Local AdaGrad-Type Algorithm for Stochastic Convex-Concave Minimax Problems

Luofeng Liao
JD Explore Academy
liaoluofeng96@gmail.com

Li Shen*
JD Explore Academy
mathshenli@gmail.com

Jia Duan
JD Explore Academy
xuelandj@gmail.com

Mladen Kolar
University of Chicago
mkolar@chicagobooth.edu

Dacheng Tao
JD Explore Academy
dacheng.tao@gmail.com

Abstract

Large scale convex-concave minimax problems arise in numerous applications, including game theory, robust training, and training of generative adversarial networks. Despite their wide applicability, solving such problems efficiently and effectively is challenging in the presence of large amounts of data using existing stochastic minimax methods. We study a class of stochastic minimax methods and develop a communication-efficient distributed stochastic extragradient algorithm, LocalAdaSEG, with an adaptive learning rate suitable for solving convex-concave minimax problem in the Parameter-Server model. LocalAdaSEG has three main features: (i) periodic communication strategy reduces the communication cost between workers and the server; (ii) an adaptive learning rate that is computed locally and allows for tuning-free implementation; and (iii) theoretically, a nearly linear speed-up with respect to the dominant variance term, arising from estimation of the stochastic gradient, is proven in both the smooth and nonsmooth convex-concave settings. LocalAdaSEG is used to solve a stochastic bilinear game, and train generative adversarial network. We compare LocalAdaSEG against several existing optimizers for minimax problems and demonstrate its efficacy through several experiments in both the homogeneous and heterogeneous settings.

1 Introduction

Stochastic minimax optimization problems arise in applications ranging from game theory [46], robust optimization [19], and AUC Maximization [28], to adversarial learning [52] and training of generative adversarial networks (GANs) [27]. In this work, we consider

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\left\{F(x,y):=\mathbb{E}_{\xi\sim P}\big[f(x,y,\xi)\big]\right\}, \tag{1}$$

where $\mathcal{X}\subseteq\mathbb{X}$, $\mathcal{Y}\subseteq\mathbb{Y}$ are nonempty compact convex sets, $\mathbb{X}$, $\mathbb{Y}$ are finite dimensional vector spaces, ξ is a random vector with an unknown probability distribution P supported on a set Ξ, and $f:\mathcal{X}\times\mathcal{Y}\times\Xi\to\mathbb{R}$ is a real valued function, which may be nonsmooth. Throughout the paper, we assume that the expectation $\mathbb{E}_{\xi\sim P}\big[f(x,y,\xi)\big]$ is well-defined and finite. For all $\xi\in\Xi$, we assume that the function $F(x,y)$ is convex in $x\in\mathcal{X}$ and concave in $y\in\mathcal{Y}$. In addition, we assume that $F(x,y)$ is a Lipschitz continuous function.

There are three main challenges in developing a solver for the minimax problem (1). First, the solver should provide iterates that converge. Second, the solver should be computationally efficient so that it

*Li Shen is the corresponding author.

Preprint. Under review.

can be applied to problems with large amounts of training samples. Third, the solver should provide an adaptive way to set the learning rate. We discuss these challenges in detail below.

First, from a theoretical perspective, it has been demonstrated that direct application of the (stochastic) gradient descent ascent ((S)GDA) to solve (1) may result in divergence of the iterates [41, 18, 26, 42]. Possible ways to overcome the divergence issue are to apply primal-dual hybrid gradient(PDHG) or (stochastic) extragradient method and their variants [41, 18, 25, 4, 38, 56, 55].

Second, from a computational perspective, it is desirable to have a distributed solver to solve the stochastic minimax problem (1). The minimax problem (1) is often instantiated as a finite-sum problem, where the distribution P is the empirical distribution over the data points. Storing and manipulating large-scale datasets on a single worker is challenging, while computing the exact (sub)gradient of $F(x, y)$ is also impossible. For example, when problem (1) is specified as BigGAN [8] over ImageNet [20], the number of training samples is as many as 14 million. Traditional distributed SGDA on problem (1) may suffer from considerable communication burden due to such a large amount of samples. Sometimes data are intrinsically distributed on multiple devices (such as cell-phone data) and, due to privacy concerns, local data must stay on the device, which further motivates the development of distributed solvers. Communication-efficient distributed large-scale solvers for minimax problems have been investigated only recently [7, 22, 30, 39].

Third, from an adaptive learning perspective, the performance of stochastic minimax solvers for (1) is highly dependant on the learning rate tuning mechanism [29, 1]. However, designing a solver for (1) with an adaptive learning rate is much more challenging compared to the convex case. For example, for classical minimization problems, the learning rate can be tuned based on the loss evaluated at the current iterate, which directly quantifies how close the iterate is to the minimum. However, such an approach does not extend to minimax problems since the value of F at an iterate (x, y) does not serve as a performance criterion and, therefore, a more sophisticated approach is required for tuning the learning rate. Development of adaptive learning rate tuning mechanisms for large scale stochastic minimax problems has been explored only recently [6, 5, 24, 1, 38]. Hence, we ask

Can we develop an efficient algorithm for the stochastic minimax problem (1) *that enjoys convergence guarantees, communication-efficiency and adaptivity* ***simultaneously*** *?*

We provide an affirmative answer to this question and develop LocalAdaSEG (Local Adaptive Stochastic Extragradient) algorithm. Our contributions are three-fold:

Novel communication-efficient distributed minimax algorithm. Specifically, LocalAdaSEG algorithm falls under the umbrella of the Parameter-Server model [50] and adopts a periodic communication mechanism to reduce the communication cost between the server and workers, similar to Local SGD/FedAvg [54, 51, 35] in federated learning [40]. In addition, in each worker, a local stochastic extragradient algorithm with an adaptive learning rate is performed with multiple iterations independently. Every once in a while, current iterates and adaptive learning rates from all workers are sent to the server. Then a weighted average of the iterates is computed, where the weights are constructed from the received local adaptive learning rates. We emphasize that the adaptive learning in each worker is distinct from others and is automatically updated according to local data as is done in [11, 7], and different from the existing adaptive distributed algorithms [53, 47, 12].

Theoretically optimal convergence rate. We establish the optimal $\tilde{O}\left(\gamma GD/\sqrt{T} + \sigma D/\sqrt{MT}\right)$ convergence rate in the nonsmooth stochastic convex-concave minimax setting and optimal $\tilde{O}\left(\sigma D/\sqrt{MT} + DM^{3/2} \cdot \mathcal{V}_1(T)/T + \gamma^2 LD^2/T + \gamma GDM^{3/2}/T\right)$ convergence rate in the smooth stochastic convex-concave minimax setting with respect to duality gap [44, 36], where M is the number of workers and $\mathcal{V}_1(T)$ is the cumulative growth of stochastic gradients. Therefore, LocalAdaSEG algorithm enjoys the linear speed-up property on the stochastic gradient variance term thanks to the periodic communication mechanism.

Experimental verification. We conduct several experiments on stochastic bilinear game and Wasserstein GAN [3] to verify the efficiency and effectiveness of LocalAdaSEG algorithm. We also extend LocalAdaSEG algorithm to solve the challenging federated GANs in a heterogeneous setting. Experiment results agree with the theoretical guarantees and demonstrate the superiority of LocalAdaSEG against several existing minimax optimizers, such as SEGDA [44], UMP [6], ASMP [24], LocalSEGDA [7], LocalSGDA [22], and Local Adam [7].

2 Related Work

While there has been a lot of work on minmax optimization, due to space constraints, we summarize only the most closely related work. Our work is related to literature on stochastic minimax algorithms, adaptive minimax algorithms, and distributed minimax algorithms.

Stochastic minimax algorithms. Stochastic convex-concave minimax problems (1) have been extensively studied in the optimization literature and are usually solved via variants of PDHG or extragradient methods, e.g., [9, 56, 44, 45, 32, 31, 15, 7]. [16] and [32] adopted mirror-prox-type methods to tackle the stochastic convex-concave minimax problem with $\mathcal{O}(\frac{1}{\sqrt{T}})$ convergence rate. [56] proposed an accelerated stochastic PDHG-type algorithm with Bergman divergence for solving stochastic convex-concave minimax problem with a similar $\mathcal{O}(\frac{1}{\sqrt{T}})$ convergence rate dominated by the stochastic variance term. However, while all these algorithms [16, 32, 56] have achieved the optimal rate according to the low and upper bound for the stochastic convex-concave minimax problem [7], their performance is highly influenced by the choice of learning rate.

Adaptive minimax algorithms. Adaptive learning rate in stochastic optimization is first developed for minimization problems [23]. Its variants [33, 48] are extensively used for training deep learning models. The key feature of adaptive learning rate is that it can automatically tune the learning rate during the training process and achieve faster convergence. Recently, adaptive learning rate has also been developed for minimax algorithms to accelerate the training process, since the learning rate in stochastic minimax algorithm is hard to tune based on the minimax loss, as compared to minimization problems. Several recent papers have tried to analyze convergence rate of adaptive extragradient in the convex-concave minimax setting. The universal mirror-prox method [6] proposed a new adaptive learning rate technique that adapts to problem parameters, such as the unknown Lipschitz parameter, and achieves optimal convergence rates in the stochastic setting. [5] extended the universal mirror-prox of [6] by replacing the norm dependence in the learning rate with a general Bregman divergence dependence. [24] proposed an adaptive stochastic single-call extragradient algorithm for variational inequality problems. [1] proposed a similar adaptive mirror-prox algorithm, but their method handles unbounded domain by introducing the notion of local norms in the deterministic setting. Training of a GAN model [27] corresponds to solving a specific non-convex non-concave minimax problem. Several works have heuristically adopted stochastic adaptive extragradient for training GANs [25, 41, 7]. Recently, [38] studied the convergence behavior of an adaptive optimistic stochastic gradient algorithm for a class of non-convex non-concave minimax problems under the MVI condition for training GANs.

Distributed minimax algorithms. As datasets and deep learning architectures become larger and larger distributed minimax algorithms are needed for GANs and adversarial training. [7] established upper and lower bounds for iteration complexity for strongly-convex-strongly-concave and convex-concave minimax problems in both the centralized and decentralized setting. However, convergence rate for their Extra Step Local SGD is established only in a strongly-convex-strongly-concave setting with a linear speed-up property with respect to the number of works; while for their proposed local Adam no convergence results are provided. [22] provided convergence guarantees for a primal-dual local stochastic gradient algorithm in the strongly-convex-strongly-concave-setting and several non-convex settings with PL-inequality-type conditions. [14] and [39] studied convergence of a distributed optimistic stochastic gradient algorithm for non-convex non-concave minimax problem under the pseudomonotonicity condition and MVI condition, respectively. However, their convergence rates hold only for a sufficient large mini-batch size or a sufficiently large amount of workers. In addition, there also exist several decentralized or federated algorithms for stochastic strongly-convex-strongly-concave minimax problems [30, 49]. In this work, we mainly focus on the centralized setting for the stochastic convex-concave minimax problems.

Our work and the proposed LocalAdaSEG contributes to the above described literature. To the best of our knowledge, the proposed LocalAdaSEG algorithm is the first communication-efficient distributed algorithm for stochastic minimax problem and simultaneously supports adaptive learning rate and mini-batch size. Moreover, LocalAdaSEG communicates only periodically to improve the communication efficiency and uses a local adaptive learning rate, computed on local data in each worker, to improve the computation efficiency. In addition, LocalAdaSEG can be also applied in the nonsmooth setting with convergence guarantee. LocalAdaSEG can be seen as distributed extension of [6] with period communication as local SGD [51]. We note that only very recently, a local adaptive

stochastic minimax algorithm, called Local Adam, has been heuristically used for training GANs without convergence guarantee [7]. We summarize relationship to existing literature in Table 1.

Stochastic minimax algorithms	Nonsmooth ?	Comm. eff. ?	Adaptive ?
Mirror SA [45], SMP [32], SAMP [16], Optimal Stochastic PDHG-type [56]	✓	✗	✗
SCAFFOLD-Catalyst-S [30], Local SGDA [22], Extra Step Local SGD [7]	✗	✓	✗
Universal Mirror-prox [6], Adaptive Single-gradient Mirror-prox [24], Geometry-Aware Universal Mirror-prox [5], AdaProx [1]	✓	✗	✓
Optimistic AdaGrad [38]	✗*	✗	✓
Our LocalAdaSEG	✓	✓	✓

Table 1: Comparison to related works on adaptive or communication-efficient approaches to stochastic minimax problems. Here "Nonsmooth ?" asks whether the algorithm enjoys theoretical guarantees in the nonsmooth convex-concave setting; "Comm. eff. ?" asks whether the proposed algorithm is communication-efficient; "Adaptive ?" asks whether the proposed algorithm requires knowledge of problem parameters. "*": The work of [38] discusses non-convex non-concave minimax problems.

