1 \leq \frac{\tau_2}{68\kappa^2}$, then

$$
\frac{\tau_2\alpha}{4} - \frac{\tau_1}{2}(1+\alpha)\kappa^2 - 3\tau_1\alpha\kappa^2 - \tau_2\alpha l^2\tau_1^2\kappa^2 \geq \frac{1}{64}\tau_2 \geq \frac{17}{16}\kappa^2\tau_1.
$$

Also, as $\alpha = \frac{1}{8}$, $\tau_2 \leq \frac{1}{l}$ and $\tau_1 = \frac{\tau_2}{68\kappa^2}$

$$
\frac{L(1+\alpha)}{2}\tau_1^2 + \frac{l\tau_2^2\alpha}{2} + \frac{l\tau_1^2\alpha}{2} + \frac{\tau_2}{2}\alpha l^2\tau_1^2 \leq 292\kappa^4 l\tau_1^2.
$$

Therefore,

$$
\mathbb{E}V_t - \mathbb{E}V_{t+1} \geq \frac{\tau_1}{16}\mathbb{E}\left\|\nabla\Phi\left(x_t\right)\right\|^2 + \frac{17}{16}\kappa^2\tau_1\mathbb{E}\|\nabla_y f(x_t, y_t)\|^2 - 292\kappa^4 l\tau_1^2\sigma^2. \qquad (18)
$$

Telescoping and rearraging, with $a_0 \triangleq \Phi(x_0) - f(x_0, y_0)$,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(x_t\right)\right\|^2 \leq & \frac{16}{\tau_1 T}[V_0 - \min_{x,y} V(x,y)] + 4762\kappa^4 l\tau_1\sigma^2 \\ \leq & \frac{16}{\tau_1 T}[\Phi(x_0) - \Phi^*] + \frac{2}{\tau_1 T}a_0 + 4672\kappa^4 l\tau_1\sigma^2,\end{aligned}$$

where in the second inequality we note that since for any x we can find y such that $\Phi(x) = f(x,y)$,

$$V_0 - \min_{x,y} V(x,y) = \Phi(x_0) + \alpha[\Phi(x_0) - f(x_0,y_0)] - \min_{x,y}\{\Phi(x) + \alpha[\Phi(x) - f(x,y)]\} = \Phi(x_0) - \Phi^* + \alpha[\Phi(x_0) - f(x_0,y_0)].$$

Picking $\tau_1 = \min\left\{\frac{\sqrt{\Phi(x_0)-\Phi^*}}{4\sigma\kappa^2\sqrt{Tl}}, \frac{1}{68l\kappa^2}\right\}$,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(x_t\right)\right\|^2 \leq & \max\left\{\frac{4\sigma\kappa^2\sqrt{Tl}}{\sqrt{\Phi(x_0)-\Phi^*}}, 68l\kappa^2\right\}\frac{16}{T}[\Phi(x_0)-\Phi^*] + \max\left\{\frac{4\sigma\kappa^2\sqrt{Tl}}{\sqrt{\Phi(x_0)-\Phi^*}}, 68l\kappa^2\right\}\frac{2}{T}a_0 + \\ & \frac{\sqrt{\Phi(x_0)-\Phi^*}}{4\sigma\kappa^2\sqrt{Tl}}4672\kappa^4 l\sigma^2 \\ \leq & \frac{1088l\kappa^2}{T}[\Phi(x_0)-\Phi^*] + \frac{136l\kappa^2}{T}a_0 + \frac{8\kappa^2\sqrt{l}a_0}{\sqrt{[\Phi(x_0)-\Phi^*]T}}\sigma + \frac{1232\kappa^2\sqrt{l[\Phi(x_0)-\Phi^*]}}{\sqrt{T}}\sigma.\end{aligned}$$

Here we can pick $\tau_2 = \min\left\{\frac{17\sqrt{\Phi(x_0)-\Phi^*}}{\sigma\sqrt{Tl}}, \frac{1}{l}\right\}$.

■

Proof of Corollary 3.1

Proof Similar to the proof of part (a) in Proposition 2.1, fixing $y^*(x_t)$ to be the projection of x_t to $\operatorname{Arg\,max}_y f(x_t, y)$, we have

$$\begin{aligned}\|\nabla_x f(x_t,y_t)\|^2 \leq & 2\|\nabla_x f(x_t, y^*(x_t))\|^2 + 2\|\nabla_x f(x_t,y_t) - \nabla_x f(x_t, y^*(x_t))\|^2 \\ \leq & 2\|\nabla\Phi(x_t)\|^2 + 2l^2\|y_t - y^*(x_t)\|^2 \\ \leq & 2\|\nabla\Phi(x_t)\| + 2\kappa^2\|\nabla_y f(x_t,y_t)\|^2,\end{aligned}$$

where in the first inequality we use Lemma A.3 and in the last inequality we use Lemma A.2. Plugging into (18),

$$\mathbb{E}V_t - \mathbb{E}V_{t+1} \geq \frac{\tau_1}{32}\mathbb{E}\left\|\nabla\Phi\left(x_t\right)\right\|^2 + \kappa^2\tau_1\mathbb{E}\|\nabla_y f(x_t,y_t)\|^2 - 292\kappa^4 l\tau_1^2\sigma^2.$$

By the same reasoning as the proof of Theorem 3.1 (after equation (18)), with the same stepsizes, we can show

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla_x f\left(x_t,y_t\right)\right\|^2 + & 32\kappa^2\mathbb{E}\left\|\nabla_y f\left(x_t,y_t\right)\right\|^2 \leq \\ & \frac{d_0 l\kappa^2}{T}[\Phi(x_0)-\Phi^*] + \frac{d_1 l\kappa^2}{T}a_0 + \frac{d_2\kappa^2\sqrt{l}a_0}{\sqrt{[\Phi(x_0)-\Phi^*]T}}\sigma + \frac{d_3\kappa^2\sqrt{l[\Phi(x_0)-\Phi^*]}}{\sqrt{T}}\sigma,\end{aligned}$$

where d_0, d_1, d_2 and d_3 are $O(1)$ constants.

■

C PROOFS FOR STOCHASTIC SMOOTHED-AGDA

Before we present the theorem and converge, we adopt the following notations.

- $\hat{f}(x, y; z) = f(x, y) + \frac{p}{2}\|x - z\|^2$: the auxiliary function;
- $\Psi(y; z) = \min_x \hat{f}(x, y; z)$: the dual function of the auxiliary problem;
- $\Phi(x; z) = \max_y \hat{f}(x, y; z)$: the primal function of the auxiliary problem;
- $P(z) = \min_x \max_y \hat{f}(x, y; z)$: the optimal value for the auxiliary function fixing z;
- $x^*(y, z) = \arg\min_x \hat{f}(x, y; z)$: the optimal x w.r.t y and z in the auxiliary function;
- $x^*(z) = \arg\min_x \Phi(x; z)$: the optimal x w.r.t z in the auxiliary function when y is already optimal w.r.t x;
- $Y^*(z) = \operatorname{Arg}\max_y \Psi(y; z)$: the optimal set of y w.r.t z when x is optimal to y;
- $y^+(z) = y + \tau_2 \nabla_y \hat{f}(x^*(y, z), y; z)$: y after one step of gradient ascent in y with the gradient of the dual function;
- $x^+(y, z) = x - \tau_1 \nabla_x \hat{f}(x, y; z)$: x after one step of gradient descent with gradient at current point;
- $\hat{G}_x(x, y, \xi; z) = G_x(x, y, \xi) + p(x - z)$: the stochastic gradient for regularized auxiliary function.

Lemma C.1 *We have the following inequalities as $p > l$*

$$
\begin{aligned}
&\|x^*(y, z) - x^*(y, z')\| \leq \gamma_1 \|z - z'\|, \\
&\|x^*(z) - x^*(z') \leq \gamma_1 \|z - z'\|, \\
&\|x^*(y, z) - x^*(y', z)\| \leq \gamma_2 \|y - y'\|, \\
&\mathbb{E}\|x_{t+1} - x^*(y_t, z_t)\|^2 \leq \gamma_3^2 \tau_1^2 \mathbb{E}\|\nabla_x \hat{f}(x_t, y_t; z_t)\|^2 + 2\tau_1^2 \sigma^2,
\end{aligned}
$$

where $\gamma_1 = \frac{p}{-l+p}$, $\gamma_2 = \frac{l+p}{-l+p}$ *and* $\gamma_3^2 = \frac{2}{\tau_1^2(-l+p)^2} + 2$.

