An Efficient Stochastic Algorithm for Decentralized Nonconvex-Strongly-Concave Minimax Optimization

Lesi Chen[1] **Haishan Ye**[2] **Luo Luo**[3,*]

[1]Institute for Interdisciplinary Information Sciences, Tsinghua Univerisity
[2]School of Management, Xi'an Jiaotong University; SGIT AI Lab, State Grid Corporation of China
[3]School of Data Science, Fudan University; Shanghai Key Laboratory for Contemporary Applied Mathematics

Abstract

This paper studies the stochastic nonconvex-strongly-concave minimax optimization over a multi-agent network. We propose an efficient algorithm, called Decentralized Recursive gradient descEnt Ascent Method (`DREAM`), which achieves the best-known theoretical guarantee for finding the ϵ-stationary points. Concretely, it requires $\mathcal{O}(\min(\kappa^3\epsilon^{-3}, \kappa^2\sqrt{N}\epsilon^{-2}))$ stochastic first-order oracle (SFO) calls and $\tilde{\mathcal{O}}(\kappa^2\epsilon^{-2})$ communication rounds, where κ is the condition number and N is the total number of individual functions. Our numerical experiments also validate the superiority of `DREAM` over previous methods.

1 Introduction

This paper studies the decentralized minimax optimization problem, where m agents in a network collaborate to solve the problem

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

We suppose that $f(x,y)$ is μ-strongly-concave in y; $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ is closed and convex; each local function on the i-th agent has the following stochastic form

$$f_i(x,y) \triangleq \mathbb{E}[F_i(x,y;\xi_i)]; \tag{2}$$

and the stochastic component $F_i(x,y;\xi_i)$ indexed by the random variable ξ_i is L-average smooth. The nonconvex-strongly-concave minimax problem (1) plays an important role in many machine learning applications, such as adversarial training (Farnia and Ozdaglar, 2021; Nagarajan and Kolter, 2017), distributional robust optimization (Jin et al., 2021; Levy et al., 2020; Sinha et al., 2018), AUC maximization (Guo et al., 2023; Liu et al., 2019; Yuan et al., 2021), reinforcement learning (Jin and Sidford, 2020; Qiu et al., 2020; Wai et al., 2018; Zhang et al., 2019), learning with non-decomposable loss (Fan et al., 2017; Rafique et al., 2021) and so on. Following existing non-asymptotic analysis for nonconvex-strongly-concave minimax optimization problems (Lin et al., 2020a,b; Luo et al., 2020), we focus on the task of finding an ϵ-stationary point of the primal function $P(x) \triangleq \max_{y\in\mathcal{Y}} f(x,y)$.

In this paper, we also consider a popular special case of problem (1) when each random variable ξ_i is finitely sampled from $\{\xi_{i,1},\dots,\xi_{i,n}\}$. That is, we can write the local function as

$$f_i(x,y) \triangleq \frac{1}{n}\sum_{j=1}^{n} F_{ij}(x,y). \tag{3}$$

We refer to the general form (2) as the online (stochastic) case and refer to the special case (3) as the offline (finite-sum) case. We define $N = mn$ as the total number of individual functions for the offline case.

Nonconvex-strongly-concave minimax optimization has received increasing attentions in recent years (Chen et al., 2023; Lin et al., 2020a,b; Luo et al., 2020, 2022; Nagarajan and Kolter, 2017; Xu et al., 2020; Zhang et al., 2020, 2021a, 2022). In the scenario of single machine, Lin et al. (2020a) showed the Stochastic Gradient Descent Ascent (`SGDA`) requires $\mathcal{O}(\kappa^3\epsilon^{-4})$ SFO complexity to find an ϵ-stationary point of $P(x)$, where the condition number is defined by $\kappa \triangleq L/\mu$. Luo et al. (2020) proposed the Stochastic Recursive gradiEnt Descent Ascent (`SREDA`), which uses the variance reduction technique of Stochastic Path Integrated Differential Estimator (`SPIDER`) (Fang et al., 2018) to establish a SFO complexity of $\mathcal{O}(\min(\kappa^3\epsilon^{-3}, \sqrt{n}\kappa^2\epsilon^{-2}))$.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. *Corresponding Author

It is worth noting that the $\mathcal{O}(\min(\epsilon^{-3}, \sqrt{n}\epsilon^{-2}))$ dependency on ϵ and n matches the lower bound of stochastic nonconvex optimization under the average-smooth assumption (Arjevani et al., 2023; Fang et al., 2018).

Distributed optimization is a popular setting for training large-scale machine learning models. It allows all agents on a given network to collaboratively optimize the global objective. In the decentralized scenario, each agent on the network only communicates with its neighbors. The decentralized training fashion avoids the communication traffic jam on the central node (Lu and De Sa, 2021). There have been a lot of works focusing on the complexity of decentralized stochastic minimization problems (Chen et al., 2021; Hendrikx et al., 2021; Kovalev et al., 2020; Li et al., 2020, 2022a,b; Shi et al., 2015; Sun et al., 2020; Uribe et al., 2020; Wang et al., 2021; Xin et al., 2022; Yuan et al., 2016).