3 Notations and Assumptions

A point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is called a saddle-point for the minimax problem in (1) if

$$F(x^*, y) \leqslant F(x^*, y^*) \leqslant F(x, y^*) \qquad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{2}$$

Under the assumptions stated in Section 1, the corresponding primal, $\min_x\{\max_y F(x,y)\}$, and dual problem, $\max_y\{\min_x F(x,y)\}$, have optimal solutions and equal optimal values, denoted F^*. The pairs of optimal solutions (x^*, y^*) form the set of saddle-points of F on $\mathcal{X} \times \mathcal{Y}$. We denote $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $z = (x, y) \in \mathcal{Z}$, and $z^* = (x^*, y^*) \in \mathcal{Z}$. We use $\|\cdot\|_{\mathcal{X}}$, $\|\cdot\|_{\mathcal{Y}}$, and $\|\cdot\|_{\mathcal{Z}}$ to denote the Euclidean norms on $\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}$, respectively, and let $\|\cdot\|_{\mathcal{X},*}$, $\|\cdot\|_{\mathcal{Y},*}$ and $\|\cdot\|_{\mathcal{Z},*}$ denote the corresponding dual norms. With this notation, $\|z\|_{\mathcal{Z}} = \sqrt{\|x\|_{\mathcal{X}}^2 + \|y\|_{\mathcal{Y}}^2}$ and $\|z\|_{\mathcal{Z},*} = \sqrt{\|x\|_{\mathcal{X},*}^2 + \|y\|_{\mathcal{Y},*}^2}$. Throughout the paper, we focus on the Euclidean setting, but note that the results can readily generalize to non-Euclidean cases.

We are interested in finding a saddle-point of F over $\mathcal{X} \times \mathcal{Y}$. For a candidate solution $\tilde{z} = (\tilde{x}, \tilde{y}) \in \mathcal{Z}$, we measure its quality by the duality gap, defined as

$$\mathrm{DualGap}(\tilde{z}) := \max_{y \in \mathcal{Y}} F(\tilde{x}, y) - \min_{x \in \mathcal{X}} F(x, \tilde{y}). \tag{3}$$

The duality gap is commonly used as a performance criterion for general convex-concave minimax problems (see, e.g., [44, 36]). Note that for all $z \in \mathcal{Z}$ it holds $\mathrm{DualGap}(z) \geqslant 0$ and $\mathrm{DualGap}(z) = 0$ if and only if z is a saddle-point.

For the stochastic minimax problem (1), we assume that neither the function $F(x, y)$ nor its sub/supgradients in x and y are available. Instead, we assume access to an unbiased stochastic oracle $G(x, y, \xi) = [G_x(x, y, \xi), -G_y(x, y, \xi)]$, such that the vector $\mathbb{E}_\xi[G(x, y, \xi)]$ is well-defined and $\mathbb{E}_\xi[G(x, y, \xi)] \in [\partial_x F(x, y), -\partial_y F(x, y)]$. For notational convenience, we let

$$\tilde{G}(z) := G(x, y, \xi), \quad G(z) := \mathbb{E}_\xi[G(x, y, \xi)]. \tag{4}$$

Below, we impose assumptions on the minimax problem (1) and the stochastic gradient oracle (4).

Assumption A.1 (Bounded Domain)**.** *There exists D such that* $\sup_{z \in \mathcal{Z}} \frac{1}{2}\|z\|^2 \leqslant D^2$.

Assumption A.2 (Bounded Stochastic Gradients)**.** *There exists G such that* $\sup_{z \in \mathcal{Z}} \|\tilde{G}(z)\|_* \leqslant G$*, P-almost surely.*

Domain boundedness A.1 is commonly assumed in the convex-concave minimax literature; see the references in §1. However, we note that the assumption might be removed in certain settings. For

example, [15, 43] use a perturbation-based variant of the duality gap as the convergence criterion, [1] handles unbounded domains via the notion of local norms, while [56] handles unbounded domains with an access to a convex optimization oracle. The almost sure boundedness assumption A.2 on the gradient oracle seems restrictive, but is common in the literature on adaptive stochastic gradient methods (see, e.g., [23, 13, 6, 38]). In Remark 2 we discuss how to extend our analysis to unbounded oracles.

Assumption A.3 (Bounded Variance)**.** *There exists* σ *such that* $\sup_{z\in\mathcal{Z}}\mathbb{E}_\xi\big[\|G(z)-\tilde{G}(z)\|_*^2 \mid z\big] \leqslant \sigma^2$.

We analyze separately the case when the saddle function F is differentiable with Lipschitz gradients.

Assumption A.4 (Smoothness)**.** *Assume for all* $z, z' \in \mathcal{Z}$*, we have* $\|G(z) - G(z')\|_* \leqslant L\|z - z'\|$.

4 LocalAdaSEG Algorithm

In this section we introduce LocalAdaSEG algorithm used to solve (1) and describe its main features. Algorithm 1 details the procedure.

Algorithm 1 LocalAdaSEG$(G_0, D; K, M, R; \alpha)$

1: **Input**: G_0, a guess on the upper bound of gradients, D, the diameter of the set $\mathcal{Z}$, K, communication interval, M, the number of workers, R, number of rounds, α, base learning rate.
2: **Initialize**: $\eta_1^m = D\alpha/G_0$, $\tilde{z}_0 = \tilde{z}_0^m = \tilde{z}_0^{m,*} = 0$ for all m, and $S := \{0, K, 2K, \dots, RK\}$.
3: **for** $t = 1, \dots, T = RK$, parallel for workers $m = 1, \dots, M$ **do**
4: update learning rate

$$\eta_t^m = D\alpha\Big/\sqrt{G_0^2 + \textstyle\sum_{\tau=1}^{t-1}\big(\|z_\tau^m - \tilde{z}_{\tau-1}^{m,*}\|^2 + \|z_\tau^m - \tilde{z}_\tau^m\|^2\big)/\big(5(\eta_\tau^m)^2\big)}$$

5: **if** $t-1 \in S$ **then**
6: worker m: send $(\eta_t^m, \tilde{z}_{t-1}^m)$ to server
7: server: compute $\tilde{z}_{t-1}^\circ$, the weighted average of $\{\tilde{z}_{t-1}^m\}_{m\in[M]}$, and broadcast it to workers

$$w^m = (\eta_t^m)^{-1}\Big/\textstyle\sum_{m'=1}^M(\eta_t^{m'})^{-1} \quad \text{and} \quad \tilde{z}_{t-1}^\circ = \sum_{m=1}^M w_m \cdot \tilde{z}_{t-1}^m$$

8: worker m: set $\tilde{z}_{t-1}^{m,*} = \tilde{z}_{t-1}^\circ$
9: **else**
10: set $\tilde{z}_{t-1}^{m,*} = \tilde{z}_{t-1}^m$
11: **end if**
12: update

$$z_t^m = \Pi_{\mathcal{Z}}[\tilde{z}_{t-1}^{m,*} - \eta_t^m M_t^m] \qquad \text{with } M_t^m = \tilde{G}(\tilde{z}_{t-1}^{m,*})$$
$$\tilde{z}_t^m = \Pi_{\mathcal{Z}}[\tilde{z}_{t-1}^{m,*} - \eta_t^m g_t^m] \qquad \text{with } g_t^m = \tilde{G}(z_t^m)$$

13: **end for**
14: **Output**: $\frac{1}{TM}\sum_{m=1}^M\sum_{t=1}^T z_t^m$

The Parameter-Server model. LocalAdaSEG uses M parallel workers which, in each of R rounds, independently execute K steps of extragradient updates (Line 1). The adaptive learning rate is computed solely based on iterates occurred in the local worker (Line 1). Let $S := \{0, K, 2K, \dots, RK = T\}$ denote the time points of communication. At a time of communication ($t \in S+1$, Lines 1–1), the workers communicate and compute the weighted iterate, $\tilde{z}_{t-1}^\circ$, defined in Line 1. Then the next round begins with a common iterate $\tilde{z}_{t-1}^\circ$. Finally, LocalAdaSEG outputs the average of the sequence $\{z_t^m\}_{m\in[M],t\in[T]}$. Overall, each worker computes $T = KR$ extragradient steps locally, for a total of $2MT$ stochastic gradient calls (since each extragradient step, Line 1, requires two calls of gradient oracle) with R rounds of communication (every K steps of computation).

Extragradient step. At the time when no communication happens ($t-1 \notin S$), Line 1 reduces to

$$z_t^m = \Pi_{\mathcal{Z}}[\tilde{z}_{t-1}^m - \eta_t^m M_t^m] \qquad \text{with } M_t^m = \tilde{G}(\tilde{z}_{t-1}^m),$$
$$\tilde{z}_t^m = \Pi_{\mathcal{Z}}[\tilde{z}_{t-1}^m - \eta_t^m g_t^m] \qquad \text{with } g_t^m = \tilde{G}(z_t^m),$$

which is just the extragradient (EG) algorithm [34] that is commonly used to solve minimax problems; see references in §1.

Periodic averaging weights. The proposed weighted averaging scheme in Line 1 is different from existing works on local SGD and Local Adam [7]. At the time of averaging ($t-1\in S$), LocalAdaSEG

pulls the averaged iterate towards the local iterate with a smaller learning rate. For the homogeneous case studied in this paper, we expect $w^m \sim 1/M$.

Intuition of local adaptive learning rate scheme. The adaptive learning rate scheme (Line 1) follows that of Bach and Levy [6] closely. To develop intuition, consider the deterministic setting where $\sigma = 0$ and define $(\delta_t^m)^2 := \|g_t^m\|_*^2 + \|M_t^m\|_*^2$. If we ignore the projection operation, the learning rate η_t^m would look like $\eta_t^m \sim 1/(1 + \sum_{\tau=1}^{t-1}(\delta_\tau^m)^2)^{1/2}$. In the nonsmooth case, the subgradients might not vanish as we approach the solution (in the case of convex optimization, consider the function $f(x) = |x|$ near 0), and thus we only have $\liminf_{t\to\infty} \delta_t^m > 0$. This implies η_t^m will vanish at the rate $1/\sqrt{t}$, which is the optimal learning rate scheme for nonsmooth convex-concave problems [6, 1]. For the smooth case, one might expect the sequence $\{\delta_t^m\}_t$ to be square-summable and thus $\eta_t^m \to \eta_\infty^m > 0$, in which case the learning rate does not vanish. Additionally, the adaptive learning rate for each worker is locally updated to exploit the problem structure available in worker's local dataset. This makes our local adaptive learning rate scheme distinct compared to existing distributed adaptive algorithms for minimization problems [53, 47, 12]. Very recently, [7] used local Adam for training conditional GANs efficiently, but they provide theoretical guarantees only for the local extragradient without adaptivity.

Adaptivity to (G, L, σ). Our algorithm does not require knowledge of problem parameters such as the size of the gradients G, the smoothness L, or the variance of gradient estimates σ. Instead, we only need an initial guess of G, denoted G_0, and the diameter of the feasible set, D. Define

$$\gamma := \max\{G/G_0, G_0/G\} \geqslant 1. \tag{5}$$

This quantity measures how good our guess is and appears in the convergence guarantees for the algorithm. Our algorithm still requires knowledge of the problem class, as we need to use different base learning rate, α, for smooth and nonsmooth problems; see Theorems 5.1 and 5.2, respectively.

5 Convergence Results

We state two theorems characterizing the convergence rate of LocalAdaSEG for the smooth and nonsmooth problems. We use the notation $\tilde{O}$ to hide absolute constants and logarithmic factors of $T = KR$ and problem parameters. The proofs are given in §B.1 and §B.2 of the appendix. Recall the definition of γ in (5). [6].