Proof The first and second inequality is the same as Proposition B.4 in (Zhang et al., 2020). The third inequality is a direct result of Lemma A.1. Now we show the last inequality.

$$
\begin{aligned}
\|x_{t+1} - x^*(y_t, z_t)\|^2 \leq& 2\|x_t - x^*(y_t, z_t)\|^2 + 2\|x_{t+1} - x_t\|^2 \\
\leq& \frac{2}{(-l+p)^2}\|\nabla_x \hat{f}(x_t, y_t; z_t)\|^2 + 2\tau_1^2 \|\hat{G}_x(x_t, y_t, \xi_1^t; z_t)\|^2.
\end{aligned}
$$

where the second inequality use $(-l+p)$-strong convexity of $\hat{f}(\cdot, y_t; z_t)$. Taking expectation

$$
\begin{aligned}
\mathbb{E}\|x_{t+1} - x^*(y_t, z_t)\|^2 \leq& \frac{2}{(-l+p)^2}\mathbb{E}\|\nabla_x \hat{f}(x_t, y_t; z_t)\|^2 + 2\tau_1^2 \mathbb{E}\|\nabla_x \hat{f}(x_t, y_t; z_t)\|^2 + 2\tau_1^2 \sigma^2 \\
\leq& 2\left[\frac{1}{(-l+p)^2} + \tau_1^2\right] \mathbb{E}\|\nabla_x \hat{f}(x_t, y_t; z_t)\|^2 + 2\tau_1^2 \sigma^2.
\end{aligned}
$$

■

Lemma C.2 *The following inequality holds*

$$
\|x^*(z) - x^*(y^+(z), z)\|^2 \leq \frac{1}{(p-l)\mu}\left(1 + \tau_2 l + \frac{\tau_2 l(p+l)}{p-l}\right)^2 \|\nabla_y \hat{f}(x^*(y, z), y; z)\|^2. \tag{19}
$$

Proof By the $(p-l)$-strong convexity of $\Phi(\cdot; z)$, we have

$$
\begin{aligned}
&\|x^*(z) - x^*(y^+(z), z)\|^2 \\
\leq& \frac{2}{p-l}\left[\Phi(x^*(y^+(z), z); z) - \Phi(x^*(z); z)\right] \\
\leq& \frac{2}{p-l}\left[\Phi(x^*(y^+(z), z); z) - \hat{f}(x^*(y^+(z), z), y^+(z); z) + \hat{f}(x^*(y^+(z), z), y^+(z); z) - \Phi(x^*(z); z)\right] \\
\leq& \frac{1}{(p-l)\mu}\|\nabla_y \hat{f}(x^*(y^+(z), z), y^+(z); z)\|^2,
\end{aligned}
$$

where in the last inequality we use μ-PL of $\hat{f}(x,\cdot;z)$ and $\hat{f}(x^*(y^+(z),z),y^+(z);z) \le \Phi(x^*(z);z)$. Then

$$\begin{aligned}\|\nabla_y \hat{f}(x^*(y^+(z),z),y^+(z);z)\| \le& \|\nabla_y \hat{f}(x^*(y,z),y;z)\| + \|\nabla_y \hat{f}(x^*(y,z),y;z) - \nabla_y \hat{f}(x^*(y^+(z),z),y^+(z);z)\| \\ \le& \|\nabla_y \hat{f}(x^*(y,z),y;z)\| + l\|x^*(y,z) - x^*(y^+(z),z)\| + l\|y - y^+(z)\| \\ \le& \left(1 + \frac{\tau_2 l(p+l)}{p-l} + \tau_2 l\right)\|\nabla_y \hat{f}(x^*(y,z),y;z)\|,\end{aligned}$$

where in the last inequality we use Lemma C.1 and $\|y - y^+(z)\| = \tau_2\|\nabla_y \hat{f}(x^*(y,z),y;z)\|$. We reach our conclusion by combining with the previous inequality.

■

Proof of Theorem 4.1

Proof We separate our proof into several parts: we first present three descent lemmas, then we show the descent property for a potential function, later we discuss the relation between our stationary measure and the potential function, and last we put things together.

Primal descent: By the $(p+l)$-smoothness of $\hat{f}(\cdot,y_t;z_t)$,

$$\begin{aligned}\hat{f}(x_{t+1},y_t;z_t) \le& \hat{f}(x_t,y_t;z_t) + \langle \nabla_x \hat{f}(x_t,y_t;z_t), x_{t+1}-x_t\rangle + \frac{p+l}{2}\|x_{t+1}-x_t\|^2 \\ =& \hat{f}(x_t,y_t;z_t) - \tau_1\langle \nabla_x \hat{f}(x_t,y_t;z_t), \hat{G}_x(x_t,y_t,\xi_1^t;z_t)\rangle + \frac{p+l}{2}\tau_1^2\|\hat{G}_x(x_t,y_t,\xi_1^t;z_t)\|^2,\end{aligned}$$

We can easily verify that $\mathbb{E}\hat{G}_x(x_t,y_t,\xi_1^t;z_t) = \nabla_x \hat{f}(x_t,y_t;z_t)$, and $\mathbb{E}\|\hat{G}_x(x_t,y_t,\xi_1^t;z_t) - \mathbb{E}\hat{G}_x(x_t,y_t,\xi_1^t;z_t)\|^2 = \mathbb{E}\|G_x(x_t,y_t,\xi_1^t) - \nabla_x f(x_t,y_t)\|^2 \le \sigma^2$. Taking expectation of both sides,

$$\mathbb{E}\hat{f}(x_{t+1},y_t;z_t) \le \mathbb{E}\hat{f}(x_t,y_t;z_t) - \tau_1\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2 + \frac{p+l}{2}\tau_1^2\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2 + \frac{p+l}{2}\tau_1^2\sigma^2.$$

As $\tau_1 \le \frac{1}{p+l}$,

$$\mathbb{E}\hat{f}(x_t,y_t;z_t) - \mathbb{E}\hat{f}(x_{t+1},y_t;z_t) \ge \frac{\tau_1}{2}\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2 - \frac{p+l}{2}\tau_1^2\sigma^2. \tag{20}$$

Also, because $\hat{f}(x_{t+1},\cdot;z_t)$ is smooth,

$$\begin{aligned}\hat{f}(x_{t+1},y_t;z_t) - \hat{f}(x_{t+1},y_{t+1};z_t) \ge& \langle \nabla_y \hat{f}(x_{t+1},y_t;z_t), y_t - y_{t+1}\rangle - \frac{l}{2}\|y_t - y_{t+1}\|^2 \\ =& -\tau_2\langle \nabla_y \hat{f}(x_{t+1},y_t;z_t), G_y(x_{t+1},y_t,\xi_2^t)\rangle - \frac{l}{2}\tau_2^2\|G_y(x_{t+1},y_t,\xi_2^t)\|^2.\end{aligned}$$

Taking expectation of both sides,

$$\begin{aligned}\mathbb{E}\hat{f}(x_{t+1},y_t;z_t) - \mathbb{E}\hat{f}(x_{t+1},y_{t+1};z_t) \ge& -\tau_2\mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 - \frac{l}{2}\tau_2^2\mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 - \frac{l}{2}\tau_2^2\sigma^2 \\ =& -\left(1 + \frac{l\tau_2}{2}\right)\tau_2\mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 - \frac{l}{2}\tau_2^2\sigma^2.\end{aligned} \tag{21}$$

Furthermore, by definition of $\hat{f}$ and z_{t+1}, as $0 < \beta < 1$

$$\begin{aligned}&\hat{f}(x_{t+1},y_{t+1};z_t) - \hat{f}(x_{t+1},y_{t+1};z_{t+1}) \\ =& \frac{p}{2}[\|x_{t+1}-z_t\|^2 - \|x_{t+1}-z_{t+1}\|^2] = \frac{p}{2}\left[\frac{1}{\beta^2}\|(z_{t+1}-z_t)\|^2 - \|(1-\beta)(x_{t+1}-z_t)\|^2\right] \\ =& \frac{p}{2}\left[\frac{1}{\beta^2}\|z_{t+1}-z_t\|^2 - \frac{(1-\beta)^2}{\beta^2}\|z_{t+1}-z_t\|^2\right] \ge \frac{p}{2\beta}\|z_t - z_{t+1}\|^2.\end{aligned} \tag{22}$$

Combining (20), (21) and (22),

$$\begin{aligned}&\mathbb{E}\hat{f}(x_t,y_t;z_t) - \mathbb{E}\hat{f}(x_{t+1},y_t;z_t) \ge \\ &\frac{\tau_1}{2}\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2 - \left(1 + \frac{l\tau_2}{2}\right)\tau_2\mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 + \frac{p}{2\beta}\mathbb{E}\|z_t - z_{t+1}\|^2 - \frac{l}{2}\tau_2^2\sigma^2 - \frac{p+l}{2}\tau_1^2\sigma^2.\end{aligned} \tag{23}$$

Dual Descent: Since the dual function $\Psi(y;z)$ is L_Ψ smooth with $L_\Psi = l + l\gamma_2$ by Lemma B.3 in (Zhang et al., 2020) or Lemma A.3,

$$\begin{aligned}\Psi(y_{t+1};z_t) - \Psi(y_t;z_t) \geq & \langle \nabla_y \Psi(y_t;z_t), y_{t+1} - y_t\rangle - \frac{L_\Psi}{2}\|y_{t+1}-y_t\|^2 \\ = & \langle \nabla_y \hat{f}(x^*(y_t,z_t), y_t; z_t), y_{t+1} - y_t\rangle - \frac{L_\Psi}{2}\|y_{t+1}-y_t\|^2.\end{aligned}$$

Taking expectation,

$$\mathbb{E}\Psi(y_{t+1};z_t) - \mathbb{E}\Psi(y_t;z_t) \geq \tau_2 \mathbb{E}\langle \nabla_y \hat{f}(x^*(y_t,z_t), y_t; z_t), \nabla_y f(x_{t+1}, y_t)\rangle - \frac{L_\Psi}{2}\tau_2^2 \mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 - \frac{L_\Psi}{2}\tau_2^2\sigma^2. \quad (24)$$

Also,

$$\begin{aligned}\Psi(y_{t+1};z_{t+1}) - \Psi(y_{t+1};z_t) = & \hat{f}(x^*(x_{t+1},z_{t+1}), y_{t+1}; z_{t+1}) - \hat{f}(x^*(y_{t+1},z_t), y_{t+1}; z_t) \\ \geq & \hat{f}(x^*(x_{t+1},z_{t+1}), y_{t+1}; z_{t+1}) - \hat{f}(x^*(y_{t+1},z_{t+1}), y_{t+1}; z_t) \\ = & \frac{p}{2}\left[\|z_{t+1} - x^*(y_{t+1},z_{t+1})\|^2 - \|z_t - x^*(y_{t+1},z_{t+1})\|^2\right] \\ = & \frac{p}{2}\langle z_{t+1} - z_t, z_{t+1} + z_t - 2x^*(y_{t+1},z_{t+1})\rangle. \quad (25)\end{aligned}$$