However, the understanding of the complexity of first-order methods decentralized nonconvex-strongly-concave minimax problems is still limited. For offline decentralized nonconvex-strongly-concave minimax problems, Tsaknakis et al. (2020) proposed the Gradient-Tracking Descent-Ascent (`GT-DA`), which combines multi-step gradient descent ascent gradient (Nouiehed et al., 2019) with gradient tracking (Nedic et al., 2017; Qu and Li, 2019). `GT-DA` can provably find an ϵ-stationary point of $P(x)$ within $\tilde{\mathcal{O}}(N\epsilon^{-2})$ SFO calls and $\tilde{\mathcal{O}}(\epsilon^{-2})$ communication rounds. Zhang et al. (2021b) proposed the Gradient-Tracking Gradient Descent Ascent (`GT-GDA`) which updates both variable x and y simultaneously. The removal of the inner loop with respect to y also leads to the removal of the additional $\mathcal{O}(\log(1/\epsilon))$ factor in the complexity, and `GT-DA` can provably find an ϵ-stationary point of $P(x)$ within $\mathcal{O}(N\epsilon^{-2})$ SFO calls and $\mathcal{O}(\epsilon^{-2})$ communication rounds. However, both `GT-DA` and `GT-GDA` access the exact local gradients on each agent, which may be quite expensive when n is very large. To reduce the computation complexity, Zhang et al. (2021b) further proposed the Gradient Tracking-Stochastic Recursive Variance Reduction (`GT-SRVR`) by combining `GT-GDA` with `SPIDER` (Fang et al., 2018). `GT-SRVR` requires $\mathcal{O}(N + \sqrt{mN}\epsilon^{-2})$ SFO complexity to find an ϵ-stationary point of $P(x)$, outperforming `GT-DA` and `GT-GDA` when $n \gtrsim m$. However, for fixed N (total number of individual functions), the SFO upper bound of `GT-SRVR` will increase with the increase of m (number of agents). This trend seems somewhat unreasonable since involving more agents in the computation intuitively should not result in a higher overall computational cost.

For the online case, Xian et al. (2021) introduced the variance reduction technique of STOchastic Recursive Momentum (`STORM`) (Cutkosky and Orabona, 2019) and proposed the Decentralized Minimax Hybrid Stochastic Gradient Descent (`DM-HSGD`) with an upper complexity bound of $\mathcal{O}(\kappa^3\epsilon^{-3})$ for both SFO calls and communication rounds. For this online case, The SFO complexity of `DM-HSGD` recovers the result of `SREDA` (Luo et al., 2020) in the scenario of single-machine (when $m = 1$). Although the SFO complexity of `DM-HSGD` is better than those for `GT-GDA` and `GT-SRVR` when N has a higher order of magnitude compared to ϵ^{-1}, the communication complexity of $\mathcal{O}(\epsilon^{-3})$ in `DM-HSGD` is worse than the communication complexity of $\mathcal{O}(\epsilon^{-2})$ in `GT-GDA` and `GT-SRVR`. This implies `DM-HSGD` may has no advantage when the bottleneck is the cost of communication.

In many applications, the inner variable y in minimax problem (1) is often subject to some constraints, meaning that $\mathcal{Y}$ typically represents a given convex set endowed with a specific model. For instance, in adversarial training (Goodfellow et al., 2014), the variable y represents perturbations for the input, which typically lie within the box $\mathcal{Y} = \{y \in \mathbb{R}^{d_y} : \|y\|_\infty \le c\}$ for some positive constant c. In distributionally robust optimization (Yan et al., 2019), the variable y represents the probability distribution in the simplex $\mathcal{Y} = \{y \in \mathbb{R}^{d_y} : \sum_k y_k = 1, y_k \ge 0\}$. Dealing with the constraint on variable y typically requires additional steps like projection, which lead to extra consensus error in the decentralized setting. However, existing variance-reduced methods for decentralized nonconvex-strongly-concave minimax problems including `GT-SRVR` and `DM-HSGD` only successfully deal with the unconstrained case.

In this paper, we propose a novel method called Decentralized Recursive-gradient dEscent Ascent Method (`DREAM`) for solving decentralized nonconvex-strongly-concave minimax problems. We provide a unified convergence analysis for both the online and offline setups. Our analysis indicates that the proposed `DREAM` achieves the best-known complexity guarantee of both cases. We summarize the advantage of `DREAM` as follows.