Theorem 5.1 (Nonsmooth Case)**.** *Assume A.1, A.2, and A.3 hold. Let* $\bar{z} = \text{LocalAdaSEG}(G_0, D; K, M, R; 1)$. *Then*

$$\mathbb{E}[\text{DualGap}(\bar{z})] = \tilde{O}\left(\frac{\gamma GD}{\sqrt{T}} + \frac{\sigma D}{\sqrt{MT}}\right).$$

Theorem 5.2 (Smooth Case)**.** *Assume A.1, A.2, A.3, and A.4 hold. Let* $\bar{z} = \text{LocalAdaSEG}(G_0, D; K, M, R; 1/\sqrt{M})$. *Define the cumulative norms of stochastic gradients occurred on worker m:*

$$\mathcal{V}_m(T) := \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|g_t^m\|_*^2 + \|M_t^m\|_*^2}\right]. \tag{6}$$

Then

$$\mathbb{E}[\text{DualGap}(\bar{z})] = \tilde{O}\left(\frac{\sigma D}{\sqrt{MT}} + \frac{DM^{3/2} \cdot \mathcal{V}_1(T)}{T} + \frac{\gamma^2 LD^2 + \gamma GDM^{3/2}}{T}\right). \tag{7}$$

Remark 1 (The cumulative stochastic gradient growth $\mathcal{V}_1(T)$.)**.** *Although a trivial bound on $\mathcal{V}_1(T)$ is $\mathcal{V}_1(T) \leqslant G\sqrt{2T}$, typically we have $\mathcal{V}_1(T) \ll \sqrt{T}$ in practice [23, 48, 17, 13, 38], especially in the sparse data scenarios. For example, consider the bilinear saddle-point problem $\min_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}} x^\top \big(\sum_{i=1}^n p_i M_i\big) y$, where a larger weight $p_i > 0$ means the matrix M_i appears more frequently in the dataset. When most of matrices with large weights are row-sparse and column-sparse, the quantity $\mathcal{V}_1(T)$ is much smaller than $G\sqrt{2T}$. Theorem B.3, in the appendix, shows that with a different choice of the base learning rate α one can obtain a near linear speed-up result, which removes the dependence on $\mathcal{V}_1(T)$: for large T,*

$$\mathbb{E}[\text{DualGap}(\bar{z})] = \tilde{O}\left(\frac{\sigma D}{\sqrt{MT^{1-2\epsilon}}} + \frac{\gamma^2 LD^2}{T^{1-2\epsilon}} + \frac{LD^2 M}{T} + \frac{\gamma GDM^{3/2}}{T^{1+\epsilon}}\right), \text{ for any } \epsilon \in (0, \tfrac{1}{2}).$$

Following the discussion in [17, 38], when the cumulative growth of the stochastic gradient is slow, i.e., $\mathcal{V}_1(T) = O(T^b)$ for some $0 < b < \frac{1}{2}$, then the second term in (7) is $O(DM^{3/2}/T^{1-b})$ and linear speed-up is achieved, since as $T \to \infty$, the dominating term become $O(\sigma D/\sqrt{MT})$.

Remark 2 (Unbounded stochastic gradient oracle)**.** *Our analysis can be extended to unbounded homogeneous and light-tailed oracles using the following argument. Let $\|G\|_\infty := \sup_{z\in\mathcal{Z}} \|G(z)\|_* < \infty$, which upper bounds the expectation of the SG oracle. Assume $\|\tilde{G}(z) - G(z)\|_*/\|G\|_\infty$ is independent of z and follows the distribution of the absolute value of a standard normal. Define the set $\mathcal{Z}' := \{z_t^m, \tilde{z}_{t-1}^{m,*}\}_{t,m}$ of all iterates. For any $0 < \delta < 1$, define the event*

$$\mathcal{E} := \left\{ \max_{z'\in\mathcal{Z}'} \|\tilde{G}(z') - G(z')\|_* \leqslant \|G\|_\infty \cdot \left(\sqrt{2\log(4MT)} + \sqrt{2\log(2/\delta)}\right) := G_{prob} \right\}.$$

Then $\mathbb{P}(\mathcal{E}) \geqslant 1 - \delta$; see Appendix A.1. We can repeat the proof of Theorem 5.1 and Theorem 5.2 on the event $\mathcal{E}$ and interpret our results with G replaced by G_{prob}, which effectively substitutes G with $\|G\|_\infty$ at the cost of an extra $\log(T)$ factor.

Remark 3 (Minibatch EG as a baseline)**.** *We comment on a performance of an obvious baseline that implements minibatch stochastic EG using M workers. Suppose the algorithm takes R extragradient steps, with each step using a minibatch of size KM, resulting in a procedure that communicates exactly R times. The performance of such a minibatch EG for general nonsmooth and smooth minimax problems [6, 24] is, respectively,* [2]

$$O\left(\frac{\sigma D}{\sqrt{KMR}} + \frac{\|G\|_\infty D}{\sqrt{R}}\right) \quad \text{and} \quad O\left(\frac{\sigma D}{\sqrt{KMR}} + \frac{LD^2}{R}\right). \tag{8}$$

Under the same computation and communication structure, our algorithm enjoys adaptivity, achieves the same linear speed-up in the variance term $\frac{\sigma D}{\sqrt{KMR}}$, and improves dependence on the gradient upper bound $\|G\|_\infty$ and the smoothness parameter L, which is a desirable property for problems where these parameters are large.

Remark 4 (Single-worker mode)**.** *Another natural baseline is to run EG on a single worker for T iterations with batchsize equal to one. The convergence rates for this procedure in nonsmooth and smooth cases are $O(\sigma D/\sqrt{T} + \|G\|_\infty D/\sqrt{T})$ and $O(\sigma D/\sqrt{T} + LD^2/T)$, respectively. In the smooth case, the single-worker mode is inferior to minibatch EG, since the dominant term for the former is $1/\sqrt{T}$, but it is $1/\sqrt{MT}$ for the latter. On the other hand, in the nonsmooth case, minibatch EG reduces the variance term, but the term involving the deterministic part degrades. Therefore, in the nonsmooth case, we can only claim that the minibatch EG is better than the single-worker mode in the noise-dominant regime $\sigma = \Omega(\|G\|_\infty\sqrt{M})$.*

6 Experiments

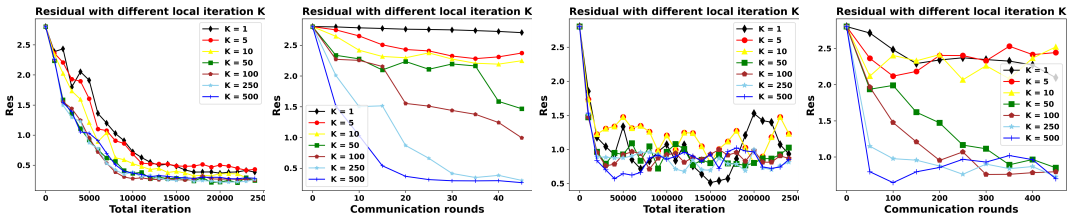
We apply LocalAdaSEG to the stochastic bilinear minimax problem introduced in [25, 7] and to train Wasserstein generative adversarial neural network (Wasserstein GAN) [3]. For the homogeneous setting, to demonstrate the efficiency of our proposed algorithm, we compare LocalAdaSEG with mini-batch stochastic extragradient gradient descent (MB-SEGDA) [44], mini-batch universal mirror-prox (MB-UMP) [6], mini-batch adaptive single-gradient mirror-Prox (MB-ASMP) [24], extra step local SGD (LocalSEGDA) [7], and local stochastic gradient descent ascent (LocalSGDA) [22]. We further extend the proposed LocalAdaSEG algorithm to solve federated WGANs with a heterogeneous dataset to verify its efficiency. In this setting, we additionally compare LocalAdaSEG with Local Adam [7]. We emphasize here that whether Local Adam converges is still an open question, even for the stochastic convex-concave setting.

6.1 Stochastic bilinear minimax problem

We consider the stochastic bilinear minimax problem with box constraints

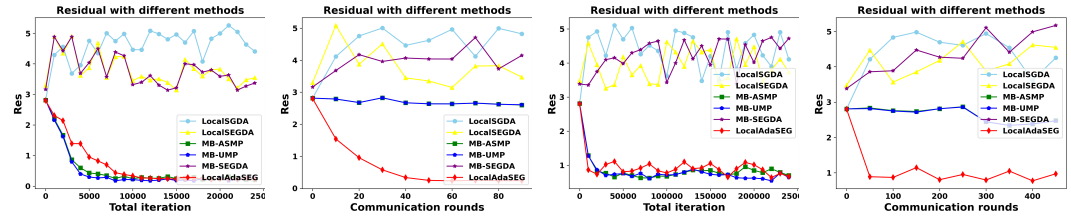
$$\min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^n} F(x,y) = \mathbb{E}_{\xi\sim P}\big[x^\top A y + (b+\xi)^\top x + (c+\xi)^\top y\big], \tag{9}$$

[2] These bounds hold due to Theorem 4 of [24], whose rates for nonsmooth and smooth problems are of the form $O(R(G+\sigma)/\sqrt{T})$ and $O(\beta R^2/T + R\sigma/\sqrt{T})$, respectively. Eq. (8) follows with σ in the original theorem statement replaced by $\sigma/\sqrt{KM}$, β by L, R by D, G by $\|G\|_\infty$, and T by R.



(a) (b) (c) (d)

Figure 1: Subfigures (a)-(b) and (c)-(d) plot the residual of LocalAdaSEG against the total number of iterations T and communications R, with varying numbers of local iterations K. We also investigate the effect of noise level ($\sigma = 0.1$ in (a)(b) and $\sigma = 0.5$ in (c)(d)).



(a) (b) (c) (d)

Figure 2: Subfigures (a)-(b) and (c)-(d) compare LocalAdaSEG with existing optimizers. We plot the residuals against the total number of iterations T and communications R with different noise levels ($\sigma = 0.1$ in (a)(b) and $\sigma = 0.5$ in (c)(d)).

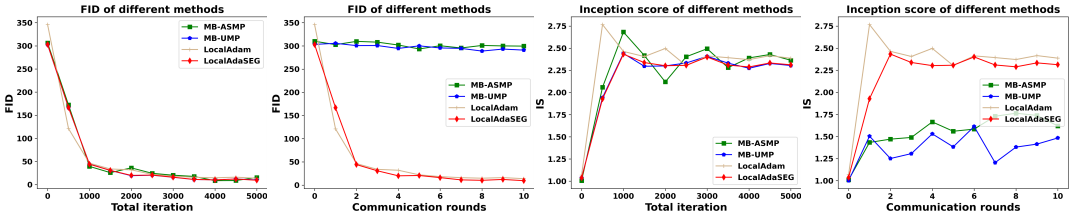
where $x, y \in [-1, 1]^n \in \mathbb{R}^n$, (A, b, c) are data, random variable $\xi \sim N(0, \sigma^2 \mathcal{I})$, and σ is the variance. We define the KKT residual $\mathrm{Res}(x, y)$ as:

$$\mathrm{Res}(x, y) = \sqrt{\left\|x - \Pi_{[-1,1]}\big(x - (Ay + b)\big)\right\|^2 + \left\|y - \Pi_{[-1,1]}\big(y + (Ax + c)\big)\right\|^2}. \tag{10}$$

It is not hard to verify that given $(x^*, y^*) \in \mathbb{R}^n \times \mathbb{R}^n$, $\mathrm{Res}(x^*, y^*) = 0$ if and only if (x^*, y^*) belongs to the saddle-points of the bilinear minimax problem (9). During the experiments, we use $\mathrm{Res}(x, y)$ to measure the quality of approximated solution obtained by different optimizers.

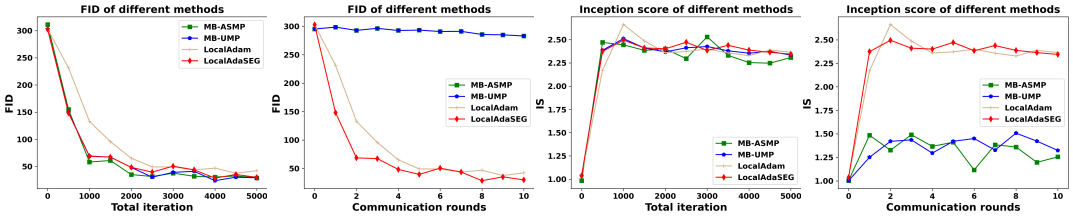
Dataset Generation. We uniformly generate b and c in $[-1, 1]^n$ with $n = 10$. The symmetric matrix A is constructed as $A = \bar{A}/\max\big(|b|_{\max}, |c|_{\max}\big)$, where $\bar{A} \in [-1, 1]^{n \times n}$ is a random symmetric matrix. We emphasize that A is merely symmetric, but not semi-definite. To simulate the distributed environment, we distribute (A, b, c) to M workers, where $M = 4$. Each worker solves the above bilinear problem with an optimization algorithm locally. We instantiate LocalAdaSEG with different numbers of local iterations $K \in \{1, 5, 10, 50, 100, 250, 500\}$, and different noise levels $\sigma \in \{0.1, 0.5\}$, shown in Figure 1. A larger σ indicates more noise in the stochastic gradients, making problem (9) harder. In addition, we further compare LocalAdaSEG by setting local iteration $K = 50$ against several existing optimizers, illustrated in Figure 2.