Combining with (24), we have

$$\begin{aligned}\mathbb{E}\Psi(y_{t+1};z_{t+1}) - \mathbb{E}\Psi(y_t;z_t) \geq & \tau_2 \mathbb{E}\langle \nabla_y \hat{f}(x^*(y_t,z_t), y_t; z_t), \nabla_y f(x_{t+1}, y_t)\rangle - \frac{L_\Psi}{2}\tau_2^2 \mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 + \\ & \frac{p}{2}\mathbb{E}\langle z_{t+1} - z_t, z_{t+1} + z_t - 2x^*(y_{t+1},z_{t+1})\rangle - \frac{L_\Psi}{2}\tau_2^2\sigma^2. \quad (26)\end{aligned}$$

Proximal Descent: for all $y^*(z_{t+1}) \in Y^*(z_{t+1})$ and $y^*(z_t) \in Y^*(z_t)$,

$$\begin{aligned}P(z_{t+1}) - P(z_t) = & \Psi(y^*(z_{t+1}); z_{t+1}) - \Psi(y^*(z_t); z_t) \\ \leq & \Psi(y^*(z_{t+1}); z_{t+1}) - \Psi(y^*(z_{t+1}); z_t) \\ = & \hat{f}(x^*(y^*(z_{t+1}), z_{t+1}), y^*(z_{t+1}); z_{t+1}) - \hat{f}(x^*(y^*(z_{t+1}), z_t), y^*(z_{t+1}); z_t) \\ \leq & \hat{f}(x^*(y^*(z_{t+1}), z_t), y^*(z_{t+1}); z_{t+1}) - \hat{f}(x^*(y^*(z_{t+1}), z_t), y^*(z_{t+1}); z_t) \\ = & \frac{p}{2}\langle z_{t+1} - z_t, z_{t+1} - z_t - 2x^*(y^*(z_{t+1}), z_t)\rangle. \quad (27)\end{aligned}$$

Potential Function We use the potential function $V_t = V(x_t, y_t, z_t) = \hat{f}(x_t, y_t; z_t) - 2\Psi(y_t; z_t) + 2P(z_t)$. By three descent steps above, we have

$$\begin{aligned}\mathbb{E}V_t - \mathbb{E}V_{t+1} \geq & \frac{\tau_1}{2}\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2 - \left(1 + \frac{l\tau_2}{2}\right)\tau_2 \mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 + \frac{p}{2\beta}\mathbb{E}\|z_t - z_{t+1}\|^2 + \\ & 2\tau_2 \mathbb{E}\langle \nabla_y \hat{f}(x^*(y_t,z_t), y_t; z_t), \nabla_y f(x_{t+1}, y_t)\rangle - L_\Psi \tau_2^2 \mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 + \\ & p\mathbb{E}\langle z_{t+1} - z_t, z_{t+1} + z_t - 2x^*(y_{t+1},z_{t+1})\rangle - p\mathbb{E}\langle z_{t+1} - z_t, z_{t+1} - z_t - 2x^*(y^*(z_{t+1}), z_t)\rangle - \\ & \frac{l}{2}\tau_2^2\sigma^2 - \frac{p+l}{2}\tau_1^2\sigma^2 - L_\Psi\tau_2^2\sigma^2 \\ \geq & \frac{\tau_1}{2}\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2 + \left(1 - \frac{l\tau_2}{2} - L_\Psi\tau_2\right)\tau_2 \mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 + \frac{p}{2\beta}\mathbb{E}\|z_t - z_{t+1}\|^2 + \\ & 2\tau_2 \mathbb{E}\langle \nabla_y \hat{f}(x^*(y_t,z_t), y_t; z_t) - \nabla_y f(x_{t+1}, y_t), \nabla_y f(x_{t+1}, y_t)\rangle + \\ & p\mathbb{E}\langle z_{t+1} - z_t, 2x^*(y^*(z_{t+1}), z_t) - 2x^*(y_{t+1},z_{t+1})\rangle - \frac{l}{2}\tau_2^2\sigma^2 - \frac{p+l}{2}\tau_1^2\sigma^2 - L_\Psi\tau_2^2\sigma^2 \\ \geq & \frac{\tau_1}{2}\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2 + \frac{\tau_2}{2}\mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 + \frac{p}{2\beta}\mathbb{E}\|z_t - z_{t+1}\|^2 + \\ & 2\tau_2 \mathbb{E}\langle \nabla_y \hat{f}(x^*(y_t,z_t), y_t; z_t) - \nabla_y f(x_{t+1}, y_t), \nabla_y f(x_{t+1}, y_t)\rangle + \\ & 2p\mathbb{E}\langle z_{t+1} - z_t, x^*(y^*(z_{t+1}), z_t) - x^*(y_{t+1},z_{t+1})\rangle - \frac{l}{2}\tau_2^2\sigma^2 - \frac{p+l}{2}\tau_1^2\sigma^2 - L_\Psi\tau_2^2\sigma^2, \quad (28)\end{aligned}$$

where in the last inequality we use $1-\frac{l\tau_2}{2}-L_\Psi\tau_2 \geq \frac{1}{2}$ since $L_\Psi = 4l$ by our choice of τ_2 and p. Now we denote $A = 2\tau_2\langle \nabla_y \hat{f}(x^*(y_t,z_t),y_t;z_t) - \nabla_y f(x_{t+1},y_t), \nabla_y f(x_{t+1},y_t)\rangle$ and $B = 2p\langle z_{t+1}-z_t), x^*(y^*(z_{t+1}),z_t) - x^*(y_{t+1},z_{t+1})\rangle$.

$$\begin{aligned}
B =& 2p\langle z_{t+1}-z_t, x^*(y^*(z_{t+1}),z_t) - x^*(y^*(z_{t+1}),z_{t+1})\rangle + 2p\langle z_{t+1}-z_t, x^*(y^*(z_{t+1}),z_{t+1}) - x^*(y_{t+1},z_{t+1})\rangle \\
\geq& -2p\gamma_1\|z_{t+1}-z_t\|^2 + 2p\langle z_{t+1}-z_t, x^*(y^*(z_{t+1}),z_{t+1}) - x^*(y_{t+1},z_{t+1})\rangle \\
\geq& -\left(2p\gamma_1 + \frac{p}{6\beta}\right)\|z_{t+1}-z_t\|^2 - 6p\beta\|x^*(y^*(z_{t+1}),z_{t+1}) - x^*(y_{t+1},z_{t+1})\|^2, \qquad (29)
\end{aligned}$$

where we use C.1 in the first inequality. Also,

$$\begin{aligned}
A \geq& -2\tau_2\|\nabla_y \hat{f}(x^*(y_t,z_t),y_t;z_t) - \nabla_y f(x_{t+1},y_t)\|\|\nabla_y f(x_{t+1},y_t)\| \\
\geq& -2\tau_2 l\|x_{t+1} - x^*(y_t,z_t)\|\|\nabla_y f(x_{t+1},y_t)\| \\
\geq& -\tau_2^2 l\nu\|\nabla_y f(x_{t+1},y_t)\|^2 - l\nu^{-1}\|x_{t+1} - x^*(y_t,z_t)\|^2, \qquad (30)
\end{aligned}$$

where in the second inequality we use $\nabla_y \hat{f}(x^*(y_t,z_t),y_t;z_t) = \nabla_y f(x^*(y_t,z_t),y_t)$ and in the third inequality $\nu > 0$ and we will choose it later. Taking expectation and applying Lemma C.1

$$\mathbb{E}A \geq -\tau_2^2 l\nu\mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 - l\tau_1^2\nu^{-1}\gamma_3^2\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2 - 2l\nu^{-1}\tau_1^2\sigma^2. \qquad (31)$$

Plugging (31) and (29) into (28),

$$\begin{aligned}
\mathbb{E}V_t - \mathbb{E}V_{t+1} \geq& \left(\frac{\tau_1}{2} - l\tau_1^2\nu^{-1}\gamma_3^2\right)\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2 + \left(\frac{\tau_2}{2} - \tau_2^2 l\nu\right)\mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 + \\
&\left(\frac{p}{2\beta} - 2p\gamma_1 - \frac{p}{6\beta}\right)\mathbb{E}\|z_t - z_{t+1}\|^2 - 6p\beta\mathbb{E}\|x^*(y^*(z_{t+1}),z_{t+1}) - x^*(y_{t+1},z_{t+1})\|^2 - \\
&\left(\frac{p+l}{2} + 2l\nu^{-1}\right)\tau_1^2\sigma^2 - \left(\frac{l}{2} + L_\Psi\right)\tau_2^2\sigma^2, \qquad (32)
\end{aligned}$$

We rewrite $\|\nabla_y f(x_{t+1},y_t)\|^2$ as:

$$\begin{aligned}
\|\nabla_y f(x_{t+1},y_t)\|^2 =& \|\nabla_y \hat{f}(x^*(y_t,z_t),y_t;z_t) + \nabla_y f(x_{t+1},y_t) - \nabla_y \hat{f}(x^*(y_t,z_t),y_t;z_t)\|^2 \\
\geq& \|\nabla_y \hat{f}(x^*(y_t,z_t),y_t;z_t)\|^2/2 - \|\nabla_y f(x_{t+1},y_t) - \nabla_y \hat{f}(x^*(y_t,z_t),y_t;z_t)\|^2 \\
\geq& \|\nabla_y \hat{f}(x^*(y_t,z_t),y_t;z_t)\|^2/2 - l^2\|x_{t+1} - x^*(y_t,z_t)\|^2. \qquad (33)
\end{aligned}$$