- For the offline case, `DREAM` achieves the SFO complexity of $\mathcal{O}(N + \sqrt{N}\kappa^2\epsilon^{-2})$ and the communication complexity of $\tilde{\mathcal{O}}(\kappa^2\epsilon^{-2})$. The SFO complexity of `DREAM` is strictly better than existing methods, and achieves a linear speed-up with respect to the number of agents m, which is better than `GT-SRVR`. The communication complexity of `DREAM` remains state-of-the-art, matching those of `GT-DA`/`GT-GDA`/`GT-SRVR` in the dependency of ϵ, up to logarithmic factors.
- For the online case, `DREAM` achieves the $\mathcal{O}(\kappa^3\epsilon^{-3})$ SFO complexity and the $\tilde{\mathcal{O}}(\kappa^2\epsilon^{-2})$ communication complexity. The SFO complexity of `DREAM` is optimal in the dependency of ϵ (Arjevani et al.,

2023). And the communication complexity of `DREAM` strictly improves that of `DM-HSGD` in the dependency of both κ and ϵ.

- Moreover, `DREAM` is capable of working in both unconstrained and constrained scenarios, which gives it a wider range of applicability than previous variance-reduced methods `GT-SRVR` and `DM-HSGD`.

We compare our theoretical results with previous work in Table 1.

Notations. Throughout this paper, we denote $\|\cdot\|$ as the Frobenius norm of a matrix or the Euclidean norm of a vector. We use $I_m \in \mathbb{R}^{m\times m}$ to present a m by m identity matrix, $O_m \in \mathbb{R}^{m\times m}$ to present a m by m zero matrix, and let $\mathbf{1} = [1, \cdots, 1]^\top \in \mathbb{R}^m$. We define aggregated variables $\mathbf{x} \in \mathbb{R}^{m\times d_x}$, $\mathbf{y} \in \mathbb{R}^{m\times d_y}$ and $\mathbf{z} \in \mathbb{R}^{m\times(d_x+d_y)}$ for all agents as

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}(1) \\ \vdots \\ \mathbf{x}(m) \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}(1) \\ \vdots \\ \mathbf{y}(m) \end{bmatrix} \quad \text{and} \quad \mathbf{z} = \begin{bmatrix} \mathbf{z}(1) \\ \vdots \\ \mathbf{z}(m) \end{bmatrix},$$

where row vectors $\mathbf{x}(i) \in \mathbb{R}^{d_x}$ and $\mathbf{y}(i) \in \mathbb{R}^{d_y}$ are local variables on the i-th agent; and we also denote $\mathbf{z}(i) = [\mathbf{x}(i); \mathbf{y}(i)] \in \mathbb{R}^d$ with $d = d_x + d_y$. We use the lowercase with the bar to represent mean vector, e.g.,

$$\bar{x} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}(i), \;\; \bar{y} = \frac{1}{m}\sum_{i=1}^m \mathbf{y}(i) \text{ and } \bar{z} = \frac{1}{m}\sum_{i=1}^m \mathbf{z}(i).$$

Similarly, we also introduce the aggregated gradient as

$$\nabla \mathbf{f}(\mathbf{z}) = \begin{bmatrix} \nabla f_1(\mathbf{x}(1), \mathbf{y}(1))^\top \\ \vdots \\ \nabla f_m(\mathbf{x}(m), \mathbf{y}(m))^\top \end{bmatrix} \in \mathbb{R}^{m\times d}.$$

We use $\mathbb{I}[\,\cdot\,]$ to represent the indicator function of an event and define $n = +\infty$ for the online case.

2 Assumptions and Preliminaries

Throughout this paper, we suppose the stochastic NC-SC decentralized optimization problem (1) satisfies the following standard assumptions.

Assumption 2.1. *We suppose $P(x) \triangleq \max_{y\in\mathcal{Y}} f(x, y)$ is lower bounded. That is, we have*

$$P^* = \inf_{x\in\mathbb{R}^{d_x}} P(x) > -\infty.$$

Assumption 2.2. *We suppose the stochastic component functions $F(x, y; \xi_i)$ on each agent is L-average smooth for some $L > 0$. That is, we have*

$$\mathbb{E}\left\|\nabla F_i(x, y; \xi_i) - \nabla F_i(x', y'; \xi_i)\right\|^2 \le L^2\big(\left\|x - x'\right\|^2 + \left\|y - y'\right\|^2\big)$$

for any $(x, y), (x', y') \in \mathbb{R}^{d_x\times d_y}$ and random index ξ_i.

Algorithm 1 `FastMix`$(\mathbf{a}^{(0)}, K)$

1: **Initialize:** $\mathbf{a}^{(-1)} = \mathbf{a}^{(0)}$
2: $\eta_a = 1/\big(1 + \sqrt{1 - \lambda_2^2(W)}\big)$
3: **for** $k = 0, 1, \dots, K$ **do**
4: $\quad \mathbf{a}^{(k+1)} = (1 + \eta_a)W\mathbf{a}^{(k)} - \eta_a\mathbf{a}^{(k-1)}$
5: **end for**
6: **Output:** $\mathbf{a}^{(K)}$

Assumption 2.3. *We suppose each local function $f_i(x, y)$ is μ-strongly-concave in y. That is, there exists some constant $\mu > 0$ such that we have*

$$f_i(x, y) \le f_i(x, y') + \nabla_y f_i(x, y')^\top(y - y') - \frac{\mu}{2}\|y - y'\|^2$$

for any $x \in \mathbb{R}^{d_x}$ and $y, y' \in \mathbb{R}^{d_y}$.