Experimental results. In Figure 1, LocalAdaSEG provides stable convergence results under different configurations of local iterations K and noise levels σ. Figure (b)(d) illustrate that a suitably large K could accelerate the convergence speed of LocalAdaSEG. Figure (a)(c) illustrate that a large variance would result in unstable optimization trajectories. The experiment findings agree with our theoretical predictions: (i) a larger $T = KR$ improves the convergence; (ii) the variance term dominates the convergence rate of LocalAdaSEG, large variance term will slowdown LocalAdaSEG. In Figure 2, (a)(c) illustrate that adaptive variants of stochastic minimax optimizers, i.e., LocalAdaSEG, MB-UMP and MB-ASMP, achieve better performance compared to standard ones such as LocalSGDA, LocalSEGDA and MB-SEGDA, whose learning rates are hard to tune for minimax problems. In addition, when compared in terms of communication rounds in (b)(d), LocalAdaSEG converges faster than other distributed stochastic minimax optimizers, demonstrating the superiority of LocalAdaSEG.



(a) (b) (c) (d)

Figure 3: Subfigures (a)-(b) and (c)-(d) show the results of WGAN trained with LocalAdaSEG and existing optimizers. We plot FID and IS against the number of iterations and communications, respectively.



(a) (b) (c) (d)

Figure 4: Subfigures (a)-(b) and (c)-(d) show the results of Federated WGAN trained with LocalAdaSEG and existing optimizers. We plot FID and IS against the number of iterations and communications, respectively.

6.2 Wasserstein GAN

We train Wasserstein GAN (WGAN) to validate the efficiency of LocalAdaSEG on a real-world application task. This is a challenging minimax problem as the objectives of both generator and discriminator are non-convex and non-concave. Problem description is placed in Appendix D.1.

Implementation details. Experiments are conducted on the MNIST datasets of digits from 0 to 9, with 60000 training images of size 28×28. We adopt the same network architecture of WGAN as that of DCGAN [3]. We simulate $M = 4$ parallel workers and run LocalAdaSEG with the batch size 128 and local iteration steps $K = 500$. In the homogeneous setting, the local data in each worker is uniformly sampled from the entire dataset. In the heterogeneous setting, we partition the MNIST dataset into 4 subsets using the partition methods in [37]. Then each worker is loaded with a fraction of the dataset. Due to non-adaptive learning rates, LocalSGDA, LocalSEGDA, and MB-SEGDA are hard to tune and do not achieve a satisfactory performance for training WGAN. For a better illustration, we only show the performance of LocalAdaSEG, MB-UMP, MB-ASMP and Local Adam. To measure the efficacy of the compared optimizers, we plot FID and IS [29] against the number iterations and communications, respectively.

Experimental results. Figures 3 and 4 compare MB-UMP, MB-ASMP, LocalAdam and LocalAdaSEG in the homogeneous and heterogeneous setting, respectively. In Figure 3(a) and Figure 4(a), MB-UMP, MB-ASMP, LocalAdam and LocalAdaSEG quickly converge to a solution with a low FID value. However, when compared in terms of communication rounds in Figure 3(b) and Figure 4(b), LocalAdaSEG and Local Adam converge faster than other optimizers and reach a satisfactory solution within just a few rounds. In Figure 3(c) and Figure 4(c), all the listed optimizers achieve a high IS. Notably, the IS of LocalAdaSEG and Local Adam increase much faster with less communication than MB-UMP, MB-ASMP as shown in Figure 3(d) and Figure 4(d).

7 Conclusion, Discussion, and Broader Impact

We proposed an adaptive communication-efficient distributed stochastic extragradient algorithm in the Parameter-Server model for stochastic convex-concave minimax problem, LocalAdaSEG. We theoretically showed LocalAdaSEG that achieves the optimal convergence rate with a linear speed-up property for both nonsmooth and smooth objectives. Experiments verify our theoretical results and demonstrate the efficiency of LocalAdaSEG.

One main limitation of this work is that the current analysis merely holds for the homogeneous setting. A future direction is to extend the theoretical result of LocalAdaSEG to the heterogeneous setting

that better models various real-world applications, such as federated GANs [7] and robust federated learning [21]. In addition, extending theoretical results from the stochastic convex-concave setting to the stochastic nonconvex-(non)concave setting is an interesting and challenging research direction.

Training deep learning models with a large amount of parameters and data requires a lot of GPUs and energy. Our algorithm can be applied to train large-scale (federated) GANs quickly and with lower communication costs. At the same time adaptive tuning of the learning rate further saves costs compared to trial-and-error approach by saving computational resource and energy.

References

[1] K. Antonakopoulos, V. Belmega, and P. Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. In *International Conference on Learning Representations*, 2021.

[2] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017.

[4] W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International Conference on Artificial Intelligence and Statistics*, pages 2863–2873. PMLR, 2020.

[5] R. Babanezhad and S. Lacoste-Julien. Geometry-aware universal mirror-prox. *arXiv preprint arXiv:2011.11203*, 2020.

[6] F. Bach and K. Y. Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on Learning Theory*, pages 164–194. PMLR, 2019.

[7] A. Beznosikov, V. Samokhin, and A. Gasnikov. Distributed saddle-point problems: Lower bounds, optimal algorithms and federated gans. *arXiv preprint arXiv:2010.13112*, 2021.

[8] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[9] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, Dec. 2010.

[10] S. Chatterjee. *Superconcentration and Related Topics*. Springer International Publishing, 2014.

[11] C. Chen, L. Shen, H. Huang, Q. Wu, and W. Liu. Quantized adam with error feedback. *arXiv preprint arXiv:2004.14180*, 2020.

[12] T. Chen, Z. Guo, Y. Sun, and W. Yin. Cada: Communication-adaptive distributed adam. In *International Conference on Artificial Intelligence and Statistics*, pages 613–621. PMLR, 2021.

[13] X. Chen, S. Liu, R. Sun, and M. Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019.

[14] X. Chen, S. Yang, L. Shen, and X. Pang. A distributed training algorithm of generative adversarial networks with quantized gradients. *arXiv preprint arXiv:2010.13359*, 2020.

[15] Y. Chen, G. Lan, and Y. Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, Jan. 2014.

[16] Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, June 2017.

[17] Z. Chen, Z. Yuan, J. Yi, B. Zhou, E. Chen, and T. Yang. Universal stagewise learning for non-convex problems with convergence on averaged solutions. In *International Conference on Learning Representations*, 2019.

[18] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018.

[19] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[21] Y. Deng, M. M. Kamani, and M. Mahdavi. Distributionally robust federated averaging. In *Advances in Neural Information Processing Systems*, volume 33, pages 15111–15122. Curran Associates, Inc., 2020.

[22] Y. Deng and M. Mahdavi. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1387–1395. PMLR, 13–15 Apr 2021.

[23] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[24] A. Ene and H. L. Nguyen. Adaptive and universal single-gradient algorithms for variational inequalities. *arXiv preprint arXiv:2010.07799*, 2020.

[25] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[26] G. Gidel, R. A. Hemmat, M. Pezeshki, R. Le Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1802–1811. PMLR, 2019.

[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

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

[29] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[30] C. Hou, K. K. Thekumparampil, G. Fanti, and S. Oh. Efficient algorithms for federated saddle point optimization, 2021.

[31] A. Juditsky, A. Nemirovski, et al. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, 30(9):149–183, 2011.

[32] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, June 2011.

[33] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2017.

[34] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 1976.

[35] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020.

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

[37] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi. Don't use large mini-batches, use local sgd. In *International Conference on Learning Representations*, 2020.

[38] M. Liu, Y. Mroueh, J. Ross, W. Zhang, X. Cui, P. Das, and T. Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *International Conference on Learning Representations*, 2020.

[39] M. Liu, W. Zhang, Y. Mroueh, X. Cui, J. Ross, T. Yang, and P. Das. A decentralized parallel algorithm for training generative adversarial nets. volume 33, 2020.

[40] H. B. McMahan et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1), 2021.

[41] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019.

[42] P. Mertikopoulos, C. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018.

[43] R. D. C. Monteiro and B. F. Svaiter. Complexity of variants of tseng's modified f-b splitting and korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM J. Optimization*, 21:1688–1720, 2011.

[44] A. Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[45] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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

[47] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

[48] S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[49] A. Rogozin, A. Beznosikov, D. Dvinskikh, D. Kovalev, P. Dvurechensky, and A. Gasnikov. Decentralized distributed optimization for saddle point problems, 2021.

[50] A. Smola and S. Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710, 2010.

[51] S. U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

[52] J. Wang, T. Zhang, S. Liu, P.-Y. Chen, J. Xu, M. Fardad, and B. Li. Towards a unified min-max framework for adversarial exploration and robustness. *arXiv preprint arXiv:1906.03563*, 2019.

[53] C. Xie, O. Koyejo, I. Gupta, and H. Lin. Local adaalter: Communication-efficient stochastic gradient descent with adaptive learning rates. *arXiv preprint arXiv:1911.09030*, 2019.

[54] H. Yu, R. Jin, and S. Yang. On the linear speed-up analysis of communication efficient momentum SGD for distributed non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7184–7193. PMLR, 09–15 Jun 2019.

[55] J. Zhang, P. Xiao, R. Sun, and Z. Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 7377–7389. Curran Associates, Inc., 2020.

[56] R. Zhao. Accelerated stochastic algorithms for convex-concave saddle-point problems. *arXiv preprint arXiv:1903.01687*, 2021.

A Appendix to Main Text

A.1 Extension to Unbounded Stochastic Gradient Oracle

Let $\{Z_i\}_{i=1}^n$ be a sequence of i.i.d. standard normals. We have the following well-known results (see Appendix A of [10])

$$\mathbb{P}\big(\max_i Z_i > \mathbb{E}[\max_i Z_i] + t\big) \leqslant \exp(-t^2/2) \text{ for all } t > 0,$$
$$\mathbb{E}[\max_i |Z_i|] \leqslant \sqrt{2\log(2n)}.$$

It can be shown $\mathbb{P}(\max_i |Z_i| \geqslant \sqrt{2\log(2n)} + t) \leqslant 2\exp(-t^2/2)$. We apply this result with the sequence $\big\{\|G(z_t^m) - \tilde{G}(z_t^m)\|_*/\|G\|_\infty, \|G(\tilde{z}_{t-1}^{m,*}) - \tilde{G}(\tilde{z}_{t-1}^{m,*})\|_*/\|G\|_\infty\big\}_{m,t}$, which is a sequence of $2MT$ i.i.d. standard normals by the homogeneity of the oracle.