Taking expectation and applying Lemma C.1

$$\mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 \geq \mathbb{E}\|\nabla_y \hat{f}(x^*(y_t,z_t),y_t;z_t)\|^2/2 - l^2\gamma_3^2\tau_1^2\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2 - 2l^2\tau_1^2\sigma^2. \qquad (34)$$

Note that $x^*(y^*(z_{t+1}),z_{t+1}) = x^*(z_{t+1})$. We rewrite $\|x^*(y^*(z_{t+1}),z_{t+1}) - x^*(y_{t+1},z_{t+1})\|^2$ as

$$\begin{aligned}
&\|x^*(z_{t+1}) - x^*(y_{t+1},z_{t+1})\|^2 \\
\leq& 4\|x^*(z_{t+1}) - x^*(z_t)\|^2 + 4\|x^*(z_t) - x^*(y_t^+(z_t),z_t)\|^2 + \\
&4\|x^*(y_t^+(z_t),z_t) - x^*(y_{t+1},z_t)\|^2 + 4\|x^*(y_{t+1},z_t) - x^*(y_{t+1},z_{t+1})\|^2 \\
\leq& 4\gamma_1^2\|z_{t+1}-z_t\|^2 + 4\|x^*(z_t) - x^*(y_t^+(z_t),z_t)\|^2 + 4\gamma_2^2\|y_t^+(z_t) - y_{t+1}\|^2 + 4\gamma_1^2\|z_t - z_{t+1}\|^2 \\
\leq& 4\|x^*(z_t) - x^*(y_t^+(z_t),z_t)\|^2 + 8\gamma_2^2\tau_2^2\|\nabla_y \hat{f}(x^*(y_t),z_t),y_t;z_t) - \nabla_y f(x_{t+1},y_t)\|^2 + \\
&8\gamma_2^2\tau_2^2\|\nabla_y f(x_{t+1},y_t) - G_y(x_{t+1},y_t,\xi_2^t)\|^2 + 8\gamma_1^2\|z_t - z_{t+1}\|^2 \\
\leq& 4\|x^*(z_t) - x^*(y_t^+(z_t),z_t)\|^2 + 8\gamma_2^2\tau_2^2 l^2\|x^*(y_t) - x_{t+1}\|^2 + \\
&8\gamma_2^2\tau_2^2\|\nabla_y f(x_{t+1},y_t) - G_y(x_{t+1},y_t,\xi_2^t)\|^2 + 8\gamma_1^2\|z_t - z_{t+1}\|^2,
\end{aligned}$$

where in the second and last inequality we use Lemma C.1, and in the third inequality we use the definition of $y_t^+(z_t)$. Taking expectation and applying Lemma C.1

$$\begin{aligned}
\mathbb{E}\|x^*(z_{t+1}) - x^*(y_{t+1},z_{t+1})\|^2 \leq& 8\gamma_1^2\mathbb{E}\|z_t - z_{t+1}\|^2 + 4\mathbb{E}\|x^*(z_t) - x^*(y_t^+(z_t),z_t)\|^2 + \\
&8\gamma_2^2\tau_2^2 l^2\gamma_3^2\tau_1^2\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2 + 16\gamma_2^2\tau_2^2 l^2\tau_1^2\sigma^2 + 8\gamma_2^2\tau_2^2\sigma^2. \qquad (35)
\end{aligned}$$

Plugging (35) and (34) into (32), we have

$$\begin{aligned}\mathbb{E}V_t - \mathbb{E}V_{t+1} \geq & \left[\frac{\tau_1}{2} - l\tau_1^2\nu^{-1}\gamma_3^2 - \left(\frac{\tau_2}{2} - \tau_2^2 l\nu\right) l^2\gamma_3^2\tau_1^2 - 48p\beta\gamma_2^2\tau_2^2 l^2\gamma_3^2\tau_1^2\right]\mathbb{E}\|\nabla_x \hat{f}(x_t, y_t; z_t)\|^2 - \\ & 24p\beta\mathbb{E}\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\|^2 + \left(\frac{\tau_2}{4} - \frac{\tau_2^2 l\nu}{2}\right)\mathbb{E}\|\nabla_y \hat{f}(x^*(y_t, z_t), y_t; z_t)\|^2 + \\ & \left[\frac{p}{2\beta} - 2p\gamma_1 - \frac{p}{6\beta} - 48p\beta\gamma_1^2\right]\mathbb{E}\|z_t - z_{t+1}\|^2 - \left[\frac{l}{2} + L_\Psi + 48p\beta\gamma_2^2\right]\tau_2^2\sigma^2 - \\ & \left[\frac{p+l}{2} + 2l\nu^{-1} + 96p\beta\gamma_2^2\tau_2^2 l^2 + 2l^2\left(\frac{\tau_2}{2} - \tau_2^2 l\nu\right)\right]\tau_1^2\sigma^2 \\ \geq & \frac{\tau_1}{4}\mathbb{E}\|\nabla_x \hat{f}(x_t, y_t; z_t)\|^2 + \frac{\tau_2}{8}\mathbb{E}\|\nabla_y \hat{f}(x^*(y_t, z_t), y_t; z_t)\|^2 + \frac{p}{4\beta}\mathbb{E}\|z_t - z_{t+1}\|^2 - \\ & 24p\beta\mathbb{E}\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\|^2 - 2l\tau_1^2\sigma^2 - 5l\tau_2^2\sigma^2, \end{aligned} \tag{36}$$

where in the last inequality we note that by our choice of τ_1, τ_2, p and β we have $\gamma_1 = 2$, $\gamma_2 = 3$ and $\gamma_3 = \frac{2}{\tau_1^2 l^2} + 2$ and therefore as we choose $\nu = \frac{1}{4l\tau_2} = \frac{12}{l\tau_1}$ we have $\frac{\tau_2}{4} - \frac{\tau_2^2 l\nu}{2} = \frac{\tau_2}{8}$ and

$$\begin{aligned} & l\tau_1^2\nu^{-1}\gamma_3^2 + \left(\frac{\tau_2}{2} - \tau_2^2 l\nu\right) l^2\gamma_3^2\tau_1^2 + 48p\beta\gamma_2^2\tau_2^2 l^2\gamma_3^2\tau_1^2 \\ = & \left[\nu^{-1}(l\tau_1\gamma_3^2) - \frac{1}{\tau_1}\frac{\tau_2}{4}(l^2\tau_1^2\gamma_3^2) + 486 l\beta\frac{\tau_2^2}{\tau_1}(l^2\tau_1^2\gamma_3^2)\right]\tau_1 \\ \leq & \left[2\nu^{-1}\left(\frac{1}{\tau_1 l} + \tau_1 l\right) + \frac{1}{96}\left(1 + \tau_1^2 l^2\right) + \frac{486 \times 2}{48 \times 1600} l\mu\tau_2^2\left(1 + \tau_1^2 l^2\right)\right]\tau_1 \\ \leq & \left[\frac{20}{9\nu}\frac{1}{\tau_1 l} + \frac{1}{96}\left(1 + \frac{1}{9}\right) + \frac{486 \times 2}{48 \times 1600}\left(1 + \frac{1}{9}\right) l\mu\tau_2^2\right]\tau_1 \leq \frac{\tau_1}{4}, \end{aligned}$$

and

$$\frac{p+l}{2} + 2l\nu^{-1} + 96p\beta\gamma_2^2\tau_2^2 l^2 + 2l^2\left(\frac{\tau_2}{2} - \tau_2^2 l\nu\right) \leq \left[\frac{3}{2} + \frac{\tau_1 l}{12} + \frac{96 \times 2 \times 9}{1600} l^2\mu\tau_2^3 + \frac{\tau_2 l}{2}\right] l \leq 2l,$$

and

$$\frac{l}{2} + L_\Psi + 48p\beta\gamma_2^2 \leq \left[\frac{1}{2} + 4 + 48 \times 2 \times 4 \times 9\beta\right] l \leq 5l,$$

and

$$\frac{p}{2\beta} - 2p\gamma_1 - \frac{p}{6\beta} - 48p\beta\gamma_1^2 \geq \left[\frac{1}{3} - 4\beta - 192\beta^2\right]\frac{p}{\beta} \geq \frac{p}{4\beta}.$$

Stationary Measure: First we note that

$$\begin{aligned}\|\nabla_x f(x_t, y_t)\| \leq \|\nabla_x \hat{f}(x_t, y_t; z_t)\| + p\|x_t - z_t\| \leq & \|\nabla_x \hat{f}(x_t, y_t; z_t)\| + p\|x_t - x_{t+1}\| + p\|x_{t+1} - z_t\| \\ \leq & \|\nabla_x \hat{f}(x_t, y_t; z_t)\| + p\tau_1\|\hat{G}_x(x_t, y_t, \xi_1^t; z_t)\| + p\|x_{t+1} - z_t\|.\end{aligned}$$

Taking square and expectation

$$\begin{aligned}\mathbb{E}\|\nabla_x f(x_t, y_t)\|^2 \leq & 6\mathbb{E}\|\nabla_x \hat{f}(x_t, y_t; z_t)\|^2 + 6p^2\tau_1^2\mathbb{E}\|\nabla_x \hat{f}(x_t, y_t; z_t)\|^2 + 6p^2\mathbb{E}\|x_{t+1} - z_t\|^2 + 6p^2\tau_1^2\sigma^2 \\ = & 6(1 + p^2\tau_1^2)\mathbb{E}\|\nabla_x \hat{f}(x_t, y_t; z_t)\|^2 + 6p^2\mathbb{E}\|x_{t+1} - z_t\|^2 + 6p^2\tau_1^2\sigma^2.\end{aligned} \tag{37}$$

