Machine Learning
https://doi.org/10.1007/s10994-022-06239-z

Local AdaGrad-type algorithm for stochastic convex-concave optimization

Luofeng Liao[1,2] · **Li Shen**[2] · **Jia Duan**[2] · **Mladen Kolar**[3] · **Dacheng Tao**[2]

Received: 20 May 2022 / Revised: 6 August 2022 / Accepted: 12 September 2022

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 problems in the Parameter-Server model. LocalAdaSEG has three main features: (1) a periodic communication strategy that reduces the communication cost between workers and the server; (2) an adaptive learning rate that is computed locally and allows for tuning-free implementation; and (3) theoretically, a nearly linear speed-up with respect to the dominant variance term, arising from the 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 a generative adversarial network. We compare LocalAdaSEG against several existing optimizers for minimax problems and demonstrate its efficacy through several experiments in both homogeneous and heterogeneous settings.

Editors: Yu-Feng Li, Prateek Jain.

✉ Li Shen
mathshenli@gmail.com

Luofeng Liao
ll3530@columbia.edu

Jia Duan
xuelandj@gmail.com

Mladen Kolar
mkolar@chicagobooth.edu

Dacheng Tao
dacheng.tao@gmail.com

[1] IEOR, Columbia Unviersity, New York, NY, USA

[2] JD Explore Academy, JD.com Inc, Beijing, China

[3] Booth School of Business, University of Chicago, Chicago, IL, USA

Published online: 23 November 2022

Keywords Stochastic minimax problem · Adaptive optimization · Distributed computation

1 Introduction

Stochastic minimax optimization problems arise in applications ranging from game theory (Neumann, 1928), robust optimization (Delage & Ye, 2010), and AUC Maximization (Guo et al., 2020), to adversarial learning (Wang et al., 2019) and training of generative adversarial networks (GANs) (Goodfellow et al., 2014). In this work, we consider

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\left\{F(x,y)=\int_{\Xi}f(x,y,\xi)P(\mathrm{d}\xi)\right\},\tag{1}$$

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 $\mathbb{E}_{\xi\sim P}[f(x,y,\xi)]$ 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 an efficient solver for the large-scale minimax problem (1). First, the solver should generate converging iterates. In contrast to convex optimization, convergence results for minimax problems are harder to obtain. Second, the solver should be able to take advantage of parallel computing in a communication-efficient way. Only then can it be applied to problems with large-scale datasets, which are often distributed across multiple workers. Third, it is desirable for the solver to choose learning rates in an adaptive manner. It is well known that, in minimax problems, solver performance is susceptible to learning rates. We discuss these challenges in detail below.

First, it has been shown that direct application of the (stochastic) gradient descent ascent ((S)GDA) to solve (1) may result in divergence of the iterates (Mertikopoulos et al., 2019; Daskalakis et al., 2018; Gidel et al., 2019; Mertikopoulos et al., 2018). Possible ways to overcome the divergence issue are to apply the primal-dual hybrid gradient (PDHG) or (stochastic) extragradient method and their variants (Mertikopoulos et al., 2019; Daskalakis et al., 2018; Gidel et al., 2019; Azizian et al., 2020; Liu et al., 2020; Zhao, 2021; Zhang et al., 2020).

Second, it is often desirable to have a communication-efficient distributed solver to solve the stochastic minimax problem (1). The first reason being that the minimax problem (1) is often instantiated as a finite-sum problem with large-scale datasets (with the distribution P being the empirical distribution over millions of data points), and thus storing and manipulating datasets on multiple workers is a must. For example, when problem (1) is specified as BigGAN Brock et al. (2019) over ImageNet (Deng et al., 2009), the number of training samples is as many as 14 million. Traditional distributed SGDA on the problem (1) may suffer from a considerable communication burden; reducing communication complexity of the algorithm is a major concern in our paper. The second reason is that, in some scenarios, data are distributed on mobile devices (such as cell phones or smart watches), and due to privacy concerns, local data must stay on the device. Furthermore, frequent communication among devices is not feasible due to failures of mobile devices (network connectivity, battery level, etc.). This further motivates the design of communication-efficient distributed solvers to eliminate central data storage and improve communication efficiency. For these reasons, communication-efficient distributed solvers for minimax problems have

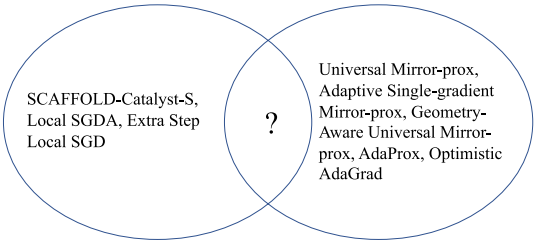

Fig. 1 A Venn Diagram for related works. Left circle: communication-efficient methods for stochastic minimax problems. Right circle: Adaptive methods for stochastic minimax problems

been investigated recently (Beznosikov et al., 2021; Deng & Mahdavi, 2021; Hou et al., 2021; Liu et al., 2020).