Based on the smoothness and strong concavity assumptions, we can define the condition number of our optimization problem as follows.

Definition 2.1. *We define $\kappa \triangleq L/\mu$ as the condition number of problem (1), where L and μ are defined in Assumption 2.2 and 2.3 respectively.*

The differentiability of the primal function $P(x)$ can be proved by Danskin's theorem (Lin et al., 2020a, Lemma 4.3).

Proposition 2.1. *Under Assumptions 2.2 and 2.3, the function $P(x)$ is L_P-smooth with $L_P \triangleq (\kappa + 1)L$ and its gradient can be written as $\nabla P(x) = \nabla_x f(x, y^*(x))$, where we define $y^*(x) \triangleq \arg\max_{y\in\mathcal{Y}} f(x, y)$.*

The differentiability of $P(x)$ allows us to define the ϵ-stationary point.

Definition 2.2. *We call $\hat{x}$ an ϵ-stationary point if it holds that $\|\nabla P(\hat{x})\| \le \epsilon$.*

The goal of algorithms is to find an ϵ stationary point of $P(x)$. In the setting of stochastic optimization, we suppose an algorithm can get access to the stochastic first-order oracle that satisfies the following assumptions. Note that the bounded variance assumption is only required for the online case.

Assumption 2.4. *We suppose each stochastic first-order oracle (SFO) $\nabla F_i(x, y; \xi_i)$ is unbiased and has bounded variance. That is, we have*

$$\mathbb{E}[\nabla F_i(x, y; \xi_i)] = \nabla f_i(x, y)$$

and

$$\mathbb{E}\left\|\nabla f_i(x, y) - \nabla F_i(x, y; \xi_i)\right\|^2 \le \sigma^2$$

with $\sigma^2 < +\infty$.

Table 1: We compare the theoretical results of `DREAM` with previous methods for decentralized nonconvex-strongly-concave minimax optimization for both the offline and online settings. Notations κ^p and κ^q are used when the polynomial dependency on κ is not explicitly provided (Tsaknakis et al., 2020; Zhang et al., 2021b). The notation $\tilde{\mathcal{O}}(\,\cdot\,)$ hides logarithmic factors in complexity. Note that `GT-GDA` and `GT-SRVR` only consider the unconstrained problem, which corresponds to the specific case in our setting where $\mathcal{Y} = \mathbb{R}^{d_y}$. The design of `DM-HSGD` includes the general constrained setting, but its convergence analysis for the constrained case looks problematic and we provide more detailed discussions in Appendix B.

Setup	Algorithm	#SFO	#Communication	Constraint
Offline	`GT-DA` Tsaknakis et al. (2020)	$\tilde{\mathcal{O}}\left(N\kappa^p\epsilon^{-2}\right)$	$\tilde{\mathcal{O}}\left(\kappa^q\epsilon^{-2}\right)$	✓
	`GT-GDA` Zhang et al. (2021b)	$\mathcal{O}\left(N\kappa^p\epsilon^{-2}\right)$	$\mathcal{O}\left(\kappa^q\epsilon^{-2}\right)$	✗
	`GT-SRVR` Zhang et al. (2021b)	$\mathcal{O}\left(N+\sqrt{mN}\kappa^p\epsilon^{-2}\right)$	$\mathcal{O}\left(\kappa^q\epsilon^{-2}\right)$	✗
	`DREAM` (Ours) Theorem 3.1	$\mathcal{O}(N+\sqrt{N}\kappa^2\epsilon^{-2})$	$\tilde{\mathcal{O}}\left(\kappa^2\epsilon^{-2}\right)$	✓
Online	`DM-HSGD` Xian et al. (2021)	$\mathcal{O}\left(\kappa^3\epsilon^{-3}\right)$	$\mathcal{O}\left(\kappa^3\epsilon^{-3}\right)$	✗
	`DREAM` (Ours) Theorem 3.1	$\mathcal{O}(\kappa^3\epsilon^{-3})$	$\tilde{\mathcal{O}}\left(\kappa^2\epsilon^{-2}\right)$	✓

The mixing matrix tells us how the agents in the network communicate with their neighbors. We assume it satisfies the following assumption (Song et al., 2023).

Assumption 2.5. *We suppose the matrix $W \in \mathbb{R}^{m\times m}$ have the following properties:*

a. supported on the network: $W_{i,j} \geq 0$ if and only if i and j are connected in the network.

b. irreducible: W cannot be conjugated into block upper triangular form by a permutation matrix.

c. symmetric: $W = W^\top$.

d. doubly stochastic: $W\mathbf{1} = W^\top\mathbf{1} = \mathbf{1}$.

e. positive semidefinite: $W \succeq O_m$.