Also,

$$\begin{aligned}\|\nabla_y f(x_t, y_t)\| \leq & \|\nabla_y f(x_{t+1}, y_t)\| + \|\nabla_y f(x_t, y_t) - \nabla_y f(x_{t+1}, y_t)\| \\ \leq & \|\nabla_y f(x_{t+1}, y_t)\| + l\|x_{t+1} - x_t\| \\ \leq & l\tau_1\|\hat{G}_x(x_t, y_t, \xi_1^t; z_t)\| + \|\nabla_y \hat{f}(x^*(y_t, z_t), y_t; z_t)\| + \|\nabla_y \hat{f}(x^*(y_t, z_t), y_t; z_t) - \nabla_y f(x_{t+1}, y_t)\| \\ \leq & l\tau_1\|\hat{G}_x(x_t, y_t, \xi_1^t; z_t)\| + \|\nabla_y \hat{f}(x^*(y_t, z_t), y_t; z_t)\| + l\|x_{t+1} - x^*(y_t, z_t)\|.\end{aligned}$$

Taking square, taking expectation and applying Lemma C.1

$$
\begin{aligned}
&\mathbb{E}\|\nabla_y f(x_t,y_t)\|^2\\
\le& 6l^2\tau_1^2\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2+6l^2\tau_1^2\sigma^2+6\mathbb{E}\|\nabla_y\hat{f}(x^*(y_t,z_t),y_t;z_t)\|^2+6l^2\gamma_3^2\tau_1^2\mathbb{E}\|\nabla_x\hat{f}(x_t,y_t;z_t)\|^2+12l^2\tau_1^2\sigma^2\\
\le& 6l^2\tau_1^2(1+\gamma_3^2)\mathbb{E}\|\nabla_x \hat{f}(x_t,y_t;z_t)\|^2+6\mathbb{E}\|\nabla_y\hat{f}(x^*(y_t,z_t),y_t;z_t)\|^2+18l^2\tau_1^2\sigma^2. \qquad (38)
\end{aligned}
$$

Combining with (37),

$$
\begin{aligned}
&\mathbb{E}\|\nabla_x f(x_t,y_t)\|^2+\kappa\mathbb{E}\|\nabla_y f(x_t,y_t)\|^2\\
\le& 6(1+p^2\tau_1^2+\kappa l^2\tau_1^2+\kappa l^2\gamma_3^2\tau_1^2)\mathbb{E}\|\nabla_x\hat{f}(x_t,y_t;z_t)\|^2+6\kappa\mathbb{E}\|\nabla_y\hat{f}(x^*(y_t,z_t),y_t;z_t)\|^2+\\
&6p^2\mathbb{E}\|x_{t+1}-z_t\|^2+(6p^2+18\kappa l^2)\tau_1^2\sigma^2\\
\le& 24\kappa\mathbb{E}\|\nabla_x\hat{f}(x_t,y_t;z_t)\|^2+6\kappa\mathbb{E}\|\nabla_y\hat{f}(x^*(y_t,z_t),y_t;z_t)\|^2 v+6p^2\mathbb{E}\|x_{t+1}-z_t\|^2+42\kappa l^2\tau_1^2\sigma^2, \qquad (39)
\end{aligned}
$$

where in the last inequality we use $6p^2+18\kappa l^2=24l^2+18\kappa l^2\le 42\kappa l^2$ and

$$
\begin{aligned}
1+p^2\tau_1^2+kl^2\tau_1^2+\kappa l^2\gamma_3^2\tau_1^2 =& 1+4l^2\tau_1^2+\kappa l^2\tau_1^2+2\kappa(1+\tau_1^2l^2)\\
\le& \frac{13}{9}+2\kappa+3\kappa l^2\tau_1^2\le 4\kappa.
\end{aligned}
$$

Putting pieces together: From Lemma C.2,

$$
\begin{aligned}
24p\beta\|x^*(z)-x^*(y^+(z),z)\|^2\le& \frac{24p\beta}{(p-l)\mu}\left(1+\tau_2 l+\frac{\tau_2 l(p+l)}{p-l}\right)^2\|\nabla_y\hat{f}(x^*(y,z),y;z)\|^2\\
\le& \frac{1}{16}\tau_2\|\nabla_y\hat{f}(x^*(y_t,z_t),y_t;z_t)\|^2,
\end{aligned}
$$

where in the second inequality we use

$$
\frac{24p\beta}{(p-l)\mu}\left(1+\tau_2 l+\frac{\tau_2 l(p+l)}{p-l}\right)^2=\frac{48\beta}{\mu}\left(1+\tau_2 l+3\tau_2 l\right)^2\le\frac{96\beta}{\mu}\le\frac{1}{16}\tau_2.
$$

Plugging into (36),

$$
\mathbb{E}V_t-\mathbb{E}V_{t+1}\ge\frac{\tau_1}{4}\mathbb{E}\|\nabla_x\hat{f}(x_t,y_t;z_t)\|^2+\frac{\tau_2}{16}\mathbb{E}\|\nabla_y\hat{f}(x^*(y_t,z_t),y_t;z_t)\|^2+\frac{p\beta}{4}\mathbb{E}\|z_t-x_{t+1}\|^2-2l\tau_1^2\sigma^2-5l\tau_2^2\sigma^2.
$$

Plugging into (39),

$$
\begin{aligned}
&\mathbb{E}\|\nabla_x f(x_t,y_t)\|^2+\kappa\mathbb{E}\|\nabla_y f(x_t,y_t)\|^2\\
\le& 24\kappa\mathbb{E}\|\nabla_x\hat{f}_x(x_t,y_t;z_t)\|^2+6\kappa\mathbb{E}\|\nabla_y\hat{f}(x^*(y_t,z_t),y_t;z_t)\|+6p^2\mathbb{E}\|x_{t+1}-z_t\|^2+42\kappa l^2\tau_1^2\sigma^2\\
\le& \max\left\{\frac{96\kappa}{\tau_1},\frac{96\kappa}{\tau_2},\frac{24p}{\beta}\right\}\left[\mathbb{E}V_t-\mathbb{E}V_{t+1}+2l\tau_1^2\sigma^2+5l\tau_2^2\sigma^2\right]+42\kappa l^2\tau_1^2\sigma^2\\
\le& \frac{O(1)\kappa}{\tau_2}[\mathbb{E}V_t-\mathbb{E}V_{t+1}]+\frac{O(1)\kappa l\tau_1^2}{\tau_2}\sigma^2+O(1)\kappa l\tau_2\sigma^2+O(1)\kappa l^2\tau_1^2\sigma^2\\
\le& \frac{O(1)\kappa}{\tau_1}[\mathbb{E}V_t-\mathbb{E}V_{t+1}]+O(1)\kappa l\tau_1\sigma^2+O(1)\kappa l^2\tau_1^2\sigma^2\\
\le& \frac{O(1)\kappa}{\tau_1}[\mathbb{E}V_t-\mathbb{E}V_{t+1}]+O(1)\kappa l\tau_1\sigma^2, \qquad (40)
\end{aligned}
$$

where in the second and fourth inequality we use $\tau_1=48\tau_2$ and $p/\beta=3200\kappa/\tau_2$. Telescoping,

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x f(x_t,y_t)\|^2+\kappa\mathbb{E}\|\nabla_y f(x_t,y_t)\|^2\le\frac{O(1)\kappa}{T\tau_1}[V_0-\min_{x,y,z}V(x,y,z)]+O(1)\kappa l\tau_1\sigma^2.
$$

Note that since for any z we can find x, y such that $(\hat{f}(x,y;z) - \Psi(y;z)) + (P(z) - \Psi(y;z)) = 0$,

$$
\begin{aligned}
&V_0 - \min_{x,y,z} V(x,y,z) \\
=&P(z_0) + (\hat{f}(x_0,y_0;z_0) - \Psi(y_0;z_0)) + (P(z_0) - \Psi(y_0;z_0)) - \min_{x,y,z}[P(z) + (\hat{f}(x,y;z) - \Psi(y;z)) + (P(z) - \Psi(y;z))] \\
\leq&(P(z_0) - \min_z P(z)) + (\hat{f}(x_0,y_0;z_0) - \Psi(y_0;z_0)) + (P(z_0) - h(y_0;z_0)).
\end{aligned}
$$

Note that for any z

$$
P(z) = \min_x \max_y f(x,y) + l\|x-z\|^2 = \min_x \Phi(x) + l\|x-z\|^2 = \Phi_{1/2l}(z) \leq \Phi(z),
$$

and $P(z) = \Phi_{1/2l}(z)$ also implies $\min_z P(z) = \min_x \Phi(x)$. Hence

$$
V_0 - \min_{x,y,z} V(x,y,z) \leq (\Phi(z_0) - \min_x \Phi(x)) + (\hat{f}(x_0,y_0;z_0) - \Psi(y_0;z_0)) + (P(z_0) - \Psi(y_0;z_0)). \tag{41}
$$

With $b = (\hat{f}(x_0,y_0;z_0) - \Psi(y_0;z_0)) + (P(z_0) - \Psi(y_0;z_0))$, we write

$$
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla_x f(x_t,y_t)\|^2 + \kappa \mathbb{E}\|\nabla_y f(x_t,y_t)\|^2 \leq \frac{O(1)\kappa}{T\tau_1}[\Delta + b] + O(1)\kappa l \tau_1 \sigma^2.
$$

with $\Delta = \Phi(z_0) - \Phi^*$. Picking $\tau_1 = \min\left\{\frac{\sqrt{\Phi(x_0)-\Phi^*}}{2\sigma\sqrt{Tl}}, \frac{1}{3l}\right\}$,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla_x f(x_t,y_t)\|^2 + \kappa \mathbb{E}\|\nabla_y f(x_t,y_t)\|^2 &\leq \max\left\{\frac{2\sigma\sqrt{Tl}}{\sqrt{\Delta}}, 3l\right\}\frac{O(1)\kappa}{T}[\Phi(z_0) - \Phi^* + b] + \frac{O(1)\sqrt{\Delta}}{2\sigma\sqrt{Tl}} \cdot \kappa l \tau_1 \sigma^2 \\
&\leq \frac{O(1)\kappa}{T}[\Delta + b] + \frac{O(1)\kappa\sqrt{l}b}{\sqrt{\Delta T}}\sigma + \frac{O(1)\kappa\sqrt{l\Delta}}{\sqrt{T}}\sigma.
\end{aligned}
$$

We reach our conclusion by noting that $b \leq 2\,\mathrm{gap}_{\hat{f}(\cdot,\cdot;z_0)}(x_t,y_t)$.