Third, the performance of stochastic minimax solvers for (1) is highly dependent on the learning rate tuning mechanism (Heusel et al., 2017; Antonakopoulos et al., 2021). And yet, designing a solver for (1) with an adaptive learning rate is much more challenging compared to the convex case; the value of F at an iterate (x, y) does *not* serve as a performance criterion. 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 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 (Bach & Levy, 2019; Babanezhad & Lacoste-Julien, 2020; Ene & Nguyen, 2020; Antonakopoulos et al., 2021; Liu et al., 2020). 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 Fig. 1 illustrates the difference between LocalAdaSEG algorithm and the existing works. LocalAdaSEG falls under the umbrella of the Parameter-Server model (Smola & Narayanamurthy, 2010) and adopts a periodic communication mechanism to reduce the communication cost between the server and the workers, similar to local SGD/FedAvg (Yu et al., 2019; Stich, 2019; Li et al., 2020) in federated learning (McMahan et al., 2021). In addition, in each worker, a local stochastic extragradient algorithm with an adaptive learning rate is performed independently with multiple iterations. Every once in a while, current iterates and adaptive learning rates from all workers are sent to the server. The server computes a weighted average of the iterates, where the weights are constructed from the received local adaptive learning rates. We emphasize that adaptive learning in each worker is distinct from others and is automatically updated according to local data as is done in (Chen et al., 2021; Beznosikov et al., 2021), and different from the existing adaptive distributed algorithms (Xie et al., 2019; Reddi et al., 2021; Chen et al., 2021).

Theoretically optimal convergence rate Let M denote the number of workers, σ denote the variance of stochastic gradients, and T denote the number of local iterations on each worker. For stochastic convex-concave minimax problems, we establish the rate in terms of

the duality gap metric (Nemirovski, 2004; Lin et al., 2020) as $\tilde{O}(\sigma/\sqrt{MT})$ in the *nonsmooth and noise-dominant* case and the rate $\tilde{O}(\sigma/\sqrt{MT} +$ higher-order terms) in *smooth case with slow cumulative gradient growth*. The terms depending on the variance σ achieve the statistical lower bound and are not improvable without further assumptions. Therefore, the LocalAdaSEG algorithm enjoys the linear speed-up property in the stochastic gradient variance term due to the periodic communication mechanism.

Experimental verification. We conduct several experiments on the stochastic bilinear game and the Wasserstein GAN (Arjovsky et al., 2017) to verify the efficiency and effectiveness of the LocalAdaSEG algorithm. We also extend the LocalAdaSEG algorithm to solve the challenging federated GANs in a heterogeneous setting. The experimental results agree with the theoretical guarantees and demonstrate the superiority of LocalAdaSEG against several existing minimax optimizers, such as SEGDA (Nemirovski, 2004), UMP (Bach & Levy, 2019), ASMP (Ene & Nguyen, 2020), LocalSEGDA (Beznosikov et al., 2021), LocalSGDA (Deng & Mahdavi, 2021), and Local Adam (Beznosikov et al., 2021).

2 Related work

Although there has been a lot of work on minimax optimization, due to space constraints, we summarize only the most closely related work. Our work is related to the literature on stochastic minimax algorithms, adaptive minimax algorithms, and distributed minimax algorithms. We defer a detailed discussion of related work to Section A in the appendix.

Our work and the proposed LocalAdaSEG contribute to the literature described above. To our knowledge, the proposed LocalAdaSEG algorithm is the first distributed communication-efficient algorithm for the stochastic minimax problem and simultaneously supports the adaptive learning rate and minibatch size. Moreover, LocalAdaSEG communicates only periodically to improve communication efficiency and uses a local adaptive learning rate, computed on local data in each worker, to improve the efficiency of computation. In addition, LocalAdaSEG can also be applied in a non-smooth setting with the convergence guarantee. LocalAdaSEG can be seen as a distributed extension of Bach and Levy (2019) with period communication as local SGD (Stich, 2019). We note that only very recently a local adaptive stochastic minimax algorithm, called Local Adam, has been used heuristically to train GANs without a convergence guarantee (Beznosikov et al., 2021). We summarize the relationship with the existing literature in Table 1.

3 Methodology

3.1 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 for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$F(x^*, y) \leq F(x^*, y^*) \leq F(x, y^*). \tag{2}$$

Under the assumptions stated in Sect. 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

Table 1 Comparison to related works on adaptive or communication-efficient approaches to stochastic minimax problems

Stochastic minimax algorithms	Nonsmooth ?	Comm. eff. ?	Adaptive ?
Mirror SA Nemirovski et al. (2009), SMP Juditsky et al. (2011), SAMP Chen et al. (2017), optimal stochastic PDHG-type Zhao (2021)	✓	×	×
SCAFFOLD-Catalyst-S Hou et al. (2021), local SGDA Deng and Mahdavi (2021), extra step local SGD Beznosikov et al. (2021)	×	✓	×
Universal mirror-prox Bach and Levy (2019), adaptive single-gradient mirror-prox Ene and Nguyen (2020), geometry-aware universal mirror-prox Babanezhad and Lacoste-Julien (2020), AdaProx Antonakopoulos et al. (2021)	✓	×	✓
Optimistic AdaGrad Liu et al. (2020)	×*	×	✓
Our LocalAdaSEG	✓	✓	✓

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 Liu et al. (2020) discusses non-convex non-concave minimax problems

$\mathcal{X}\times\mathcal{Y}$. We denote $\mathcal{Z}=\mathbb{X}\times\mathbb{Y}$, $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$, $z=(x,y)\in\mathcal{Z}$, 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 $\mathbb{X}$, $\mathbb{Y}$, $\mathcal{Z}$, 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

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

The duality gap is commonly used as a performance criterion for general convex-concave minimax problems (see, e.g., Nemirovski (2004); Lin et al. (2020)). Note that for all $z\in\mathcal{Z}$ it holds $\text{DualGap}(z)\geq 0$ and $\text{DualGap}(z)=0$ 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 1 *(Bounded domain)* There exists D such that $\sup_{z\in\mathcal{Z}}\frac{1}{2}\|z\|^2\leq D^2$.