Note that if a matrix W satisfies Assumption 2.5 a-d, we can let Assumption 2.5e be automatically satisfied by choosing $(W + I_m)/2$ to be the new mixing matrix.

By the Perron–Frobenius theorem, the eigenvalues of W can be sorted by

$$0 \leq \lambda_m(W) \leq \cdots \leq \lambda_2(W) < \lambda_1(W) = 1.$$

We then define the spectral gap of W as follows.

Definition 2.3. *For a matrix W that satisfies Assumption 2.5, we define the spectral gap as $\delta \triangleq 1 - \lambda_2(W)$.*

It is well known that the spectral gap of W is related to the mixing rate on the network.

Proposition 2.2 (Koloskova et al. (2019, Lemma 16))**.** *Given a matrix W that satisfies Assumption 2.5, for any vector $\mathbf{a} \in \mathbb{R}^{m\times d}$, for the standard mixing iterate given by $\mathbf{a}^{(k+1)} = W\mathbf{a}^{(k)}$, we have*

$$\|W\mathbf{a} - \mathbf{1}\bar{\mathbf{a}}\| \leq (1-\delta)^K\|\mathbf{a} - \mathbf{1}\bar{\mathbf{a}}\|,$$

where $\|\cdot\|$ is the Frobenius norm.

This simple mixing strategy is adopted by previous works including Tsaknakis et al. (2020); Xian et al. (2021); Zhang et al. (2021b), but it would lead to an unavoidable communication complexity dependency of at least $\mathcal{O}(1/\delta)$, which is suboptimal in the dependency of δ. To accelerate the mixing rate, we introduce the `FastMix` sub-procedure (Liu and Morse, 2011), which can lead to the optimal $\mathcal{O}(1/\sqrt{\delta}\,)$ dependency.

Proposition 2.3 (Ye et al. (2020a, Proposition 1))**.** *Given a matrix W that satisfies Assumption 2.5, running Algorithm 1 ensures $\frac{1}{m}\mathbf{1}^\top\mathbf{a}^{(K)} = \bar{a}^{(0)}$ and*

$$\left\|\mathbf{a}^{(K)} - \mathbf{1}\bar{a}^{(0)}\right\| \leq c_1\left(1 - c_2\sqrt{\delta}\,\right)^K\left\|\mathbf{a}^{(0)} - \mathbf{1}\bar{a}^{(0)}\right\|,$$

where $\bar{a}^{(0)} = \frac{1}{m}\mathbf{1}^\top\mathbf{a}^{(0)}$, $\|\cdot\|$ is the Frobenius norm, $c_1 = \sqrt{14}$ and $c_2 = 1 - 1/\sqrt{2}$.

To tackle the possible constraint in y, we define the projection and the constrained reduced gradient (Nesterov, 2018).

Definition 2.4. *We define*

$$\Pi(y) \triangleq \arg\min_{y' \in \mathcal{Y}} \|y' - y\|^2 \ \text{and}\ \mathbf{\Pi}(\mathbf{y}) \triangleq \arg\min_{\mathbf{y}'(i) \in \mathcal{Y}} \|\mathbf{y}' - \mathbf{y}\|^2$$

for $y \in \mathbb{R}^{d_y}$ *and* $\mathbf{y} \in \mathbb{R}^{m \times d_y}$ *respectively.*

Definition 2.5. *We also define the constrained reduced gradient of* f *at* (x, y) *with respect to* y *as*

$$G_\eta(x, y) = \frac{\Pi(y + \eta \nabla_y f(x, y)) - y}{\eta}$$

with some $0 < \eta \le 1/L$.

3 The Proposed Algorithm

In this section, we propose a novel stochastic algorithm named Decentralized Recursive-gradient dEscent Ascent Method (`DREAM`) for decentralized nonconvex-strongly-concave minimax problems. We provide a unified framework for analyzing our `DREAM` for both online and offline cases. It shows the algorithm can find an ϵ-stationary point of $P(x)$ within at most $\mathcal{O}(\kappa^3 \epsilon^{-3})$ and $\mathcal{O}(N + \sqrt{N}\kappa^2\epsilon^{-2})$ SFO calls for the online and offline setting respectively; and both of two settings require at most $\mathcal{O}(\kappa^2 \epsilon^{-2} \log m/\sqrt{\delta})$ communication rounds.