B Proof of Theorems

Lemma B.1. *For all $m \in [M]$, consider the sequence $\{\eta_t^m, \tilde{z}_{t-1}^{m,*}, z_t^m, \tilde{z}_t^m\}_{t=1}^T$ defined in Algorithm 1. It holds*

$$\|\tilde{z}_{t-1}^{m,*} - z_t^m\|/\eta_t^m \leqslant G, \quad \|\tilde{z}_t^m - z_t^m\|/\eta_t^m \leqslant G.$$

Proof of Lemma B.1. Let $I : \mathcal{Z} \to \mathcal{Z}^*$ be the identity map which maps an element $z \in \mathcal{Z}$ to the corresponding element in the dual space $\mathcal{Z}^*$. The first-order optimality condition of the update rule $z_t^m = \Pi_{\mathcal{Z}}[\tilde{z}_{t-1}^{m,*} - \eta_t^m M_t^m]$ is

$$\langle \eta_t^m M_t^m + I(z_t^m - \tilde{z}_{t-1}^{m,*}), z - z_t^m\rangle \geqslant 0, \forall z \in \mathcal{Z}.$$

Set $z = \tilde{z}_{t-1}^{m,*}$, apply the Cauchy-Schwartz inequality and we obtain

$$\begin{aligned}\eta_t^m \|M_t^m\|_* \cdot \|\tilde{z}_{t-1}^{m,*} - z_t^m\| &\geqslant \langle \eta_t^m M_t^m, \tilde{z}_{t-1}^{m,*} - z_t^m\rangle \\ &\geqslant \langle I(\tilde{z}_{t-1}^{m,*} - z_t^m), \tilde{z}_{t-1}^{m,*} - z_t^m\rangle = \|\tilde{z}_{t-1}^{m,*} - z_t^m\|^2.\end{aligned}$$

The second inequality holds due to similar reasoning. We conclude the proof of Lemma B.1. □

Lemma B.2 (One-step analysis)**.** *For all $m \in [M]$, consider the sequence $\{\eta_t^m, \tilde{z}_{t-1}^{m,*}, M_t^m = \tilde{G}(\tilde{z}_{t-1}^{m,*}), z_t^m, g_t^m = \tilde{G}(\tilde{z}_t^m), \tilde{z}_t^m\}_{t=1}^T$ defined in Algorithm 1. It holds for all $z \in \mathcal{Z}$,*

$$\begin{aligned}\langle z_t^m - z, g_t^m\rangle \leqslant \frac{1}{\eta_t^m}\Big(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{m,*}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2\Big) - \frac{1}{\eta_t^m}\Big(\tfrac{1}{2}\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + \tfrac{1}{2}\|z_t^m - \tilde{z}_t^m\|^2\Big) \\ + \|g_t^m - M_t^m\|_* \cdot \|z_t^m - \tilde{z}_t^m\|.\end{aligned}$$

Proof of Lemma B.2. For any $c, g \in \mathcal{Z}$, consider the update of the form $a^* = \Pi_{\mathcal{Z}}[c - g] = \operatorname{argmin}_{z\in\mathcal{Z}} \langle g, z\rangle + \frac{1}{2}\|z - c\|^2$. It holds for all $b \in \mathcal{Z}$,

$$\langle g, a^* - b\rangle \leqslant \tfrac{1}{2}\|b - c\|^2 - \tfrac{1}{2}\|b - a^*\|^2 - \tfrac{1}{2}\|a^* - c\|^2.$$

By the update rule of z_t^m and $\tilde{z}_t^m$, we have (taking $a^* \leftrightarrow z_t^m$, $b \leftrightarrow \tilde{z}_t^m$, $g \leftrightarrow \eta_t^m M_t^m$, $c \leftrightarrow \tilde{z}_{t-1}^{m,*}$)

$$\langle \eta_t^m M_t^m, z_t^m - \tilde{z}_t^m\rangle \leqslant \tfrac{1}{2}\|\tilde{z}_t^m - \tilde{z}_{t-1}^{m,*}\|^2 - \tfrac{1}{2}\|\tilde{z}_t^m - z_t^m\|^2 - \tfrac{1}{2}\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2, \tag{11}$$

and for all $z \in \mathcal{Z}$ (taking $a^* \leftrightarrow \tilde{z}_t^m$, $b \leftrightarrow z$, $g \leftrightarrow \eta_t^m g_t^m$, $c \leftrightarrow \tilde{z}_{t-1}^{m,*}$)

$$\langle \eta_t^m g_t^m, \tilde{z}_t^m - z\rangle \leqslant \tfrac{1}{2}\|z - \tilde{z}_{t-1}^{m,*}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2 - \tfrac{1}{2}\|\tilde{z}_t^m - \tilde{z}_{t-1}^{m,*}\|^2. \tag{12}$$

Finally we apply the Cauchy-Schwarz inequality and plug in Eqs. (11) and (12).

$$\begin{aligned}
\langle g_t^m, z_t^m - z\rangle &= \langle g_t^m, z_t^m - \tilde{z}_t^m\rangle + \langle g_t^m, \tilde{z}_t^m - z\rangle \\
&= \langle g_t^m - M_t^m, z_t^m - \tilde{z}_t^m\rangle + \langle g_t^m, \tilde{z}_t^m - z\rangle + \langle M_t^m, z_t^m - \tilde{z}_t^m\rangle \\
&\leqslant \|g_t^m - M_t^m\|_* \cdot \|z_t^m - \tilde{z}_t^m\| + \langle g_t^m, \tilde{z}_t^m - z\rangle + \langle M_t^m, z_t^m - \tilde{z}_t^m\rangle \\
&\leqslant \|g_t^m - M_t^m\|_* \cdot \|z_t^m - \tilde{z}_t^m\| \\
&\quad + \frac{1}{\eta_t^m}\left(\tfrac{1}{2}\|\tilde{z}_t^m - \tilde{z}_{t-1}^{m,*}\|^2 - \tfrac{1}{2}\|\tilde{z}_t^m - z_t^m\|^2 - \tfrac{1}{2}\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2\right) \\
&\quad + \frac{1}{\eta_t^m}\left(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{m,*}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2 - \tfrac{1}{2}\|\tilde{z}_t^m - \tilde{z}_{t-1}^{m,*}\|^2\right) \\
&= \frac{1}{\eta_t^m}\left(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{m,*}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2\right) - \frac{1}{\eta_t^m}\left(\tfrac{1}{2}\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + \tfrac{1}{2}\|z_t^m - \tilde{z}_t^m\|^2\right) \\
&\quad + \|g_t^m - M_t^m\|_* \cdot \|z_t^m - \tilde{z}_t^m\|.
\end{aligned}$$

This finishes the proof of Lemma B.2. □

B.1 Proof of Theorem 5.1

Proof of Theorem 5.1, Non-smooth Case. The proof strategy follows closely that of Bach and Levy [6]. **Step 1.** We apply the Lemma B.2 and sum over all $m \in [M]$ and $t \in [T]$. Define

$$\xi_t^m := G(z_t^m) - g_t^m = G(z_t^m) - \tilde{G}(z_t^m).$$

For all $z \in \mathcal{Z}$,

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\langle z_t^m - z, G(z_t^m)\rangle = \sum_{t=1}^{T}\sum_{m=1}^{M}\langle z_t^m - z, \xi_t^m\rangle + \sum_{t=1}^{T}\sum_{m=1}^{M}\langle z_t^m - z, g_t^m\rangle \tag{13}$$

$$\leqslant \underbrace{\sum_{t=1}^{T}\sum_{m=1}^{M}\langle z_t^m - z, \xi_t^m\rangle}_{I(z)} \tag{14}$$

$$+ \underbrace{\sum_{t=1}^{T}\sum_{m=1}^{M}\frac{1}{\eta_t^m}\left(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{m,*}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2\right)}_{II(z)} \tag{15}$$

$$\underbrace{-\sum_{t=1}^{T}\sum_{m=1}^{M}\frac{1}{\eta_t^m}\left(\tfrac{1}{2}\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + \tfrac{1}{2}\|z_t^m - \tilde{z}_t^m\|^2\right)}_{III} \tag{16}$$

$$+ \underbrace{\sum_{t=1}^{T}\sum_{m=1}^{M}\|g_t^m - M_t^m\|_* \cdot \|z_t^m - \tilde{z}_t^m\|}_{IV}. \tag{17}$$

Now we use Lemma C.3 and obtain

$$TM \cdot \mathbb{E}[\mathrm{DualGap}(\bar{z})] \leqslant \mathbb{E}[\sup_{z\in\mathcal{Z}}\{I(z) + II(z) + III + IV\}] \tag{18}$$

$$\leqslant \mathbb{E}[\sup_{z\in\mathcal{Z}} I(z)] + \mathbb{E}[\sup_{z\in\mathcal{Z}} II(z)] + \mathbb{E}[III] + \mathbb{E}[IV] \tag{19}$$

Next we upper bound each term in turns. Steps 2–5 rely heavily on the learning rate scheme. Define

$$(Z_t^m)^2 := \frac{\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + \|z_t^m - \tilde{z}_t^m\|^2}{5(\eta_t^m)^2}$$

for all $t \in [T]$ and $m \in [M]$. By Lemma B.1 we know $Z_t^m \leqslant G$ almost surely. This is due to

$$\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + \|z_t^m - \tilde{z}_t^m\|^2 \leqslant \|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + 2\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + 2\|\tilde{z}_{t-1}^{m,*} - \tilde{z}_t^m\|^2 \leqslant 5G^2(\eta_t^m)^2.$$

Moreover, for the nonsmooth case ($\alpha = 1$), η_t^m can be expressed by

$$\eta_t^m = \frac{D}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1}(Z_\tau^m)^2}}. \tag{20}$$

Step 2. Show $\mathbb{E}[\sup_{z\in\mathcal{Z}} I(z)] = O(\sigma D\sqrt{MT})$. For all $z \in \mathcal{Z}$,

$$I(z) = \sum_{t=1}^{T}\sum_{m=1}^{M}\langle z_t^m - \tilde{z}_0^m, \xi_t^m\rangle + \sum_{t=1}^{T}\sum_{m=1}^{M}\langle \tilde{z}_0^m - z, \xi_t^m\rangle.$$

The first term is a martingale difference sequence (MDS) and is zero in expectation. For the second term we use the Cauchy-Schwarz inequality. For all $z \in \mathcal{Z}$,

$$\begin{aligned}
\mathbb{E}\Big[\sum_{t=1}^{T}\sum_{m=1}^{M}\langle \tilde{z}_0^m - z, \xi_t^m\rangle\Big] &= \mathbb{E}\Big[\Big\langle \tilde{z}_0 - z, \sum_{t=1}^{T}\sum_{m=1}^{M}\xi_t^m\Big\rangle\Big] \\
&\leqslant \sqrt{\mathbb{E}\Big[\|\tilde{z}_0 - z\|^2\Big]}\cdot\sqrt{\mathbb{E}\Big[\Big\|\sum_{t=1}^{T}\sum_{m=1}^{M}\xi_t^m\Big\|_*^2\Big]} \\
&= \sqrt{\mathbb{E}\Big[\|\tilde{z}_0 - z\|^2\Big]}\cdot\sqrt{\mathbb{E}\Big[\sum_{t=1}^{T}\sum_{m=1}^{M}\|\xi_t^m\|_*^2\Big]} \leqslant \sigma D\sqrt{MT}.
\end{aligned}$$

In the last equality we use the fact that $\{\xi_t^m\}$ is a MDS. This establishes $\mathbb{E}[\sup_z I(z)] \leqslant \sigma D\sqrt{MT}$.

Step 3. Show $\mathbb{E}[\sup_{z\in\mathcal{Z}} II(z)] = O(DG\cdot M\sqrt{T})$. For all $z \in \mathcal{Z}$,

$$\begin{aligned}
II(z) &= \sum_{t=1}^{T}\sum_{m=1}^{M}\frac{1}{\eta_t^m}\Big(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{m,*}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2\Big) \\
&= \sum_{m=1}^{M}\sum_{t\notin S+1}\frac{1}{\eta_t^m}\Big(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{m,*}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2\Big) + \sum_{m=1}^{M}\sum_{t\in S+1}\frac{1}{\eta_t^m}\Big(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{m,*}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2\Big) \\
&= \sum_{m=1}^{M}\sum_{t\notin S+1}\frac{1}{\eta_t^m}\Big(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{m}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2\Big) + \sum_{m=1}^{M}\sum_{t\in S+1}\frac{1}{\eta_t^m}\Big(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{\circ}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2\Big) \\
&= \underbrace{\sum_{m=1}^{M}\sum_{t=1}^{T}\frac{1}{\eta_t^m}\Big(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{m}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2\Big)}_{A} + \underbrace{\sum_{m=1}^{M}\sum_{t\in S+1}\frac{1}{\eta_t^m}\Big(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{\circ}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_{t-1}^m\|^2\Big)}_{B}
\end{aligned}$$

where we used the definition of $\tilde{z}_{t-1}^{m,*}$ for two cases $t \in S+1$ and $t \notin S+1$ (Line 1 and 1 in algorithm).

We upper bound A and show $B \leqslant 0$.