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

Domain boundedness Assumption 1 is commonly assumed in the convex-concave minimax literature; see references in Sect. 1. However, we note that the assumption might be removed in certain settings. For example, (Chen et al., 2014; Monteiro & Svaiter, 2011) use a perturbation-based variant of the duality gap as the convergence criterion, and (Antonakopoulos et al., 2021) handles unbounded domains via the notion of local norms, while (Zhao, 2021) handles unbounded domains with access to a convex optimization oracle. The almost sure boundedness Assumption 2 on the gradient oracle seems restrictive but is common in the literature on adaptive stochastic gradient methods [(see, e.g., Duchi et al. (2011), Chen et al. (2019a), Bach and Levy (2019), Liu et al. (2020)]. In Remark 2 we discuss how to extend our analysis to unbounded oracles.

Assumption 3 *(Bounded variance)* There exists σ such that $\mathbb{E}_\xi\left[\|G(z)-\tilde{G}(z)\|_*^2\mid z\right]\leq\sigma^2$ for P-almost every z.

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

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

3.2 LocalAdaSEG **Algorithm**

We introduce the 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 $S := \{0, K, 2K, \ldots, RK\}$.
3: **for** $t = 1, \ldots, T = RK$, parallel for workers $m = 1, \ldots, M$ **do**
4: update learning rate $\eta_t^m =$

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

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

$$w_t^m = \frac{(\eta_t^m)^{-1}}{\sum_{m'=1}^M (\eta_t^{m'})^{-1}},\ \tilde{z}_{t-1}^\circ = \sum_{m=1}^M w_t^m \cdot \tilde{z}_{t-1}^m$$

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

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

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 12). The

adaptive learning rate is computed solely based on iterates occurred in the local worker (Line 4). Let $S := \{0, K, 2K, \ldots, RK = T\}$ denote the time points of communication. At a time of communication ($t \in S + 1$, Lines 5–8), the workers communicate and compute the weighted iterate, $\tilde{z}^{\circ}_{t-1}$, defined in Line 7. Then the next round begins with a common iterate $\tilde{z}^{\circ}_{t-1}$. 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 12, 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 12 reduces to

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

where $\Pi_{\mathcal{Z}}(z) = \operatorname{argmin}_{z'\in\mathcal{Z}}\|z - z'\|_2$ is the projection operator onto the compact set $\mathcal{Z}$. The above update is just the extragradient (EG) algorithm Korpelevich (1976) that is commonly used to solve minimax problems; see references in Sect. 1.

Periodic averaging weights The proposed weighted averaging scheme in Line 7 is different from existing works on local SGD and Local Adam (Beznosikov et al., 2021). 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 4) follows that of Bach and Levy Bach and Levy (2019) 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 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 (Bach & Levy, 2019; Antonakopoulos et al., 2021). For the smooth case, one might expect the sequence $\{\delta_t^m\}_t$ to be square-summable and $\eta_t^m \to \eta_\infty^m > 0$, 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 (Xie et al., 2019; Reddi et al., 2021; Chen et al., 2021). Very recently, (Beznosikov et al., 2021) 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. Following (Bach & Levy, 2019), we define

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

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 a different base learning rate, α, for smooth and nonsmooth problems; see Theorems 1 and 2, respectively.

3.3 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 Section C.1 and Section C.2 of the appendix. Recall the definition of γ in (5).

Theorem 1 (Nonsmooth Case) *Assume that Assumptions 1, 2, and 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 2 (Smooth case) *Assume that Assumptions 1, 2, 3, and 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 as*

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

Then

$$\mathbb{E}[\text{DualGap}(\bar{z})] = \tilde{O}\left(\frac{\sigma D}{\sqrt{MT}} + \frac{D\sqrt{M}\mathcal{V}_1(T)}{T} + \frac{\gamma^2 LD^2 M^{-1/2}}{T} + \frac{\gamma GD\sqrt{M}}{T}\right). \tag{7}$$

Remark 1 (The term $\mathcal{V}_1(T)$.) Note that by symmetry $\mathcal{V}_m(T) = \mathcal{V}_1(T)$ for all m. Although a trivial bound on $\mathcal{V}_1(T)$ is $\mathcal{V}_1(T) \le G\sqrt{2T}$, typically we have $\mathcal{V}_1(T) \ll \sqrt{T}$ in practice (Duchi et al., 2011; Reddi et al., 2018; Chen et al., 2019b, a; Liu et al., 2020), especially in the sparse data scenarios. For example, consider the bilinear saddle-point problem $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \{x^\top (\sum_{i=1}^n p_i M_i) y\}$, 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 5, 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,

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

for any $\epsilon \in (0, \frac{1}{2})$. Following the discussion in Chen et al. (2019b), Liu et al. (2020), 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 *(Extension to 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

$$\mathcal{E} := \left\{ \max_{z'\in\mathcal{Z}'} \|\tilde{G}(z') - G(z')\|_* \le G_{T,\delta} := \|G\|_\infty \cdot \left(\sqrt{2\log(4MT)} + \sqrt{2\log(2/\delta)}\right) \right\}.$$

Then $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$; see Appendix B.1. We can repeat the Proof of Theorems 1 and 2 on the event $\mathcal{E}$ and interpret our results with G replaced by $G_{T,\delta}$, which effectively substitutes G with $\|G\|_\infty$ at the cost of an extra $\log(T)$ factor.

Remark 3 *(Baseline 1: Minibatch EG)* We comment on the 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 (Bach & Levy, 2019; Ene & Nguyen, 2020) is, respectively,[1]

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

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 *(Baseline 2: EG on a single worker)* Another natural baseline is to run EG on a single worker for T iterations with batch-size 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, EG on a single worker 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 $\sigma = \Omega(\|G\|_\infty \sqrt{M})$.