■

D CATALYST-AGDA

Algorithm 3 Catalyst-AGDA

Input: (x_0, y_0), step sizes $\tau_1 > 0, \tau_2 > 0$.
for all $t = 0, 1, 2, ..., T-1$ **do**
  Let $k = 0$ and $x_0^0 = x_0$.
  repeat
    $y_{k+1}^t = y_k^t + \tau_2 \nabla_y f(x_k^t, y_k^t)$
    $x_{k+1}^t = x_k^t - \tau_1[\nabla_x f(x_k^t, y_{k+1}^t) + 2l(x_k^t - x_0^t)]$
    $k = k+1$
  until $\mathrm{gap}_{\hat{f}_t}(x_k^t, y_k^t) \leq \beta\, \mathrm{gap}_{\hat{f}_t}(x_0^t, y_0^t)$ where $\hat{f}_t(x,y) \triangleq f(x,y) + l\|x - x_0^t\|^2$
  $x_0^{t+1} = x_{k+1}^t, \quad y_0^{t+1} = y_{k+1}^t$
end for
Output: $\tilde{x}_T$, which is uniformly sampled from $x_0^1, ..., x_0^T$

In this section,we present a new algorithm, called Catalyst-AGDA, in Algorithm 3. It iteratively solves an augmented auxiliary problem similar to Smoothed-AGDA:

$$
\hat{f}_t(x,y) \triangleq f(x,y) + l\|x - x_0^t\|^2,
$$

by AGDA with y update first[6]. The stopping criterion for the inner-loop is

$$\operatorname{gap}_{\hat{f}_t}(x_k^t, y_k^t) \leq \beta \operatorname{gap}_{\hat{f}_t}(x_0^t, y_0^t),$$

and we will specify β later. For Catalyst-AGDA, we only consider the deterministic case, in which we have the exact gradient of $f(\cdot,\cdot)$.

In this section, we use (x^t, y^t) as a shorthand for (x_0^t, y_0^t). We denote $(\hat{x}^t, \hat{y}^t)$ with $\hat{y}^t \in \hat{Y}^t$ as the optimal solution to the auxiliary problem at t-th iteration: $\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} \left[\hat{f}_t(x,y) \triangleq f(x,y) + l\|x - x^t\|^2\right]$. Define $\hat{\Phi}_t(x) = \max_y f(x,y) + l\|x-x^t\|^2$. We use $Y^*(x)$ to denote the set $\operatorname{Arg\,max}_y f(x,y)$. In the following lemma, we show the convergence of the Moreau envelop $\|\nabla\Phi_{1/2l}(x)\|^2$ when we choose β appropriately in the stopping criterion of the AGDA subroutine.

Lemma D.1 *Under Assumptions 2.1 and 2.2, define $\Delta = \Phi(x_0) - \Phi^*$, if we apply Catalyst-AGDA with $\beta = \frac{\mu^2}{4l^2}$ in the stopping criterion of the inner-loop, then we have*

$$\sum_{t=0}^{T-1} \|\nabla\Phi_{1/2l}(x^t)\|^2 \leq \frac{35l}{2}\Delta + 3la_0,$$

where $a_0 := \Phi(x_0) - f(x_0, y_0)$.

Proof Define $g_{t+1} = \operatorname{gap}_{\hat{f}_t}(x^{t+1}, y^{t+1})$. It is easy to observe that $\hat{x}^t = \operatorname{prox}_{\Phi/2l}(x^t)$. Define $\hat{\Phi}_t(x) = \max_y f(x,y) + l\|x - x^t\|^2$. By Lemma 4.3 in (Drusvyatskiy and Paquette, 2019),

$$\begin{aligned}
\|\nabla\Phi_{1/2l}(x^t)\|^2 = 4l^2\|x^t - \hat{x}^t\|^2 \leq& 8l[\hat{\Phi}_t(x^t) - \hat{\Phi}_t(\operatorname{prox}_{\Phi/2l}(x^t))] \\
\leq& 8l[\hat{\Phi}_t(x^t) - \hat{\Phi}_t(x^{t+1}) + b_{t+1}] \\
=& 8l\left\{\Phi(x^t) - \left[\Phi(x^{t+1}) + l\|x^{t+1} - x^t\|^2\right] + b_{t+1}\right\} \\
\leq& 8l[\Phi(x^t) - \Phi(x^{t+1}) + g_{t+1}],
\end{aligned} \tag{42}$$

where in the first inequality we use l-strongly convexity of $\hat{\Phi}_t$. Because $\hat{f}$ is $3l$-smooth, l-strongly convex in x and μ-PL in y, its primal and dual function are $18l\kappa$ and $18l$ smooth, respectively, by Lemma A.3. Then we have

$$\begin{aligned}
\operatorname{gap}_{\hat{f}_t}(x^t, y^t) =& \max_y \hat{f}_t(x^t, y) - \min_x \max_y \hat{f}_t(x,y) + \min_x \max_y \hat{f}_t(x,y) - \min_x \hat{f}_t(x, y_t) \\
\leq& 9l\kappa\|x^t - \hat{x}^t\|^2 + 9l\|y^t - \hat{y}^t\|^2,
\end{aligned} \tag{43}$$

for all $\hat{y}^t \in \hat{Y}^t$. For $t \geq 1$, by fixing $\hat{y}^{t-1}$ to be the projection of y^t to $\hat{Y}^{t-1}$, there exists $\hat{y}^t \in \hat{Y}^t$ so that

$$\begin{aligned}
\|y^t - \hat{y}^t\|^2 \leq& 2\|y^t - \hat{y}^{t-1}\|^2 + 2\|y^*(\hat{x}^{t-1}) - y^*(\hat{x}^t)\|^2 \\
\leq& 2\|y^t - \hat{y}^{t-1}\|^2 + 2\left(\frac{l}{\mu}\right)^2\|\hat{x}^t - \hat{x}^{t-1}\|^2 \\
\leq& 2\|y^t - \hat{y}^{t-1}\|^2 + 4\left(\frac{l}{\mu}\right)^2\|\hat{x}^t - x^t\|^2 + 4\left(\frac{l}{\mu}\right)^2\|x^t - \hat{x}^{t-1}\|^2 \\
\leq& \frac{8l}{\mu^2}g_t + 4\left(\frac{l}{\mu}\right)^2\|\hat{x}^t - x^t\|^2,
\end{aligned}$$

where we use Lemma A.3 in the second inequality, and strong-convexity and PL condition in the last inequality. By our stopping criterion and $\|\nabla\Phi_{1/2l}(x^t)\|^2 = 4l^2\|x^t - \hat{x}^t\|^2$, for $t \geq 1$

$$g_{t+1} \leq \beta \operatorname{gap}_{\hat{f}_t}(x^t, y^t) \leq 9l\kappa\beta\|x^t - \hat{x}^t\|^2 + 9l\beta\|y^t - \hat{y}^t\|^2 \leq 72\kappa^2\beta g_t + \frac{12\kappa^2\beta}{l}\|\nabla\Phi_{1/2l}(x^t)\|^2. \tag{44}$$

[6] We believe that updating x first in the subroutine will lead to the same convergence property. For simplicity, we update y first so that we can directly apply Theorem A.1.

For $t=0$, by fixing $y^*(x^0)$ to be the projection of y^0 to $Y^*(x^0)$,

$$\|y^0-\hat{y}^0\|^2 \leq 2\|y^0-y^*(x^0)\|^2+2\|\hat{y}^0-y^*(x^0)\|^2 \leq \frac{4}{\mu}a_0+2\kappa^2\|x^0-\hat{x}^0\|^2. \tag{45}$$

Because $\Phi(x)+l\|x-x^0\|^2$ is l-strongly convex, we have

$$\left(\Phi(\hat{x}^0)+l\|\hat{x}^0-x^0\|^2\right)+\frac{l}{2}\|\hat{x}^0-x^0\|^2 \leq \Phi(x^0)=\Phi^*+(\Phi(x^0)-\Phi^*) \leq \Phi(\hat{x}^0)+(\Phi(x^0)-\Phi^*).$$

This implies $\|\hat{x}^0-x^0\|^2 \leq \frac{2}{3l}(\Phi(x^0)-\Phi^*)$. Hence, by the stopping criterion,

$$g_1 \leq \beta\,\mathrm{gap}_{\hat{f}_0}(x^0,y^0) \leq 9l\kappa\beta\|x^0-\hat{x}^0\|^2+9l\beta\|y^0-\hat{y}^0\|^2 \leq 18\kappa^2\beta\Delta+36\kappa\beta a_0. \tag{46}$$