Recall for $t \in S+1$, we have $\tilde{z}_{t-1}^{m,*} = \tilde{z}_{t-1}^{\circ} = \sum_{m=1}^{M} w_m \cdot \tilde{z}_{t-1}^m$, and for $t \notin S+1$, we have $\tilde{z}_{t-1}^{m,*} = \tilde{z}_{t-1}^m$. For the first term A we use $\frac{1}{2}\|z - \tilde{z}_t^m\|^2 \leqslant D^2$ and then telescope.

$$\begin{aligned}
A &= \sum_{m=1}^{M}\Big\{\frac{1}{\eta_1^m}\Big(\tfrac{1}{2}\|\tilde{z}_0^m - z\|^2\Big) - \frac{1}{\eta_T^m}\Big(\tfrac{1}{2}\|\tilde{z}_T^m - z\|^2\Big) + \sum_{t=2}^{T}\Big(\frac{1}{\eta_t^m} - \frac{1}{\eta_{t-1}^m}\Big)\Big(\tfrac{1}{2}\|\tilde{z}_{t-1}^m - z\|^2\Big)\Big\} \\
&\leqslant \sum_{m=1}^{M}\Big\{\frac{D^2}{\eta_1^m} + \sum_{t=2}^{T}\Big(\frac{1}{\eta_t^m} - \frac{1}{\eta_{t-1}^m}\Big)D^2\Big\} \\
&\leqslant \sum_{m=1}^{M}\Big\{\frac{D^2}{\eta_1^m} + \frac{D^2}{\eta_T^m}\Big\}
\end{aligned}$$

For each m, we have $D^2/\eta_1^m = DG_0$. For D^2/η_T^m we use the learning rate scheme. Recall the definition of Z_t^m. Then

$$\frac{D^2}{\eta_T^m} = D\sqrt{G_0^2 + \sum_{t=1}^{T-1}(Z_t^m)^2} \leqslant D\sqrt{G_0 + G^2T} \leqslant DG_0 + DG\sqrt{T}.$$

This implies $A \leqslant M(2DG_0 + DG\sqrt{T}) = O(DG \cdot M\sqrt{T})$.

For the term B, we use the definition of $\tilde{z}_{t-1}^\circ$ and the weights $\{w_m\}$ to show $B \leqslant 0$. For each t, since $\tilde{z}_{t-1}^\circ$ the same for all workers,

$$\begin{aligned}
\sum_{m=1}^M \frac{1}{\eta_t^m}\Big(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^\circ\|^2\Big) &= \Big(\sum_{m=1}^M \frac{1}{\eta_t^m}\Big)\Big(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^\circ\|^2\Big) \\
&= \Big(\sum_{m=1}^M \frac{1}{\eta_t^m}\Big)\Big(\tfrac{1}{2}\big\|\textstyle\sum_{m=1}^M {w_m}^{1/2} \cdot {w_m}^{1/2}(z - \tilde{z}_{t-1}^m)\big\|^2\Big) \\
&\leqslant \Big(\sum_{m=1}^M \frac{1}{\eta_t^m}\Big)\Big(\sum_{m=1}^M w_m\Big)\Big(\sum_{m=1}^M w_m \cdot \tfrac{1}{2}\|z - \tilde{z}_{t-1}^m\|^2\Big) \\
&= \sum_{m=1}^M \frac{1}{\eta_t^m}\Big(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^m\|^2\Big).
\end{aligned}$$

In the last equality we use $\sum_{m=1}^M w_m = 1$ and $(\sum_{m=1}^M 1/\eta_t^m)w_{m'} = 1/\eta_t^{m'}$ for all $m' \in [M]$. This implies $B \leqslant 0$. This establishes $\mathbb{E}[\sup_z II(z)] \leqslant \mathbb{E}[\sup_z A] = O(DG \cdot M\sqrt{T})$.

Step 4. Show $\mathbb{E}[III] \leqslant 0$. This is obviously true.

Step 5. Show $\mathbb{E}[IV] = \tilde{O}(\gamma DG \cdot M\sqrt{T})$. Define $\gamma = G/G_0$. By A.2 we have $\|g_t^m - M_t^m\|_* \leqslant 2G$. It holds almost surely that

$$\begin{aligned}
IV &\leqslant 2G \sum_{m=1}^M \sum_{t=1}^T \|z_t^m - \tilde{z}_t^m\| \\
&\leqslant 2G\sqrt{T} \cdot \sum_{m=1}^M \sqrt{\sum_{t=1}^T \|z_t^m - \tilde{z}_t^m\|^2} \\
&\leqslant 2G\sqrt{T} \cdot \sum_{m=1}^M \sqrt{\sum_{t=1}^T (\eta_t^m Z_t^m)^2} \\
&= 2G\sqrt{T} \cdot D \cdot \sum_{m=1}^M \sqrt{\sum_{t=1}^T \frac{(Z_t^m)^2}{G_0^2 + \sum_{\tau=1}^{t-1}(Z_\tau^m)^2}} \\
&\leqslant 2GD\sqrt{T} \cdot \sum_{m=1}^M \sqrt{2 + 4\gamma^2 + 2\log\Big(\frac{G_0^2 + \sum_{t=1}^{T-1}(Z_t^m)^2}{G_0^2}\Big)} && \text{(Lemma C.1)} \\
&\leqslant 2GD\sqrt{T} \cdot \sum_{m=1}^M \sqrt{2 + 4\gamma^2 + 2\log\Big(\frac{G_0^2 + G^2T}{G_0^2}\Big)} \\
&\leqslant 2GD\sqrt{T} \cdot \sum_{m=1}^M \sqrt{2 + 4\gamma^2 + 2\log(1 + \gamma^2 T)} \\
&= \tilde{O}(\gamma GD \cdot M\sqrt{T})
\end{aligned}$$

Finally, we plug in the upper bounds for I–IV and continue Eq (19).

$$TM \cdot \mathbb{E}[\mathrm{DualGap}(\bar{z})] = \tilde{O}(\gamma DG \cdot M\sqrt{T} + \sigma D\sqrt{MT}).$$

This finishes the proof of Theorem 5.1 □

B.2 Proof of Theorem 5.2

Proof of Theorem 5.2, Smooth Case. The proof strategy follows closely that of Bach and Levy [6]. Using the notation for Step 1 in the proof for nonsmooth case, we have the bound

$$TM\mathbb{E}[\text{DualGap}(\bar{z})] \leqslant \mathbb{E}[\sup_{z}\{I(z) + II(z) + III + IV\}],$$

where I–IV are defined in Eqs. (14)–(17). We deal with these terms in a different manner.

For the term $I(z)$ in Eq. (14), following Step 2 we have $\mathbb{E}[\sup_z I(z)] = O(\gamma\sigma D\sqrt{MT})$.

Next we define a stopping time. For each $m \in [M]$, let

$$\tau_m^* := \{\max t \in [T] : \frac{1}{\eta_t^m} \leqslant 1/(2L)\}. \tag{21}$$

Recall our learning rate scheme for the smooth case

$$\eta_1^m = \frac{D\alpha}{G_0}, \quad \eta_t^m = \frac{D\alpha}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1}(Z_\tau^m)^2}}.$$

For the term $II(z)$ in Eq. (15), we follow Step 3 and obtain for all $z \in \mathcal{Z}$,

$$II(z) \leqslant \sum_{m=1}^{M}\left\{\frac{D^2}{\eta_1^m} + \frac{D^2}{\eta_T^m}\right\}.$$

By the definition of η_1^m, we have $\sum_{m=1}^M D^2/\eta_1^m \leqslant DMG_0/\alpha$. For the second term, for fixed $m \in [M]$,

$$\begin{aligned}
\sum_{m=1}^{M} D^2/\eta_T^m &= \sum_{m=1}^{M}\frac{D}{\alpha}\sqrt{G_0^2 + \sum_{t=1}^{T-1}(Z_t^m)^2} && (22)\\
&\leqslant \sum_{m=1}^{M}\frac{D}{\alpha}\left(G_0 + \sum_{t=1}^{T}\frac{(Z_t^m)^2}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1}(Z_\tau^m)^2}}\right) && \text{(Lemma C.2)}\\
&= \frac{MDG_0}{\alpha} + \underbrace{\sum_{m=1}^{M}\sum_{t=1}^{T}\frac{1}{\alpha^2}(\eta_t^m)^2(Z_t^m)^2}_{:=\mathcal{A}} && (23)
\end{aligned}$$

So we have $\mathbb{E}[\sup_z II(z)] \leqslant 2\gamma MDG/\alpha + \mathbb{E}[\mathcal{A}]$.

For the term III in Eq. (16), we also split it into two parts by τ_m^*.

$$\begin{aligned}
III &:= -\sum_{t=1}^{T}\sum_{m=1}^{M}\frac{1}{\eta_t^m}(\tfrac{1}{2}\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + \tfrac{1}{2}\|z_t^m - \tilde{z}_t^m\|^2) && (24)\\
&= -\sum_{t=1}^{T}\sum_{m=1}^{M}\frac{5}{2}\eta_t^m(Z_t^m)^2 && (25)\\
&= -\underbrace{\sum_{m=1}^{M}\sum_{t=1}^{\tau_m^*}\frac{5}{2}\eta_t^m(Z_t^m)^2}_{\geqslant 0} - \underbrace{\sum_{m=1}^{M}\sum_{t=\tau_m^*+1}^{T}\frac{5}{2}\eta_t^m(Z_t^m)^2}_{:=\mathcal{B}_{\text{tail}}} && (26)
\end{aligned}$$

For the term IV in defined in Eq. (17), we first introduce a margtingale difference sequence. For all $t \in [T], m \in [M]$, let

$$\zeta_t^m := \left(g_t^m - G(z_t^m)\right) + \left(M_t^m - G(\tilde{z}_{t-1}^{m,*})\right). \tag{27}$$

By the triangular inequality, we have

$$IV := \sum_{t=1}^{T}\sum_{m=1}^{M} \|g_t^m - M_t^m\|_* \cdot \|z_t^m - \tilde{z}_t^m\| \tag{28}$$

$$\leqslant \underbrace{\sum_{t=1}^{T}\sum_{m=1}^{M} \|\zeta_t^m\|_* \cdot \|z_t^m - \tilde{z}_t^m\|}_{:=V} + \sum_{t=1}^{T}\sum_{m=1}^{M} \|G(z_t^m) - G(\tilde{z}_{t-1}^{m,*})\|_* \cdot \|z_t^m - \tilde{z}_t^m\| \tag{29}$$

$$\leqslant V + \sum_{t=1}^{T}\sum_{m=1}^{M} \left(\frac{L}{2}\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + \frac{L}{2}\|z_t^m - \tilde{z}_t^m\|^2\right) \tag{30}$$

$$= V + \sum_{t=1}^{T}\sum_{m=1}^{M} \frac{5L}{2}(\eta_t^m)^2 (Z_t^m)^2 \tag{31}$$

$$= V + \underbrace{\sum_{m=1}^{M}\sum_{t=1}^{\tau_m^*} \frac{5L}{2}(\eta_t^m)^2 (Z_t^m)^2}_{:=\mathcal{C}_{\text{head}}} + \underbrace{\sum_{m=1}^{M}\sum_{t=\tau_m^*+1}^{T} \frac{5L}{2}(\eta_t^m)^2 (Z_t^m)^2}_{:=\mathcal{C}_{\text{tail}}} \tag{32}$$

Eq. (30) holds due to smoothness, i.e., for all $z, z' \in \mathcal{Z}$, $\|G(z) - G(z')\|_* \leqslant L\|z - z'\|$. Eq. (30) thus follows:

$$\begin{aligned}
&\|G(z_t^m) - G(\tilde{z}_{t-1}^{m,*})\|_* \cdot \|z_t^m - \tilde{z}_t^m\| \\
&\leqslant \frac{1}{2L}\|G(z_t^m) - G(\tilde{z}_{t-1}^{m,*})\|_*^2 + \frac{L}{2}\|z_t^m - \tilde{z}_t^m\|^2 \\
&\leqslant \frac{L}{2}\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + \frac{L}{2}\|z_t^m - \tilde{z}_t^m\|^2.
\end{aligned}$$

To summarize, we have shown

$$TM \cdot \mathbb{E}[\text{DualGap}(\bar{z})] \leqslant \mathbb{E}[\sup_z\{I(z) + II(z) + III + IV\}] \tag{33}$$

$$\leqslant O\left(\gamma\sigma D\sqrt{MT}\right) + 2\gamma MDG/\alpha \tag{34}$$

$$+ \mathbb{E}[\mathcal{A} + \mathcal{C}_{\text{head}} + (-\mathcal{B}_{\text{tail}} + \mathcal{C}_{\text{tail}}) + V]. \tag{35}$$