Remark 5 (*On the choice of K*) Consider the baseline minibatch EG (see Remark 3) which runs as follows: the algorithm takes R extragradient steps, with each step using a minibatch of size KM, resulting in a procedure that communicates exactly R times. Note this procedure has exactly the same computation and communication structure as LocalAdaSEG, facilitating a fair comparison. In the non-smooth case, our theory shows that LocalAdaSEG dominates minibatch EG regardless of the choice K. Therefore, let us focus the discussion

[1] These bounds hold due to Theorem 4 of Ene and Nguyen (2020), 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. The claim 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.

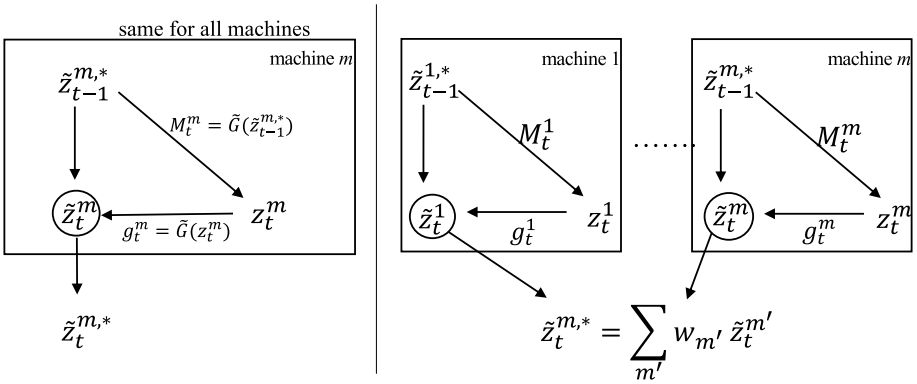

Fig. 2 The computation diagram for LocalAdaSEG. Left panel: computation on machine m when no communication ($t \notin S$). Right panel: computation on machine m when on communication round ($t \in S$)

on the *smooth loss with slow gradient growth case*. Suppose that the gradient growth term $\mathcal{V}_m(T) := \mathbb{E}\left[(\sum_{t=1}^T \|g_t^m\|_*^2 + \|M_t^m\|_*^2)^{1/2}\right]$ admits a rate $\mathcal{V}_m(T) = O(T^b)$ for some $0 < b < 1/2$. Theorem 2 then shows that LocalAdaSEG enjoys a convergence rate (ignoring problem parameters L, D and G)

$$\frac{1}{\sqrt{MKR}} + \frac{\sqrt{M}}{(KR)^{1-b}} + \frac{\sqrt{M}}{KR},$$

where M is the number of machines, R the communication rounds, and K is the length between two communications. The minibatch EG attains the convergence rate

$$\frac{1}{\sqrt{MKR}} + \frac{1}{R}.$$

Both algorithms achieve linear speedup, i.e., the dominant term is $O(\sigma/\sqrt{MKR})$. In order for LocalAdaSEG to be comparable with minibatch EG in the higher order term, we set $\sqrt{M}/(KR)^{1-b} = \Theta(1/R)$ and $\sqrt{M}/(KR) = O(1/R)$ and obtain $K = \Theta(\sqrt{MT^b})$. With this choice of K, LocalAdaSEG achieves a communication efficiency no worse than minibatch EG with the crucial advantage of being tuning-free. Compared with case of optimizing strongly-convex functions, local SGD needs $K = O(\sqrt{T})$ to achieve linear speedup (Stich, 2019). The discussion here is purely theoretical, since the exponent of gradient growth b is hard to estimate in practice.

Proof Sketch of Theorem 2 We present a proof sketch for the smooth case. Recall the update formula

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

Figure 2 provides a computation diagram and illustrates the relationship between the above variables.

We define the noise in the gradient operator G by

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

Moreover, we define a gradient-like quantity

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

If we ignore the projection operator in the update, the term (Z_t^m) will be of a similar scale as the gradients $\tilde{G}(z_t^m)$ and $\tilde{G}(\tilde{z}_t^m)$.

We begin with the following decomposition: for all $z \in \mathcal{Z}$,

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{m=1}^{M}\left\langle z_t^m - z, G(z_t^m)\right\rangle \\
&= \sum_{t=1}^{T}\sum_{m=1}^{M}\left\langle z_t^m - z, \xi_t^m\right\rangle + \sum_{t=1}^{T}\sum_{m=1}^{M}\left\langle z_t^m - z, g_t^m\right\rangle \\
&\le \underbrace{\sum_{t=1}^{T}\sum_{m=1}^{M}\left\langle z_t^m - z, \xi_t^m\right\rangle}_{I(z)} \\
&\quad + \underbrace{\sum_{t=1}^{T}\sum_{m=1}^{M}\frac{1}{\eta_t^m}\left(\tfrac{1}{2}\|z - \tilde{z}_{t-1}^{m,*}\|^2 - \tfrac{1}{2}\|z - \tilde{z}_t^m\|^2\right)}_{II(z)} \\
&\quad \underbrace{-\sum_{t=1}^{T}\sum_{m=1}^{M}\frac{1}{\eta_t^m}\left(\tfrac{1}{2}\|z_t^m - \tilde{z}_{t-1}^{m,*}\|^2 + \tfrac{1}{2}\|z_t^m - \tilde{z}_t^m\|^2\right)}_{III} \\
&\quad + \underbrace{\sum_{t=1}^{T}\sum_{m=1}^{M}\|g_t^m - M_t^m\|_* \cdot \|z_t^m - \tilde{z}_t^m\|}_{IV},
\end{aligned}$$

where we have used a descent lemma for EG updates common in the literature (Lemma 4 in our paper). The reason we care about the above quantity is that by the convexity-concavity of the problem, the duality gap metric can be upper-bounded by this term.