3.1 Method Overview

`DREAM` constructs stochastic recursive gradients $\mathbf{g}_t(i)$ (Fang et al., 2018; Li et al., 2021; Luo et al., 2020; Nguyen et al., 2017) to estimate the local gradients, *i.e.* $\mathbf{g}_t(i) \approx \nabla f_i(\mathbf{z}_t(i))$. This step (Line 22 in Algorithm 2) reduces the variance of stochastic gradient estimators on each agent, leading to the optimal $\mathcal{O}(\epsilon^{-3})$ dependency in the SFO upper complexity bound. `DREAM` then applies the gradient tracking technique (Di Lorenzo and Scutari, 2016; Qu and Li, 2019) to track the average gradient over the network via the gradient tracker $\mathbf{s}_t(i)$, *i.e.* $\mathbf{s}_t(i) \approx \nabla f(\mathbf{z}_t(i))$. The gradient tracker $\mathbf{s}_t(i)$ is also updated in a recursive way (Line 24 in Algorithm 2). This step is commonly used to achieve convergence when the data distribution on each agent does not satisfy the i.i.d assumption in decentralized optimization (Nedic et al., 2017; Song et al., 2023). With the gradient tracker, `DREAM` applies the two-timescale gradient descent ascent (Lin et al., 2020a) to solve the maximization of y and minimization of x simultaneously (Line 13-14 in Algorithm 2). The random variable $\zeta_t \sim \text{Bernoulli}(p)$ (Line 12 in Algorithm 2) decides the batch size and the number of consensus steps in the iteration. By appropriately choosing the probability p, batch sizes b, b', and consensus steps K, K', `DREAM` achieves the best-known computation complexity and communication complexity trade-off in decentralized nonconvex-strongly-concave minimax optimization.

3.2 A Novel Lyapunov Function

Different from previous works (Luo et al., 2020; Xian et al., 2021; Zhang et al., 2021b), we propose a novel Lyapunov function as follows:

$$\begin{aligned}
\Phi_t &\triangleq \Psi_t + \frac{1}{m\eta} C_t + \frac{\eta}{mp} V_t + \frac{\eta}{p} U_t, \quad \text{where} \\
\Psi_t &= P(\bar{x}_t) - P^* + \alpha(P(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)), \\
C_t &= \|\mathbf{z}_t - \mathbf{1}\bar{z}_t\|^2 + \eta^2 \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\|^2 \\
V_t &= \frac{1}{m} \|\mathbf{g}_t - \nabla \mathbf{f}(\mathbf{z}_t)\|^2 \ \text{ and} \\
U_t &= \left\| \frac{1}{m} \sum_{i=1}^{m} (\mathbf{g}_t(i) - \nabla f_i(\mathbf{z}_t(i))) \right\|^2.
\end{aligned} \tag{4}$$

We set $\alpha \in (0, 1]$ in later analysis. Below, we illustrate the meaning of each quantity in the Lyapunov function.

- Ψ_t measures the optimization error. The gradient descent ascent step ensures that Ψ_t can decrease monotonically at each iteration (Chen et al., 2022; Yang et al., 2020).
- C_t measures the consensus error, which can be bounded by the properties of gradient tracking and consensus steps (Song et al., 2023).
- V_t and U_t measure the variance of $\mathbf{g}_t$, which can be bounded by the property of martingale (Fang et al., 2018; Luo et al., 2020). We provide a detailed discussion for the roles of V_t and U_t after Lemma 3.3 and 3.4.

We can show that C_t, V_t, and U_t are sufficiently small during the iterations, which allows us to analyze `DREAM` by analogizing gradient descent ascent on the mean variables $\bar{x}_t$ and $\bar{y}_t$ with some small noise.

Remark 3.1. *Our Lyapunov function* Φ_t *characterizes the sub-optimality of the maximization problem* $\max_{y \in \mathcal{Y}} f(\bar{x}_t, y)$ *by* $P(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)$*, which is easy to be analyzed for stochastic variance reduced algorithm in the decentralized setting. In contrast, Xian et al. (2021) and Zhang et al. (2021b) measure the sub-optimality by* $\|\bar{y}_t - y^*(\bar{x}_t)\|^2$*, which leads to their analysis be more complicated than ours.*

3.3 Convergence Analysis

Our analysis starts from the following descent lemma.

Lemma 3.1. *For Algorithm 2, we set the parameters by* $\eta \le 1/(4L)$ *and*

$$\gamma = \frac{\alpha}{(1+\alpha) 128 \kappa^2}. \tag{5}$$

Algorithm 2 Decentralized Recursive-gradient dEscent Ascent Method (`DREAM`)

Notations: Let $\mathbf{z}_t = [\mathbf{x}_t, \mathbf{y}_t] \in \mathbb{R}^{m\times d}$ and $\mathbf{s}_t = [\mathbf{u}_t, \mathbf{v}_t] \in \mathbb{R}^{m\times d}$.

Input: initial parameter $\bar{z}_0 \in \mathbb{R}^d$, stepsize $\eta > 0$, stepsize ratio $\gamma \in (0, 1]$, probability $p \in (0, 1]$, small mini-batch size b, large mini-batch size b' (we set $b' = n$ for the offline case), initial communication rounds K_0, small communication rounds K, large communication rounds K'.