Recursing (44) and (46), we have for $t \geq 1$

$$\begin{aligned} g_{t+1} \leq& (72\kappa^2\beta)^t g_1+\frac{12\kappa^2\beta}{l}\sum_{k=1}^{t}(72\kappa^2\beta)^{t-k}\|\nabla\Phi_{1/2l}(x_k)\|^2 \\ \leq& 18\kappa^2\beta(72\kappa^2\beta)^t\Delta+36\kappa\beta(72\kappa^2\beta)^t a_0+\frac{12\kappa^2\beta}{l}\sum_{k=1}^{t}(72\kappa^2\beta)^{t-k}\|\nabla\Phi_{1/2l}(x_k)\|^2. \end{aligned}$$

Summing from $t=0$ to $T-1$,

$$\begin{aligned} \sum_{t=0}^{T-1}g_{t+1} &= \sum_{t=1}^{T-1}g_t+g_1 \\ &\leq 18\kappa^2\beta\sum_{t=0}^{T-1}(72\kappa^2\beta)^t\Delta+36\kappa\beta\sum_{t=0}^{T-1}(72\kappa^2\beta)^t a_0+\frac{12\kappa^2\beta}{l}\sum_{t=1}^{T-1}\sum_{k=1}^{t}(72\kappa^2\beta)^{t-k}\|\nabla\Phi_{1/2l}(x_k)\|^2 \\ &\leq \frac{18\kappa^2\beta}{1-72\kappa^2\beta}\Delta+\frac{36\kappa\beta}{1-72\kappa^2\beta}a_0+\frac{12\kappa^2\beta}{l(1-72\kappa^2\beta)}\sum_{t=1}^{T-1}\|\nabla\Phi_{1/2l}(x^t)\|^2, \end{aligned} \tag{47}$$

where in the last inequality $\sum_{t=1}^{T-1}\sum_{k=1}^{t}(72\kappa^2\beta)^{t-k}\|\nabla\Phi_{1/2l}(x_k)\|^2=\sum_{k=1}^{T-1}\sum_{t=k}^{T}(72\kappa^2\beta)^{t-k}\|\nabla\Phi_{1/2l}(x_k)\|^2 \leq \sum_{k=1}^{T-1}\frac{1}{1-(72\kappa^2\beta)}\|\nabla\Phi_{1/2l}(x_k)\|^2$. Now, by telescoping (42),

$$\frac{1}{8l}\sum_{t=0}^{T-1}\|\nabla\Phi_{1/2l}(x^t)\|^2 \leq \Phi(x^0)-\Phi^*+\sum_{t=0}^{T-1}g_{t+1}.$$

Plugging (47) in,

$$\left(\frac{1}{8l}-\frac{12\kappa^2\beta}{l(1-72\kappa^2\beta)}\right)\sum_{t=0}^{T-1}\|\nabla\Phi_{1/2l}(x^t)\|^2 \leq \left(1+\frac{18\kappa^2\beta}{1-72\kappa^2\beta}\right)\Delta+\frac{36\kappa\beta}{1-72\kappa^2\beta}a_0. \tag{48}$$

With $\beta=\frac{1}{264\kappa^4}$, we have $\frac{\kappa^2\beta}{1-72\kappa^2\beta} \leq \frac{1}{192\kappa^2}$. Therefore,

$$\sum_{t=0}^{T-1}\|\nabla\Phi_{1/2l}(x^t)\|^2 \leq \frac{35l}{2}\Delta+3la_0.$$

■

Theorem D.1 *Under Assumptions 2.1 and 2.2, if we apply Catalyst-AGDA with $\beta=\frac{1}{264\kappa^4}$ in the stopping criterion of the inner-loop, then the output from Algorithm 3 satisfies*

$$\sum_{t=1}^{T}\|\nabla\Phi(x_0^t)\|^2 \leq \frac{1}{T}\sum_{t=1}^{T}\|\nabla\Phi(x^{t+1})\|^2 \leq \frac{19l}{T}\Delta+\frac{6l}{T}a_0 \tag{49}$$

which implies the outer-loop complexity of $O(l\Delta\epsilon^{-2})$. Furthermore, if we choose $\tau_1 = \frac{1}{3l}$ and $\tau_2 = \frac{1}{486l}$, it takes $K = O(\kappa \log(\kappa))$ inner-loop iterations to satisfy the stopping criterion. Therefore, the total complexity is $O(\kappa l \Delta\epsilon^{-2}\log\kappa)$.

Proof We separate the proof into two parts: 1) outer-loop complexity 2) inner-loop convergence rate.

Outer-loop: We still denote $g_{t+1} = \mathrm{gap}_{\hat{f}_t}(x^{t+1}, y^{t+1})$. First, note that

$$\begin{aligned}\|\nabla\Phi(x^{t+1})\|^2 &\le 2\|\nabla\Phi(x^{t+1}) - \nabla\Phi(\hat{x}^t)\|^2 + 2\|\nabla\Phi(\hat{x}^t)\|^2 \\ &\le 2\left(\frac{2l^2}{\mu}\right)\|x^{t+1} - \hat{x}^t\|^2 + 2\|\nabla\Phi_{1/2l}(x^t)\|^2 \\ &\le \frac{16l^3}{\mu^2} g_{t+1} + 2\|\nabla\Phi_{1/2l}(x^t)\|^2. \end{aligned} \tag{50}$$

where in the second inequality we use Lemma A.1 and Lemma 4.3 in (Drusvyatskiy and Paquette, 2019). Summing from $t = 0$ to $T-1$, we have

$$\sum_{t=0}^{T-1}\|\nabla\Phi(x^{t+1})\|^2 \le \frac{16l^3}{\mu^2}\sum_{t=0}^{T-1} g_{t+1} + 2\sum_{t=0}^{T-1}\|\nabla\Phi_{1/2l}(x^t)\|^2. \tag{51}$$

Applying (47), we have

$$\sum_{t=0}^{T-1}\|\nabla\Phi(x^{t+1})\|^2 \le \left[\frac{16l^3}{\mu^2}\cdot\frac{12\kappa^2\beta}{l(1-72\kappa^2\beta)} + 2\right]\sum_{t=1}^{T-1}\|\nabla\Phi_{1/2l}(x^t)\|^2 + \frac{16l^3}{\mu^2}\cdot\frac{18\kappa^2\beta}{1-72\kappa^2\beta}\Delta + \frac{16l^3}{\mu^2}\cdot\frac{36\kappa\beta}{1-72\kappa^2\beta}a_0,$$

With $\beta = \frac{1}{264\kappa^4}$, we have

$$\sum_{t=0}^{T-1}\|\nabla\Phi(x^{t+1})\|^2 \le 3\sum_{t=1}^{T-1}\|\nabla\Phi_{1/2l}(x^t)\|^2 + \frac{3l}{2}\Delta + 3la_0.$$

Applying Lemma D.1,

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\Phi(x^{t+1})\|^2 \le \frac{19l}{T}\Delta + \frac{6l}{T}a_0.$$

Inner-loop: The objective of auxiliary problem $\min_x \max_y \hat{f}_t(x, y) \triangleq f(x, y) + l\|x - x_0^t\|^2$ is $3l$-smooth and (l, μ)-SC-PL. We denote the dual function of the auxiliary problem by $\hat{\Psi}^t(y) = \min_x \hat{f}_t(x, y)$. We also define

$$P_k^t \triangleq \left[\max_y \hat{\Psi}^t(y) - \hat{\Psi}^t(y_k^t)\right] + \frac{1}{10}\left[\hat{f}_t(x_k^t, y_k^t) - \hat{\Psi}^t(y_k^t)\right].$$

By Theorem A.1, AGDA with stepsizes $\tau_1 = \frac{1}{3l}$ and $\tau_2 = \frac{l^2}{18(3l)^3} = \frac{1}{486l}$ satisfies

$$P_k^t \le \left(1 - \frac{\mu}{972l}\right)^k P_0^t.$$

We denote $x_*^t(y) = \arg\min_x \hat{f}_t(x, y)$. We note that

$$\begin{aligned}\|x_k^t - \hat{x}^t\|^2 =& 2\|x_k^t - x_*^t(y_k^t)\|^2 + 2\|x_*^t(y_k^t) - \hat{x}^t\|^2 \\ =& 2\|x_k^t - x_*^t(y_k^t)\|^2 + 2\|x_*^t(y_k^t) - x_*^t(\hat{y}^t)\|^2 \\ \le& \frac{4}{l}\left[\hat{f}_t(x_k^t, y_k^t) - \hat{\Psi}^t(y_k^t)\right] + 2\left(\frac{3l}{\mu}\right)^2\|y_k^t - \hat{y}^t\|^2 \\ \le& \frac{4}{l}\left[\hat{f}_t(x_k^t, y_k^t) - \hat{\Psi}^t(y_k^t)\right] + \frac{36l^2}{\mu^3}[\hat{\Psi}^t(\hat{y}^t) - \hat{\Psi}^t(y_k^t)] \\ \le& \left(\frac{40}{l} + \frac{36l^2}{\mu^3}\right)\left(1 - \frac{\mu}{972l}\right)^k P_0^t, \end{aligned} \tag{52}$$

where in the first inequality we use l-strong convexity of $\hat{f}_t(\cdot, y_k^t)$ and Lemma A.1, and in the second inequality we use μ-PL of $\hat{\Psi}^t$ and Lemma A.2. Since $\hat{\Phi}^t$ is smooth by Lemma A.3,

$$\hat{\Phi}^t(x_k^t) - \hat{\Phi}^t(\hat{x}^t) \leq \frac{2(3l)^2}{2\mu}\|x_k^t - \hat{x}^t\|^2 \leq \frac{9l^2}{\mu}\left(\frac{40}{l} + \frac{36l^2}{\mu^3}\right)\left(1 - \frac{\mu}{972l}\right)^k P_0^t. \tag{53}$$

Therefore,

$$\begin{aligned}\operatorname{gap}_{\hat{f}_t}(x_k^t, y_k^t) = \hat{\Phi}^t(x_k^t) - \hat{\Phi}^t(\hat{x}^t) + \hat{\Psi}^t(\hat{y}^t) - \hat{\Psi}^t(y_k^t) &\leq \left[\frac{9l^2}{\mu}\left(\frac{40}{l} + \frac{36l^2}{\mu^3}\right) + 1\right]\left(1 - \frac{\mu}{972l}\right)^k P_0^t \\ &\leq 754\kappa^4\left(1 - \frac{1}{972\kappa}\right)^k \operatorname{gap}_{\hat{f}_t}(x_0^t, y_0^t).\end{aligned}$$

where in the last inequality we note that $P_0^t \leq \frac{11}{10}\operatorname{gap}_{\hat{f}_t}(x_0^t, y_0^t)$. So after $K = O(\kappa\log(\kappa))$ iterations of AGDA, the stopping criterion $\operatorname{gap}_{\hat{f}_t}(x_k^t, y_k^t) \leq \beta\operatorname{gap}_{\hat{f}_t}(x_0^t, y_0^t)$ can be satisfied.