Next, we analyze each term separately. The term $I(z)$ characterizes the noise of the problem and eventually contributes to the noise term $\frac{\sigma}{\sqrt{KMR}}$. For the term *II* we use a telescoping argument and show that it can be upper bounded by $\sum_{m,t}\eta_t^m(Z_t^m)^2$. The telescoping argument can be applied due to the averaging weights w_t^m in the algorithm. The term *III* is negative. We keep the tail part of *III* which cancels the tail part of the term *IV*. For the term *IV* we use the smoothness property of the problem and show that it can be bounded by $\sum_{m,t}(\eta_t^m)^2(Z_t^m)^2$. Finally, two sums of the form $\sum_{m,t}\eta_t^m(Z_t^m)^2$ and $\sum_{m,t}(\eta_t^m)^2(Z_t^m)^2$ remain to be handled. For this we use the well-known basic inequality $\sum_{i=1}^n a_i/(a_0 + \sum_{j=1}^{i-1} a_j) = O(\log(1 + \sum_i a_i))$ and $\sum_{i=1}^n a_i/\sqrt{a_0 + \sum_{j=1}^{i-1} a_j} = \Theta(\sqrt{\sum_i a_i})$ for positive numbers a_i's.

Nonadaptive local algorithms rely on choosing a vanishing stepsize that is usually inversely proportional to a prespecified number of total iterations T. The freedom to choose the stepsize based on a prespecified T is crucial in the proofs of these algorithms and allows canceling of the asynchronicity of updates caused by local updates and the bias in those updates caused by data heterogeneity. This is the case for both convex optimization and convex-concave optimization. However, in the adaptive algorithm regimes, such a proof technique is clearly not viable.

Our algorithm requires a carefully designed iterates averaging scheme, with weight inversely proportional to stepsize. Such averaging-scheme is designed to account for the asynchronicity of local iterates and is automatically determined by the optimization process. This is what enables the extension of an Adam-type stepsize to parallel settings, which is highly nontrivial. □

4 Experiments

We apply LocalAdaSEG to the stochastic bilinear minimax problem introduced in Gidel et al. (2019), Beznosikov et al. (2021) and train the Wasserstein generative adversarial neural network (Wasserstein GAN) (Arjovsky et al., 2017). For the homogeneous setting, to demonstrate the efficiency of our proposed algorithm, we compare LocalAdaSEG with minibatch stochastic extragradient gradient descent (MB-SEGDA) (Nemirovski, 2004), minibatch universal mirror-prox (MB-UMP) (Bach & Levy, 2019), minibatch adaptive single-gradient mirror-Prox (MB-ASMP) (Ene & Nguyen, 2020), extra step local SGD (LocalSEGDA) (Beznosikov et al., 2021), and local stochastic gradient descent ascent (LocalSGDA) (Deng & Mahdavi, 2021). We further extend the proposed LocalAdaSEG algorithm to solve federated WGANs with a heterogeneous dataset to verify its efficiency. To validate the practicality of LocalAdaSEG, we also train the BigGAN (Brock et al., 2019) over CIFAR10 dataset under the heterogeneous setting. In this setting, we also compare LocalAdaSEG with Local Adam (Beznosikov et al., 2021). We emphasize here that whether Local Adam converges is still an open question, even for the stochastic convex-concave setting.

4.1 Stochastic bilinear minimax problem

We consider the stochastic bilinear minimax problem with box constraints

$$\min_{x\in C^n}\max_{y\in C^n} F(x,y) \tag{8}$$

where

$$F(x,y) := \mathbb{E}_{\xi\sim P}\left[x^\top A y + (b+\xi)^\top x + (c+\xi)^\top y\right],$$

Here $C^n = [-1,1]^n$ is a box in $\mathbb{R}^n$, the tuple (A, b, c) is deterministic, and the perturbation variable ξ follows the normal distribution with variance σ. We define the KKT residual $\mathrm{Res}(x,y)$ as:

$$\begin{aligned}\mathrm{Res}(x,y)^2 :=& \|x - \Pi_{C^n}(x-(Ay+b))\|^2 \\ &+ \|y - \Pi_{C^n}(y+(Ax+c))\|^2.\end{aligned}$$