$\mathbf{z}_0 = \mathbf{1}\bar{z}_0$

parallel for $i = 1, \dots, m$ **do**

  Sample $\mathcal{S}_0(i) = \begin{cases} \{\xi_{i,1}, \dots, \xi_{i,b'}\} \text{ i.i.d.}, & \text{online case;} \\ \{\xi_{i,1}, \dots, \xi_{i,n}\}, & \text{offline case;} \end{cases}$

  $\mathbf{g}_0(i) = \dfrac{1}{b'} \sum_{\xi_{i,j} \in \mathcal{S}_0(i)} \nabla F_i(\mathbf{z}_0(i); \xi_{i,j})$

end parallel for

$\mathbf{s}_0 = \texttt{FastMix}(\mathbf{g}_0, K_0)$

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

  Sample $\zeta_t \sim \text{Bernoulli}(p)$

  $\mathbf{x}_{t+1} = \texttt{FastMix}\big(\mathbf{x}_t - \gamma\eta\mathbf{u}_t, K\big)$

  $\mathbf{y}_{t+1} = \texttt{FastMix}\big(\mathbf{\Pi}(\mathbf{y}_t + \eta\mathbf{v}_t), K\big)$

  parallel for $i = 1, \dots, m$ **do**

    if $\zeta_t = 1$ **do**

      Sample $\mathcal{S}'_t(i) = \begin{cases} \{\xi_{i,1}, \dots, \xi_{i,b'}\} \text{ i.i.d.}, & \text{online case;} \\ \{\xi_{i,1}, \dots, \xi_{i,n}\}, & \text{offline case;} \end{cases}$

      $\mathbf{g}_{t+1}(i) = \dfrac{1}{b'} \sum_{\xi_{i,j} \in \mathcal{S}'_t(i)} \nabla F_i(\mathbf{z}_{t+1}(i); \xi_{i,j})$

    else

      Sample $\omega_t(i) \sim \text{Bernoulli}(q)$

      Sample $\mathcal{S}_t(i) = \{\xi_{i,1}, \dots, \xi_{i,b}\}$ i.i.d.

      $\mathbf{g}_{t+1}(i) = \mathbf{g}_t(i) + \dfrac{\omega_t(i)}{bq} \sum_{\xi_{i,j} \in \mathcal{S}_t(i)} \big(\nabla F_i(\mathbf{z}_{t+1}(i); \xi_{i,j}) - \nabla F_i(\mathbf{z}_t(i); \xi_{i,j})\big)$

    end if

  end parallel for

  $\mathbf{s}_{t+1} = \begin{cases} \texttt{FastMix}\big(\mathbf{s}_t + \mathbf{g}_{t+1} - \mathbf{g}_t, K'\big), & \text{if } \zeta_t = 1; \\ \texttt{FastMix}\big(\mathbf{s}_t + \mathbf{g}_{t+1} - \mathbf{g}_t, K\big), & \text{if } \zeta_t = 0; \end{cases}$

end for

Output: x_{out} by uniformly sampling from $\{\mathbf{x}_0(1), \mathbf{x}_0(2), \cdots, \mathbf{x}_{T-1}(m)\}$

Then for any $\alpha > 0$ it holds that

$$\mathbb{E}[\Psi_{t+1}] \leq \Psi_t - \frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t)\|^2 - \frac{1}{8\gamma\eta}\mathbb{E}\|\bar{x}_{t+1} - \bar{x}_t\|^2 - \frac{\alpha}{16\eta}\mathbb{E}\|\bar{z}_{t+1} - \bar{z}_t\|^2 + 6\alpha\eta U_t + \frac{3\alpha}{m\eta}C_t.$$

If both C_t and U_t are sufficiently small, the above lemma indicates that the optimization error decreases by roughly $(\gamma\eta/2)\|\nabla P(\bar{x}_t)\|^2$ at each step in expectation and the remaining proof can follow the gradient descent (Bubeck, 2015; Nesterov, 2018) on the primal function $P(x)$.

Recall Proposition 2.3. We further define the discount factors of consensus error that arises from mixing steps (Line 10, 13, 14 and 24 in Algorithm 2) as:

$$\begin{aligned} \rho_0 &= c_1\big(1 - c_2\sqrt{\delta}\big)^{K_0}, \\ \rho' &= c_1\big(1 - c_2\sqrt{\delta}\big)^{K'}, \\ \rho &= c_1\big(1 - c_2\sqrt{\delta}\big)^{K}. \end{aligned} \tag{6}$$

Then we can bound the consensus error as follows.

Lemma 3.2. *For Algorithm 2, let $\rho^2 \leq 1/24$. Then*

$$\mathbb{E}[C_{t+1}] \leq 12c\rho^2 C_t + 6\rho'^2\eta^2 m V_t + 2c\rho^2 m\mathbb{E}\|\bar{z}_{t+1} - \bar{z}_t\|^2 + \frac{6\rho'^2 m\eta^2\sigma^2}{b'}\mathbb{I}[b' < n],$$

where we define $c = \max\{1/(bq), 1\}$.

Remark 3.2. *For convenience, we define $n = +\infty$ for the online case and $b' = n$ for the offline case. Then*

$$\mathbb{I}[b' < n] = \begin{cases} 1, & \text{for online case,} \\ 0, & \text{for offline case.} \end{cases}$$

This notation allows us to present the analysis for both cases in one unified framework.