Step a. Show $\mathbb{E}[\mathcal{A}] \leqslant 8\gamma GDM/\alpha + 3DM\mathcal{V}_1(T)/\alpha$. Recall its definition in Eq. (50).

$$\begin{aligned}
\mathcal{A} &:= \sum_{m=1}^{M}\sum_{t=1}^{T} \frac{1}{\alpha^2}(\eta_t^m)^2 (Z_t^m)^2 \\
&= \frac{D}{\alpha}\sum_{m=1}^{M}\sum_{t=1}^{T} \frac{(Z_t^m)^2}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1}(Z_\tau^m)^2}} \\
&\leqslant \frac{D}{\alpha}\sum_{m=1}^{M}\left(5\gamma G + 3\sqrt{G_0^2 + \sum_{t=1}^{T-1}(Z_t^m)^2}\right) && \text{(Lemma C.2)} \\
&\leqslant \frac{D}{\alpha}\sum_{m=1}^{M}\left(8\gamma G + 3\sqrt{\sum_{t=1}^{T-1}(Z_t^m)^2}\right)
\end{aligned}$$

Note by Lemma B.1 we know $(Z_t^m)^2 \leqslant (\|g_t^m\|_*^2 + \|M_t^m\|_*^2)/5 \leqslant \|g_t^m\|_*^2 + \|M_t^m\|_*^2$. Recall the definition of $\mathcal{V}_m(T)$ in Eq. (6). By the symmetry of the algorithm over all workers, we know

$\mathcal{V}_1(T) = \mathcal{V}_m(T)$ for all $m \in [M]$. Then

$$\begin{aligned}
\mathbb{E}[\mathcal{A}] &\leqslant 8\gamma DMG/\alpha + \frac{3D}{\alpha}\sum_{m=1}^{M}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T-1}(Z_t^m)^2}\right] \\
&\leqslant 8\gamma DMG/\alpha + \frac{3D}{\alpha}\sum_{m=1}^{M}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T-1}\|g_t^m\|_*^2 + \|M_t^m\|_*^2}\right] \\
&= 8\gamma DMG/\alpha + \frac{3D}{\alpha}\sum_{m=1}^{M}\mathcal{V}_m(T) = 8\gamma DMG/\alpha + 3DM\mathcal{V}_1(T)/\alpha.
\end{aligned}$$

By our choice of α we have $\mathbb{E}[\mathcal{A}] = O(\gamma DM^{3/2}G + DM^{3/2}\mathcal{V}_1(T))$.

Step b. Show $\mathbb{E}[\mathcal{C}_{\text{head}}] = O(1)$. Recall its definition in Eq. (32).

$$\begin{aligned}
\mathcal{C}_{\text{head}} &:= \sum_{m=1}^{M}\sum_{t=1}^{\tau_m^*}\frac{5L}{2}(\eta_t^m)^2(Z_t^m)^2 && (36)\\
&= \frac{5\alpha^2D^2L}{2}\sum_{m=1}^{M}\sum_{t=1}^{\tau_m^*}\frac{(Z_t^m)^2}{G_0^2 + \sum_{\tau=1}^{t-1}(Z_\tau^m)^2} && (37)\\
&\leqslant \frac{5\alpha^2D^2L}{2}\sum_{m=1}^{M}\left(6\gamma^2 + 2\log\left(\frac{G_0^2 + \sum_{t=1}^{\tau_m^*-1}(Z_\tau^m)^2}{G_0^2}\right)\right) && \text{(Lemma C.1)}\\
&= \frac{5\alpha^2D^2L}{2}\sum_{m=1}^{M}\left(6\gamma^2 + 2\log\left(\frac{\alpha^2D^2}{G_0^2(\eta_{\tau_m^*}^m)^2}\right)\right) && (38)\\
&\leqslant \frac{5\alpha^2D^2LM}{2}\left(6\gamma^2 + 4\log\left(\frac{\alpha D}{2G_0L}\right)\right) && (39)
\end{aligned}$$

The last inequality is due to definition of τ_m^*. By our choice of α we have $\mathbb{E}[\mathcal{C}_{\text{head}}] = \tilde{O}(\gamma^2LD^2)$.

Step c. Show $\mathcal{C}_{\text{tail}} - \mathcal{B}_{\text{tail}} \leqslant 0$. Recall $\mathcal{B}_{\text{tail}}$ is defined in Eq. (26). By definition,

$$\mathcal{C}_{\text{tail}} - \mathcal{B}_{\text{tail}} = \sum_{m=1}^{M}\sum_{t=\tau_m^*+1}^{T}\left(\frac{5L}{2}\eta_t^m - \frac{5}{2}\right)\eta_t^m(Z_t^m)^2.$$

We show $\frac{5L}{2}\eta_t^m - \frac{5}{2} \leqslant 0$ for all $t \in [T], m \in [M]$. Note that for all $t \geqslant \tau_m^* + 1$ we have $\eta_t^m \leqslant 1/(2L)$. Thus $\frac{5L}{2}\eta_t^m - \frac{5}{2} \leqslant -5/4$. Thus $\mathcal{C}_{\text{tail}} - \mathcal{B}_{\text{tail}} \leqslant 0$.

Step d. Show $\mathbb{E}[V] = \tilde{O}(\gamma\sigma D\sqrt{MT})$. Recall its definition in Eq.(29). Also note $\mathbb{E}[\|\zeta_t^m\|_*^2] \leqslant 4\sigma^2$.

$$\mathbb{E}[V] := \mathbb{E}\left[\sum_{t=1}^{T}\sum_{m=1}^{M}\|\zeta_t^m\|_* \cdot \|z_t^m - \tilde{z}_t^m\|\right] \tag{40}$$

$$\leqslant \mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\sum_{m=1}^{M}\|\zeta_t^m\|_*^2}\right] \cdot \mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\sum_{m=1}^{M}\|z_t^m - \tilde{z}_t^m\|^2}\right] \tag{41}$$

$$\leqslant \sqrt{\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}\big[\|\zeta_t^m\|_*^2\big]} \cdot \mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\sum_{m=1}^{M}\|z_t^m - \tilde{z}_t^m\|^2}\right] \tag{42}$$

$$\leqslant 2\sigma\sqrt{MT} \cdot \mathbb{E}\left[\sqrt{\sum_{m=1}^{M}\sum_{t=1}^{T}\|z_t^m - \tilde{z}_t^m\|^2}\right] \tag{43}$$

$$\leqslant 2\sigma\sqrt{MT} \cdot \mathbb{E}\left[\sqrt{\sum_{m=1}^{M}\sum_{t=1}^{T}\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + \|z_t^m - \tilde{z}_t^m\|^2}\right] \tag{44}$$

$$= 2\sigma\sqrt{MT} \cdot \mathbb{E}\left[\sqrt{\sum_{m=1}^{M}\sum_{t=1}^{T}5 \cdot (\eta_t^m)^2 (Z_t^m)^2}\right] \tag{45}$$

$$= 2\sqrt{5} \cdot \sigma\sqrt{MT} \cdot D\alpha \cdot \mathbb{E}\left[\sqrt{\sum_{m=1}^{M}\sum_{t=1}^{T}\frac{(Z_t^m)^2}{G_0^2 + \sum_{\tau=1}^{t-1}(Z_\tau^m)^2}}\right] \tag{46}$$

$$\leqslant 6 \cdot \sigma\sqrt{MT} \cdot D\alpha \cdot \mathbb{E}\left[\sqrt{\sum_{m=1}^{M}\left(6\gamma^2 + 2\log\left(\frac{G_0^2 + \sum_{t=1}^{T-1}(Z_t^m)^2}{G_0^2}\right)\right)}\right] \quad \text{(Lemma C.1)}$$

$$\leqslant 6\sigma\sqrt{MT} \cdot D\alpha \cdot \sqrt{M(6\gamma^2 + 2\log(1+\gamma^2 T))}. \tag{47}$$

By our choice of α, we have $\mathbb{E}[V] = \tilde{O}(\gamma\sigma D\sqrt{MT})$.

Continuing Eq. (53), we have

$$\begin{aligned}
&TM \cdot \mathbb{E}[\mathrm{DualGap}(\bar{z})] \\
&\leqslant O\Big(\gamma\sigma D\sqrt{MT}\Big) + 2\gamma MDG/\alpha + \mathbb{E}[\mathcal{A} + \mathcal{C}_{\text{head}} + (-\mathcal{B}_{\text{tail}} + \mathcal{C}_{\text{tail}}) + V] \\
&= \tilde{O}\Big(\gamma\sigma D\sqrt{MT} + \underbrace{\gamma DM^{3/2}G + DM^{3/2}\mathcal{V}_m(T)}_{\mathcal{A}} + \underbrace{\gamma^2 LD^2}_{\mathcal{C}_{\text{head}}} + \underbrace{\gamma\sigma D\sqrt{MT}}_{V}\Big).
\end{aligned}$$

This finishes the proof of Theorem 5.2.

□

Remark 5 (Getting rid of $\mathcal{V}_1(T)$)**.** *We could also use the free parameters α (base learning rate) and obtain the following near linear speed-up result.*

Theorem B.3 (Smooth Case, free of $\mathcal{V}_1(T)$)**.** *Assume A.1, A.2, A.3 and A.4. Let σ, D, G, L be defined therein. For any $\epsilon \in (0, \frac{1}{2})$, let $\bar{z} = \mathrm{LocalAdaSEG}(G_0, D; K, M, R; T^\epsilon/\sqrt{M})$. If $T \geqslant M^{1/(2\epsilon)}$, then*

$$\mathbb{E}[\mathrm{DualGap}(\bar{z})] = \tilde{O}\left(\frac{\sigma D}{\sqrt{MT^{1-2\epsilon}}} + \frac{\gamma^2 LD^2}{T^{1-2\epsilon}} + \frac{LD^2 M}{T} + \frac{\gamma GDM^{3/2}}{T^{1+\epsilon}}\right),$$

where $\tilde{O}$ hides absolute constants, logarithmic factors of problem parameters and logarithmic factors of T.

Proof of Theorem B.3. We deocompose the term II in Eq.(15) in a different way. Recall in Step 3 we have shown for all $z \in \mathcal{Z}$, $II(z) \leqslant \sum_{m=1}^{M} \frac{D^2}{\eta_1^m} + \frac{D^2}{\eta_T^m}$. For the second term, for fixed $m \in [M]$,

$$\sum_{m=1}^{M} D^2/\eta_T^m = \sum_{m=1}^{M} \frac{D}{\alpha} \sqrt{G_0^2 + \sum_{t=1}^{T-1} (Z_t^m)^2} \tag{48}$$

$$\leqslant \sum_{m=1}^{M} \frac{D}{\alpha} \left(G_0 + \sum_{t=1}^{T} \frac{(Z_t^m)^2}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} (Z_\tau^m)^2}} \right) \qquad \text{(Lemma C.2)}$$

$$= \frac{MDG_0}{\alpha} + \sum_{m=1}^{M} \sum_{t=1}^{T} \frac{1}{\alpha^2} (\eta_t^m)^2 (Z_t^m)^2 \tag{49}$$

$$\leqslant \frac{\gamma MDG}{\alpha} + \underbrace{\sum_{m=1}^{M} \sum_{t=1}^{\tau_m^*} \frac{1}{\alpha^2} (\eta_t^m)^2 (Z_t^m)^2}_{:=\mathcal{A}_{\text{head}}} + \underbrace{\sum_{m=1}^{M} \sum_{t=\tau_m^*+1}^{T} \frac{1}{\alpha^2} (\eta_t^m)^2 (Z_t^m)^2}_{:=\mathcal{A}_{\text{tail}}} \tag{50}$$

So we have $\mathbb{E}[\sup_z II(z)] \leqslant 2\gamma MDG/\alpha + \mathbb{E}[\mathcal{A}_{\text{head}} + \mathcal{A}_{\text{tail}}]$. Then, following the proof in the smooth case, we have

$$TM \cdot \mathbb{E}[\text{DualGap}(\bar{z})] \leqslant \mathbb{E}[\sup_z \{I(z) + II(z) + III + IV\}] \tag{51}$$

$$\leqslant O\Big(\gamma\sigma D\sqrt{MT}\Big) + 2\gamma MDG/\alpha \tag{52}$$

$$+ \mathbb{E}[\mathcal{A}_{\text{head}} + \mathcal{C}_{\text{head}} + (\mathcal{A}_{\text{tail}} - \mathcal{B}_{\text{tail}} + \mathcal{C}_{\text{tail}}) + V]. \tag{53}$$

Recall our choice of $\alpha = T^\epsilon/\sqrt{M}$.