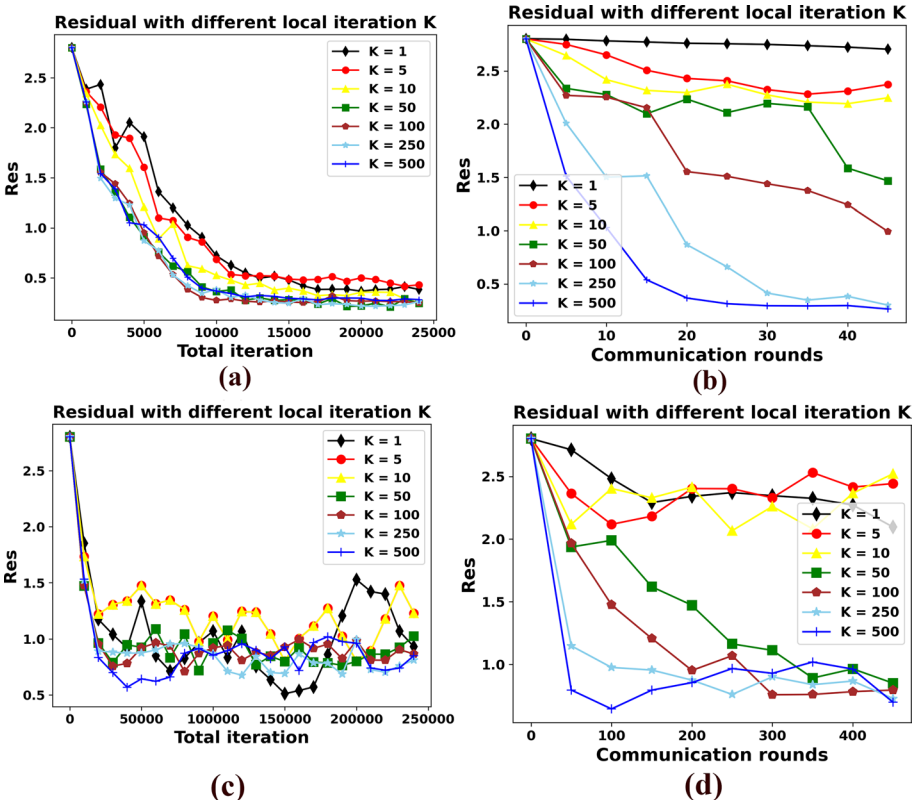

Fig. 3 Subfigures **a**, **b** and **c**, **d** plot the residual of LocalAdaSEG against the total number of iterations T and communications R, with varying numbers of local iterations K. We also investigate the effect of noise level ($\sigma = 0.1$ in (**a**), (**b**) and $\sigma = 0.5$ in (**c**), (**d**))

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 set of saddle-points of the bilinear minimax problem (8). During experiments, we use $\mathrm{Res}(x, y)$ to measure the quality of the approximate 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 $A = \bar{A}/\max\left(|b|_{\max}, |c|_{\max}\right)$, 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 locally with an optimization algorithm. 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 Fig. 3. A larger σ indicates more noise in the stochastic gradients, making problem (8) harder. Furthermore, we compare LocalAdaSEG by setting the local iteration $K = 50$ against several existing optimizers, illustrated in Fig. 4.

Experimental results In Fig. 3, LocalAdaSEG provides stable convergence results under different configurations of local iterations K and noise levels σ. Figure (b)(d) illustrates that a suitably large K could accelerate the convergence speed of LocalAdaSEG.

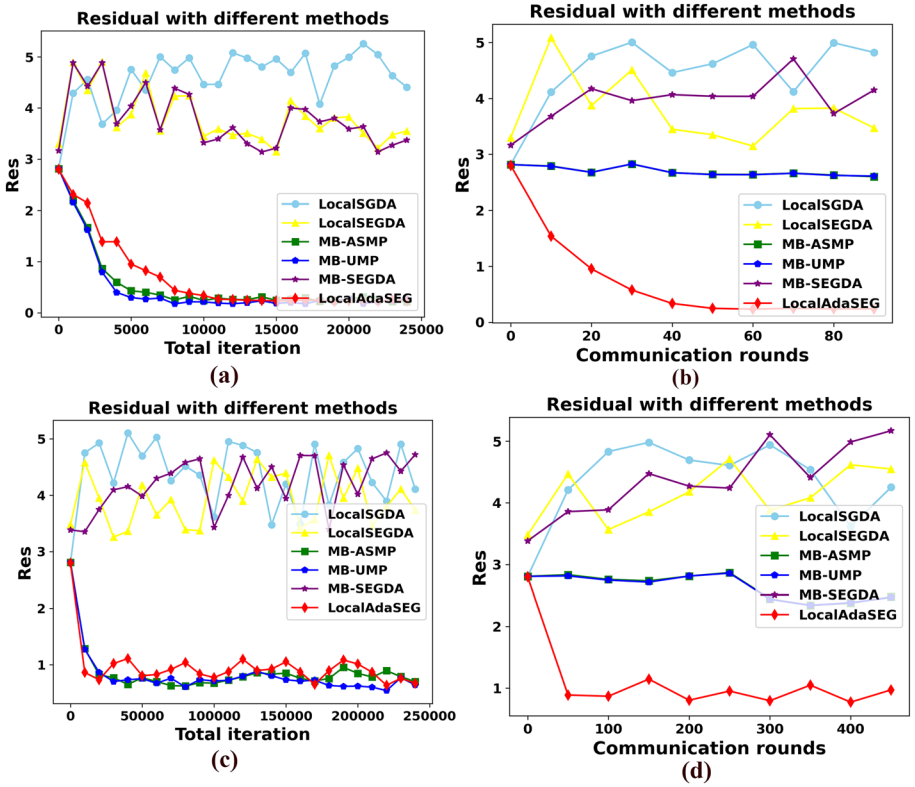

Fig. 4 Subfigures **a**, **b** and **c**, **d** compare LocalAdaSEG with existing optimizers. We plot the residuals against the total number of iterations T and communications R with different noise levels ($\sigma = 0.1$ in (**a**), (**b**) and $\sigma = 0.5$ in (**c**), (**d**)

Figure (a)(c) illustrates that a large variance would result in unstable optimization trajectories. The findings of the experiment agree with our theoretical predictions: (i) a larger $T = KR$ improves convergence; (ii) the variance term dominates the convergence rate of LocalAdaSEG; a large variance term will slow down LocalAdaSEG. In Fig. 4, (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. Furthermore, when compared in terms of communication rounds in (b)(d), LocalAdaSEG converges faster than other distributed stochastic minimax optimizers, demonstrating the superiority of LocalAdaSEG.

To validate the performance of our proposed method, we conduct the comparison of the asynchronous case and the synchronous case of LocalAdaSEG for the stochastic bilinear minimax problem. We also compare asynchronous and synchronous cases with the single-thread version (SEGDA with MKR iterations) from the aspects of residual and wallclock time. Finally, we evaluate the quantity of V_t with the update t. The experimental details are described in Appendix E.1. As can be seen in Fig. E1 (in Appendix E.1), compared with synchronous cases, asynchronicity only affects the convergence rate that is slower than the synchronous version with respect to the communication

rounds. Compared to SEGDA of MKR iterations, our proposed LocalAdaSEG can achieve more stable and better performance. Regarding the quantity of Vt, it is really much smaller than the dominant variance term.

4.2 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. The description of the problem and implementation details are placed in Section E.2.