■

Remark D.1 *The theorem above implies that Catalyst-AGDA can achieve the complexity of $\tilde{O}(\kappa l\Delta\epsilon^{-2})$ to find ϵ-stationary point of Φ in the deterministic setting, which is comparable to the complexity of Smoothed-AGDA up to a logarithmic term in κ but does not require additional translation as in Corollary 4.1.*

E ADDITIONAL EXPERIMENTS

In this section, we show the tuning of Adam, RMSprop and Stochastic AGDA (SAGDA) for the task of training a toy *regularized* linear WGAN and a toy *regularized* neural WGAN (one hidden layer). All details on these models are given in the experimental section in the main paper. This section motivates that the smoothed version of stochastic AGDA has superior performance compared to stochastic AGDA that is carefully tuned (see Figures 4 and 6). Often, the performance is comparable to Adam and RMSprop, if not better (see Figures 5 and 7). Findings are similar both for the linear and the neural net cases. We note, as in the main paper, that the stochastic nature of the gradients makes the algorithms converge fast in the beginning and slow down later on.

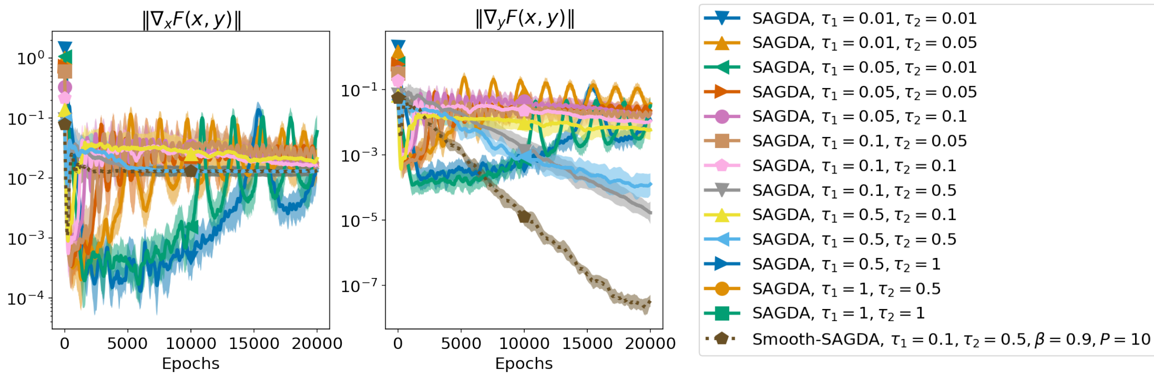

Figure 4: Training of a **Linear WGAN** (see experiment section in the main paper for details). Stochastic **AGDA** (SAGDA) is compared to the tuned version of Smoothed SAGDA (best), for different choices of learning rates. Shown is the mean of 3 independent runs and one standard deviation. Smoothing provides acceleration.

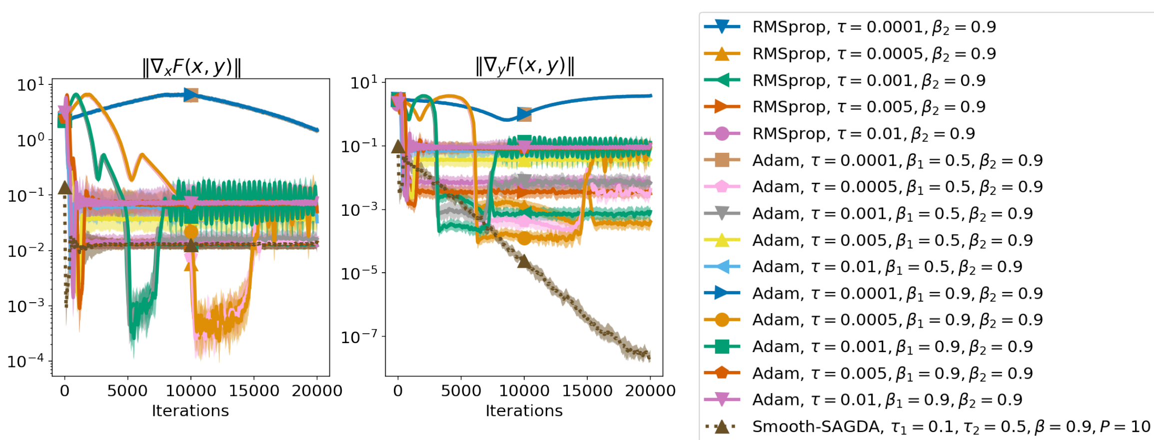

Figure 5: Training of a **Linear WGAN** (see experiment section in the main paper for details). **Adam and RMSprop** (same learning rate for generator and critic) are compared to the tuned version of Smoothed SAGDA (best), for different choices hyperparameters. Shown is the mean of 3 independent runs and one standard deviation. Smoothing also in this setting provides acceleration.

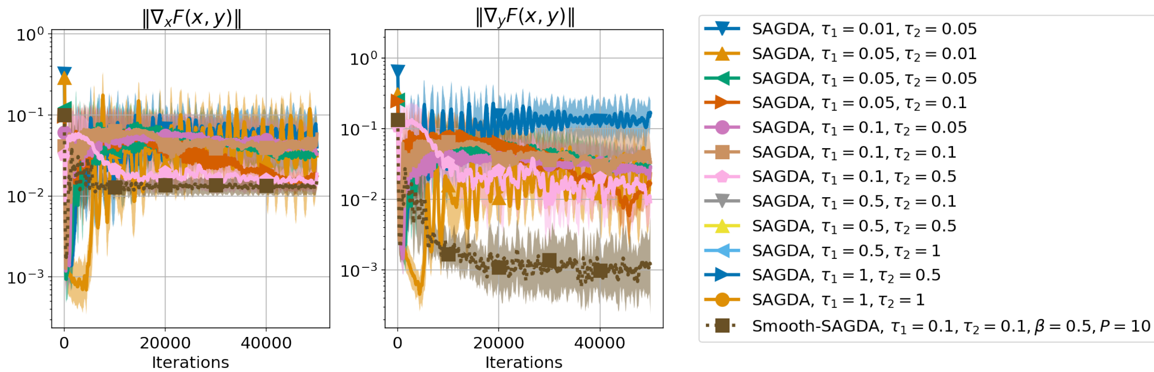

Figure 6: Training of a **Neural WGAN** (see experiment section in the main paper for details). Stochastic **AGDA** (SAGDA) is compared to the tuned version of Smoothed SAGDA (best) for different choices of learning rates. Shown is the mean of 3 independent runs and one standard deviation. Smoothing provides acceleration.

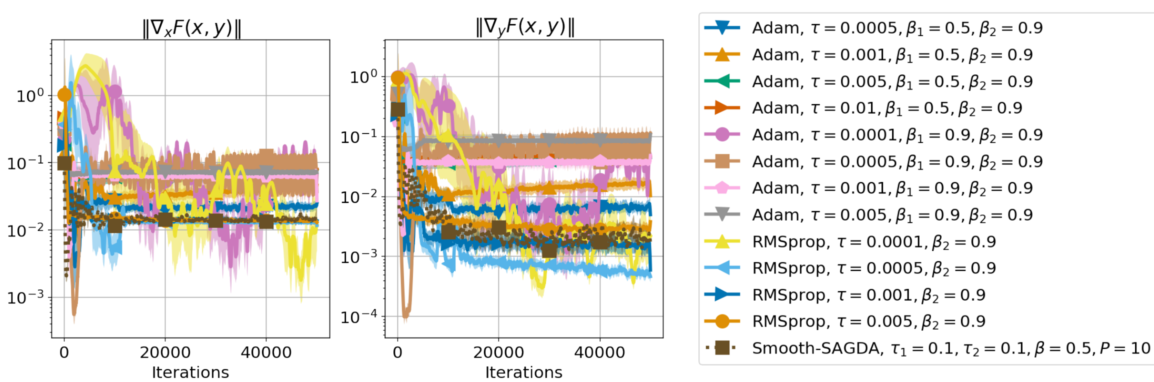

Figure 7: Training of a **Neural WGAN** (see experiment section in the main paper for details). **Adam and RMSprop** (same learning rate for the generator and critic) are compared to the tuned version of Smoothed SAGDA, for different choices of the hyperparameters. Shown is the mean of 3 independent runs and 1/2 standard deviation (for better visibility). Performance is slightly worse than RMSprop tuned at best. As mentioned in the main paper, we believe a combination of adaptive stepsizes and smoothing would lead to the best results.