Show $\mathbb{E}[\mathcal{A}_{\text{head}}] = \tilde{O}(1)$. Recall its definition in Eq. (50).

$$\begin{aligned}
\mathcal{A}_{\text{head}} &:= \sum_{m=1}^{M} \sum_{t=1}^{\tau_m^*} \frac{1}{\alpha^2} (\eta_t^m)^2 (Z_t^m)^2 \\
&= \frac{D}{\alpha} \sum_{m=1}^{M} \sum_{t=1}^{\tau_m^*} \frac{(Z_t^m)^2}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} (Z_\tau^m)^2}} \\
&\leqslant \frac{D}{\alpha} \sum_{m=1}^{M} \left(5\gamma G + 3\sqrt{G_0^2 + \sum_{t=1}^{\tau_m^*-1} (Z_t^m)^2} \right) \qquad \text{(Lemma C.2)} \\
&= \frac{D}{\alpha} \sum_{m=1}^{M} \left(5\gamma G + \frac{3D\alpha}{\eta_{\tau_m^*}^m} \right) \\
&\leqslant \frac{D}{\alpha} \sum_{m=1}^{M} \Big(5\gamma G + 6\alpha LD \Big) = \frac{5\gamma GDM}{\alpha} + 6LD^2 M.
\end{aligned}$$

By our choice of α we have $\mathbb{E}[\mathcal{A}_{\text{head}}] \leqslant 5\gamma GDM^{3/2}T^{-\epsilon} + 6LD^2M$.

For $\mathcal{C}_{\text{head}}$ defined in Eq. (32), following Eq (39), we have $\mathbb{E}[\mathcal{C}_{\text{head}}] = \tilde{O}(\gamma^2 LD^2 T^{2\epsilon})$.

Show $\mathcal{A}_{\text{tail}} + \mathcal{C}_{\text{tail}} - \mathcal{B}_{\text{tail}} \leqslant 0$. Recall $\mathcal{B}_{\text{tail}}$ is defined in Eq. (26). By definition,

$$\mathcal{A}_{\text{tail}} + \mathcal{C}_{\text{tail}} - \mathcal{B}_{\text{tail}} = \sum_{m=1}^{M} \sum_{t=\tau_m^*+1}^{T} \Big(\frac{1}{\alpha^2} + \frac{5L}{2}\eta_t^m - \frac{5}{2} \Big) \eta_t^m (Z_t^m)^2.$$

We show $\frac{1}{\alpha^2} + \frac{5L}{2}\eta_t^m - \frac{5}{2} \leqslant 0$ for all $t \in [T], m \in [M]$. Note that

$$T \geqslant M^{1/(2\epsilon)} \implies \alpha^2 = (T^\epsilon/\sqrt{M})^2 \geqslant 1,$$

and that for all $t \geqslant \tau_m^* + 1$ we have $\eta_t^m \leqslant 1/(2L)$. Thus $\frac{1}{\alpha^2} + \frac{5L}{2}\eta_t^m - \frac{5}{2} \leqslant -1/4$. Thus $\mathcal{A}_{\text{tail}} + \mathcal{C}_{\text{tail}} - \mathcal{B}_{\text{tail}} \leqslant 0$.

For V defined in Eq. (29), following Eq. (47), $\mathbb{E}[V] = \tilde{O}(\gamma\sigma D\sqrt{MT^{1+2\epsilon}})$.

Putting together we have

$$\begin{aligned}
&TM \cdot \mathbb{E}[\text{DualGap}(\bar{z})] \\
&\leqslant O\Big(\gamma\sigma D\sqrt{MT}\Big) + 2\gamma MDG/\alpha + \mathbb{E}[\mathcal{A}_{\text{head}} + \mathcal{C}_{\text{head}} + (\mathcal{A}_{\text{tail}} - \mathcal{B}_{\text{tail}} + \mathcal{C}_{\text{tail}}) + V] \\
&= \tilde{O}\Big(\gamma\sigma D\sqrt{MT} + \underbrace{\gamma GDM^{3/2}T^{-\epsilon} + LD^2M}_{\mathcal{A}_{\text{head}}} + \underbrace{\gamma^2 LD^2T^{2\epsilon}}_{\mathcal{C}_{\text{head}}} + \underbrace{\gamma\sigma D\sqrt{MT^{1+2\epsilon}}}_{V}\Big).
\end{aligned}$$

This finishes the proof of Theorem B.3 □

C Helper Lemmas

Lemma C.1. *For any non-negative real numbers $a_1, \ldots, a_n \in [0, a]$, and $a_0 > 0$, it holds*

$$\sum_{i=1}^{n} \frac{a_i}{a_0 + \sum_{j=1}^{i-1} a_j} \leqslant 2 + \frac{4a}{a_0} + 2\log\Big(1 + \sum_{i=1}^{n-1} a_i/a_0\Big).$$

Proof of Lemma C.1. See Lemma A.2 of [6].

□

Lemma C.2. *For any non-negative numbers $a_1, \ldots, a_n \in [0, a]$, and $a_0 > 0$, it holds*

$$\sqrt{a_0 + \sum_{i=1}^{n-1} a_i} - \sqrt{a_0} \leqslant \sum_{i=1}^{n} \frac{a_i}{\sqrt{a_0 + \sum_{j=1}^{i-1} a_j}} \leqslant \frac{2a}{a_0} + 3\sqrt{a} + 3\sqrt{a_0 + \sum_{i=1}^{n-1} a_i}.$$

Proof of Lemma C.1. See Lemma A.1 of [6].

□

Lemma C.3. *9 For any sequence $\{z_t\}_{t=1}^T \subset \mathcal{Z}^o$, let $\bar{z}$ denote its mean. It holds*

$$T \cdot \text{DualGap}(\bar{z}) \leqslant \sup_{z \in \mathcal{Z}} \sum_{t=1}^{T} \big\langle z_t - z, G(z_t) \big\rangle.$$

Proof of Lemma C.3. This lemma depends on the convexity-concavity of the saddle function F.

Denote $\bar{z} := [\bar{x}, \bar{y}]$, $z_t := [x_t, y_t]$. Note $\bar{x} = (1/T)\sum_{t=1}^T x_t$ and $\bar{y} = (1/T)\sum_{t=1}^T y_t$. By definition of duality gap and the convexity-concavity of F,

$$\begin{aligned}
\text{DualGap}(\bar{z}) &:= \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} F(\bar{x}, y) - F(x, \bar{y}) \\
&\leqslant \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} \frac{1}{T}\sum_{t=1}^{T} F(x_t, y) - \frac{1}{T}\sum_{t=1}^{T} F(x, y_t).
\end{aligned}$$

Let $G(z_t) = G(x_t, y_t) := [d_{x,t}, -d_{y,t}]$. Since $d_{x,t} \in \partial_x F(x_t, y_t)$, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$F(x_t, y) + \langle d_{x,t}, x - x_t \rangle \leqslant F(x, y).$$

Similarly, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, it holds

$$F(x, y_t) + \langle d_{y,t}, y - y_t \rangle \geqslant F(x, y).$$

We have

$$T \cdot \mathrm{DualGap}(\bar{z}) \leqslant \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} \sum_{t=1}^{T} \langle d_{x,t}, x_t - x \rangle - \langle d_{y,t}, y_t - y \rangle$$
$$= \sup_{z \in \mathcal{Z}} \sum_{t=1}^{T} \langle G(z_t), z_t - z \rangle.$$

This completes the proof of Lemma C.3. □

D Experiments

We implement our algorithm and conduct all the experiments on a computer with Intel Core i5 CPU @ 3.20GHz cores, 8GB RAM, and GPU @ GeForce RTX 3090. The deep learning framework we use is PyTorch 1.8.1. The OS environment was created by conda over Ubuntu 20.04. We use Python 3.7. Python library requirement is specified in the configuration file provided in the supplemental materials. Due to the hardware limitation, we simulate the distributed environment by creating object instances to simulate multiple clients and a central server on one GPU card.

D.1 Wasserstein GAN

Inspired by game theory, generative adversarial networks (GANs) have shown great performance in many generative tasks to replicate the real-world rich content, such as images, texts and music. GANs are composed of two models, a generator and a discriminator, which are competing with each other to improve the performance for a specific task. In this experiment, we aim to train a digit image generator using the MNIST dataset.

It is challenging to train a GAN model due to the slow convergence speed, instability of training or even failure to converge. [2, 3] proposed to use the Wasserstein distance as the GAN loss function to provide stable and fast training. To enforce the Lipschitz constraint on the discriminator, we adopt WGAN with gradient penalty as our experimental model. The objective can be described as

$$\min_G \max_D \mathop{\mathbb{E}}_{x \sim \mathbb{P}_r} [D(x)] - \mathop{\mathbb{E}}_{z \sim \mathbb{P}_z} [D(G(z))] - \lambda [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \tag{54}$$

where D and G denote the generator and discriminator, $\mathbb{P}_r$ is the data distribution, and $\mathbb{P}_z$ represents the noise distribution (uniform or Gaussian distribution). The point $\hat{x} \sim \mathbb{P}_{\hat{x}}$ is sampled uniformly along straight lines between pairs of points sampled from the real data distribution $\mathbb{P}_r$ and the generator distribution $\mathbb{P}_{\tilde{x}}$, expressed as $\hat{x} := \epsilon x + (1 - \epsilon)\tilde{x}$, where $\epsilon \sim U[0, 1]$.

DCGAN. We implement WGAN with the DCGAN architecture, which improves the original GAN with convolutional layers. Specifically, the generator consists of 3 blocks, which contain deconvolutional layers, batch normalization and activations. The details of the whole generator can be represented as sequential layers *{Linear, BN, ReLU, DeConv, BN, ReLU, DeConv, BN, ReLU, DeConv, Tanh}*, where *Linear*, *BN*, *DeConv* denote the linear, batch normalization and deconvolutional layer, respectively. *ReLU* and *Tanh* represent the activation functions. Similarly, the discriminator also contains 3 blocks, which can be described as sequential layers *{Conv, LReLU, Conv, LReLU, Conv, LReLU, Linear}*, where *Conv* and *LReLU* denote the convolutional layer and Leaky-ReLU activation function, respectively.

Inception score (IS). Inception score (IS) is proposed to evaluate the performance of a GAN with an inception model. IS measures GAN from two aspects simultaneously. Firstly, GAN should output a high diversity of images. Secondly, the generated images should contain clear objects. Specifically, we feed the generated images x into a well-trained inception model to obtain the output y. Then, IS can be calculated by the following equation:

$$IS = \exp(\mathop{\mathbb{E}}_{x \sim \mathbb{P}_g} [D_{KL}(p(y|x) \| p(y))]), \tag{55}$$

where $\mathbb{P}_g$ is the generator model distribution. Essentially, IS computes the mutual information $I(y; x) = H(y) - H(y|x)$, where $H()$ denotes the entropy. The higher $H(y)$ indicates higher diversity of generated images. The lower $H(y|x)$ implies the input x belongs to one class with

higher probability. In summary, IS is bounded by $1 \leqslant IS \leqslant 1000$. The higher IS implies a better performance of a GAN.

Fréchet inception distance (FID). Although IS can measure the diversity and quality of the generated images, it still has some limitations, such as losing the sight of the true data distribution, failure to measure the model generalization. FID is an improved metric for GAN, which cooperates with the training samples and generated samples to measure the performance together. Specifically, we feed the generated samples and training samples into an inception model to extract the feature vectors, respectively. Usually, we extract the logits value before the last sigmoid activation as the feature vector with dimension 2048. Essentially, FID is the Wasserstein metric between two multidimensional Gaussian distributions: $\mathcal{N}(\mu_g, \Sigma_g)$ the distribution of feature vectors from generated samples and $\mathcal{N}(\mu_r, \Sigma_r)$ the distribution of feature vectors from the training samples. It can be calculated as

$$FID = \|u_r - u_g\|^2 + tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \tag{56}$$

where $tr()$ denotes the trace of a matrix. The lower the FID, the better the performance of a GAN.