Experimental results Figure E2 and E3 (in Section E.2) compare MB-UMP, MB-ASMP, LocalAdam and LocalAdaSEG in a homogeneous and heterogeneous setting, respectively. In Figs. E2a and E3a, 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 Figs. E2b and E3b, LocalAdaSEG and Local Adam converge faster than other optimizers and reach a satisfactory solution in just a few rounds. In Figs. E2c and E3c, all the listed optimizers achieve a high IS. In particular, the IS of LocalAdaSEG and Local Adam increases much faster with less communication than MB-UMP, MB-ASMP as shown in Figs. E2d and E3d.

In Figs. E4 and E5, we show and compare the FID and IS of LocalAdaSEG with other optimizers under different data distributions. As can be seen from Fig. E4, LocalAdaSEG converges faster when the Dirichlet distribution parameter α decreases. In Fig. E5, when data distribution changes, our LocalAdaSEG can still converge faster than other existing optimizers.

4.3 BigGAN

To validate the practicability of our proposed LocalAdaSEG method, we apply LocaAdaSEG to train the large-scale BigGAN (Brock et al., 2019) model over the CIFAR10 dataset. The description of BigGAN and parameter setup are placed in Section E.3.

Experimental results Figure E6 illustrates the comparison of the FID and IS against communication rounds by using LocalAdaSEG and existing optimizers. As can be seen from Fig. E6a, LocalAdaSEG and Local Adam can reach a satisfactory FID value in a few rounds. Similarly, from Fig. E6b, we can see that the IS value of LocalAdaSEG and Local Adam is much higher than that of MB-UMP and MB-ASMP. In a word, the FID and IS values of LocalAdaSEG and Local Adam converge much faster than that of other optimizers.

Additional discussions To end this section, we briefly discuss the limitation of current work.

Theoretical limitations. Our theory is applicable to the homogeneous setting, meaning each worker has access to data from one distribution. However, in practice, data heterogeneity is a main factor practitioners must take into account for distributed learning. We briefly discuss technical challenges here. For the heterogeneous case, the theory for *non-adaptive* algorithms relies on choosing a very small stepsize, usually inverse proportional to a prespecified number of total iterations T. The freedom to choose the stepsize based on a prespecified T is crucial in those proofs and enables canceling the bias caused by local updates, a.k.a. client drifts. The same situation also occurs in the convex optimization case.

However, our goal is to have an adaptive algorithm that does not depend on the problem parameters or a prespecified T. For this reason, we leave such an important open question for future work.

Experimental limitations. In the scale of the dataset, we experimented with should be increased to showcase the computation benefit of the proposed algorithm. At the current stage we have experimented with MNIST data and further, add CIFAR 10 experiments after reviewers' suggestions. Application to other ultra-large datasets such as ImageNet requires significant engineering efforts and will be left for future investigation. We should emphasize that our paper mainly contributes to the theoretical understanding of adaptive algorithms in distributed settings.

5 Conclusion

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.

For future work, since that the current analysis merely holds for the homogeneous setting, a promising direction is to extend the theoretical result of LocalAdaSEG to the heterogeneous setting that better models various real-world applications, such as federated GANs (Beznosikov et al., 2021) and robust federated learning (Deng et al., 2020). 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.

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

Editorial Policies for:

Springer journals and proceedings: https://www.springer.com/gp/editorial-policies

Nature Portfolio journals: https://www.nature.com/nature-research/editorial-policies

Scientific Reports: https://www.nature.com/srep/journal-policies/editorial-policies

BMC journals: https://www.biomedcentral.com/getpublished/editorial-policies

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10994-022-06239-z.

Author contributions All authors contributed to the study conception and design. The first draft of the manuscript was written by Luofeng Liao, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work is supported by the Major Science and Technology Innovation 2030 "Brain Science and Brain-like Research" key project (No. 2021ZD0201405).

Data availability The data used in this work is all public.

Code availability The codes of the proposed method will be released after publishing.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval Not Applicable.

Consent to participate Not Applicable.

Consent for publication Not Applicable.

References

Antonakopoulos, K., Belmega, V., & Mertikopoulos, P. (2021). Adaptive extra-gradient methods for min-max optimization and games. In *International conference on learning representations.*

Arjovsky, M. & Bottou, L. (2017). Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In D. Precup & Y. W. Teh, (Eds.) *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, (pp. 214–223). PMLR.

Azizian, W., Mitliagkas, I., Lacoste-Julien, S., & Gidel, G. (2020). A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International conference on artificial intelligence and statistics,* (pp. 2863–2873). PMLR.

Babanezhad, R. & Lacoste-Julien, S. (2020). Geometry-aware universal mirror-prox. arXiv preprint arXiv:2011.11203.

Bach, F. & Levy, K. Y. (2019). A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on learning theory,* (pp. 164–194). PMLR.

Beznosikov, A., Samokhin, V., & Gasnikov, A. (2021). Distributed saddle-point problems: Lower bounds, optimal algorithms and federated gans. arXiv preprint arXiv:2010.13112.

Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale gan training for high fidelity natural image synthesis. In *International conference on learning representations.*

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

Chatterjee, S. (2014). *Superconcentration and Related Topics*. Springer International Publishing.

Chen, T., Guo, Z., Sun, Y., & Yin, W. (2021). Cada: Communication-adaptive distributed adam. In *International Conference on Artificial Intelligence and Statistics,* (pp. 613–621). PMLR.

Chen, X., Liu, S., Sun, R., & Hong, M. (2019). On the convergence of a class of adam-type algorithms for non-convex optimization. In *International conference on learning representations.*

Chen, C., Shen, L., Zou, F., & Liu, W. (2021). Towards practical adam: Non-convexity, convergence theory, and mini-batch acceleration. arXiv preprint arXiv:2101.05471.

Chen, X., Yang, S., Shen, L., & Pang, X. (2020). A distributed training algorithm of generative adversarial networks with quantized gradients. arXiv preprint arXiv:2010.13359.

Chen, Z., Yuan, Z., Yi, J., Zhou, B., Chen, E., & Yang, T. (2019). Universal stagewise learning for non-convex problems with convergence on averaged solutions. In *International conference on learning representations.*

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

Chen, Y., Lan, G., & Ouyang, Y. (2017). Accelerated schemes for a class of variational inequalities. *Mathematical Programming, 165*(1), 113–149.

Chen, C., Shen, L., Huang, H., & Liu, W. (2021). Quantized adam with error feedback. *ACM Transactions on Intelligent Systems and Technology (TIST), 12*(5), 1–26.

Daskalakis, C., Ilyas, A., Syrgkanis, V., & Zeng, H. (2018). Training GANs with optimism. In *International conference on learning representations.*

Delage, E., & Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research, 58*(3), 595–612.

Deng, Y., & Mahdavi, M. (Apr 2021). 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,* (pp. 1387–1395). PMLR, 13–15.

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

Deng, Y., Kamani, M. M., & Mahdavi, M. (2020). Distributionally robust federated averaging. In *Advances in neural information processing systems,* (vol 33, pp. 15111–15122). Curran Associates, Inc., .

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(7).

Ene, A., & Nguyen, H. L. (2020). Adaptive and universal single-gradient algorithms for variational inequalities. arXiv preprint arXiv:2010.07799.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., & Lacoste-Julien, S. (2019). A variational inequality perspective on generative adversarial networks. In *International conference on learning representations.*

Gidel, G., Hemmat, R. A., Pezeshki, M., Le Priol, R., Huang, G., Lacoste-Julien, S., & Mitliagkas, I. (2019). Negative momentum for improved game dynamics. In *The 22nd international conference on artificial intelligence and statistics,* (pp. 1802–1811). PMLR.

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

Guo, Z., Liu, M., Yuan, Z., Shen, L., Liu, W., & Yang, T. (2020). Communication-efficient distributed stochastic auc maximization with deep neural networks. In *International conference on machine learning,* (pp. 3864–3874). PMLR.

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

Hou, C., Thekumparampil, K. K., Fanti, G., & Oh, S. (2021). Efficient algorithms for federated saddle point optimization.

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

Juditsky, A., Nemirovski, A., & Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems, 1*(1), 17–58.

Kingma, D. P. & Adam, J. B. (2017). A method for stochastic optimization. In *International conference on learning representations.*

Korpelevich, G. M. (1976). The extragradient method for finding saddle points and other problems. Matecon.

Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2020). On the convergence of fedavg on non-iid data. In *International conference on learning representations*.

Lin, T., Jin, C., & Jordan, M. I. (2020). Near-optimal algorithms for minimax optimization. In Conference on learning theory, (pp. 2738–2779). PMLR.

Lin, T., Stich, S. U., Patel, K. K., & Jaggi, M. (2020). Don't use large mini-batches, use local sgd. In *International conference on learning representations*.

Liu, M., Mroueh, Y., Ross, J., Zhang, W., Cui, X., Das, P., & Yang, T. (2020). Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *International conference on learning representations*.

Liu, M., Zhang, W., Mroueh, Y., Cui, X., Ross, J., Yang, T., & Das, P. (2020). A decentralized parallel algorithm for training generative adversarial nets. (vol 33).

McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R. and D'Oliveira, R. G. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, *14*(1).

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C. -S., Chandrasekhar, V., & Piliouras, G. (2019). Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International conference on learning representations.*

Mertikopoulos, P., Papadimitriou, C., & Piliouras, G. (2018). Cycles in adversarial regularized learning. In *Proceedings of the twenty-ninth annual ACM-SIAM symposium on discrete algorithms*, (pp. 2703–2717). SIAM.

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

Nemirovski, A. (2004). 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.

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

Neumann, Jv. (1928). Zur theorie der gesellschaftsspiele. *Mathematische Annalen, 100*(1), 295–320.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., & McMahan, H. B. (2021). Adaptive federated optimization. In *International conference on learning representations.*

Reddi, S. J., Kale, S., & Kumar, S. (2018). On the convergence of adam and beyond. In *International conference on learning representations*.

Rogozin, A., Beznosikov, A., Dvinskikh, D., Kovalev, D., Dvurechensky, P., & Gasnikov, A. (2021). Decentralized distributed optimization for saddle point problems.

Smola, A., & Narayanamurthy, S. (2010). An architecture for parallel topic models. *Proceedings of the VLDB Endowment, 3*(1–2), 703–710.

Stich, S. U. (2019). Local SGD converges fast and communicates little. In *International conference on learning representations.*

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

Xie, C., Koyejo, O., Gupta, I., & Lin, H. (2019). Local adaalter: Communication-efficient stochastic gradient descent with adaptive learning rates. arXiv preprint arXiv:1911.09030.

Yan, Y., & Xu, Y. (2020). Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs. arXiv preprint arXiv:2012.14943.

Yu, H., Jin, R., & Yang, S. (Jun 2019). 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,* (pp. 7184–7193). PMLR, 09–15.

Zhang, J., Xiao, P., Sun, R., & Luo, Z. (2020). A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. In *Advances in neural information processing systems,* (vol 33, pp. 7377–7389). Curran Associates, Inc., .

Zhao, R. (2021). Accelerated stochastic algorithms for convex-concave saddle-point problems. arXiv preprint arXiv:1903.01687.

Zou, F., Shen, L., Jie, Z., Zhang, W., & Liu, W. (2019). A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* (pp. 11127–11135).