Lemma 3.1 and 3.2 mean the decrease of the optimization error Ψ_t and consensus error C_t requires the reasonable upper bounds of V_t and U_t, which can be characterized by the following recursions.

Lemma 3.3. *For Algorithm 2, we have*

$$\mathbb{E}[V_{t+1}] \leq (1-p)V_t + \frac{4(1-p)L^2}{mbq}C_t + \frac{p\sigma^2}{b'}\mathbb{I}[b' < n] + \frac{3(1-p)L^2}{bq}\mathbb{E}\|\bar{z}_{t+1} - \bar{z}_t\|^2.$$

Lemma 3.4. *For Algorithm 2, we have*

$$\mathbb{E}[U_{t+1}] \leq (1-p)U_t + \frac{4(1-p)L^2}{m^2bq}C_t + \frac{p\sigma^2}{mb'}\mathbb{I}[b' < n] + \frac{3(1-p)L^2}{mbq}\mathbb{E}\|\bar{z}_{t+1} - \bar{z}_t\|^2.$$

Note that for any vector sequence $a_1, \cdots, a_m$ we always have $\|\sum_{i=1}^m a_i\|^2 \leq m\sum_{i=1}^m \|a_i\|^2$, which directly implies $U_t \leq V_t$. Therefore one can use the quantity V_t only in the analysis as Zhang et al. (2021b), but the separation of U_t and V_t makes our bound tighter. As a consequence, we show a linear speed-up in the SFO complexity with respect to the number of agents m which was not shown by Zhang et al. (2021b).

Putting Lemma 3.1, 3.2 and 3.3 together, we can prove the main result for `DREAM` as follows.

Theorem 3.1. *For Algorithm 2 , we set parameters*

$$\begin{aligned} &\eta = \frac{1}{48L}, \quad b = \left\lceil \sqrt{\frac{b'}{m}}\, \right\rceil, \quad q = \frac{1}{b}\sqrt{\frac{b'}{m}}, \quad p = \frac{bq}{bq+b'} \\ &T = \left\lceil \frac{16\Psi_0}{\gamma\eta\epsilon^2} + \frac{2}{p} \right\rceil, \quad K_0 = \left\lceil \frac{\log\big(16c_1/(\gamma m\epsilon^2)\big)}{c_2\sqrt{\delta}} \right\rceil, \\ &K = \left\lceil \frac{5\log(c_1(m/b'+1))}{c_2\sqrt{\delta}} \right\rceil, \quad K' = \left\lceil \frac{5\log(c_1 m)}{c_2\sqrt{\delta}} \right\rceil, \end{aligned} \tag{7}$$

where c_1, c_2 are defined in Proposition 2.3, $\gamma = \Theta(\kappa^{-2})$ follows (5), $\alpha = 1/8$ and

$$b' = \begin{cases} \lceil 32\sigma^2/(\gamma m\epsilon^2) \rceil, & \text{for online case,} \\ n, & \text{for offline case.} \end{cases} \tag{8}$$

Then the output satisfies $\mathbb{E}\|\nabla P(x_{\text{out}})\| \leq \epsilon$ within the overall SFO complexity

$$\begin{cases} \mathcal{O}\big(\kappa^2\sigma^2\epsilon^{-2} + \kappa^3 L\sigma\epsilon^{-3}\big), & \text{for online case;} \\ \mathcal{O}\big(mn + \sqrt{mn}\kappa^2 L\epsilon^{-2}\big), & \text{for offline case;} \end{cases}$$

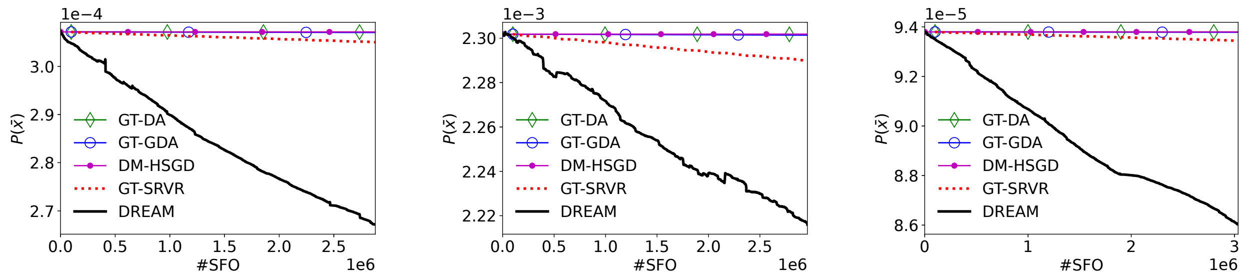
and communication complexity $\mathcal{O}((\kappa^2 L\epsilon^{-2}\log m)/\sqrt{\delta}\,)$.

This theorem shows that `DREAM` achieves the best-known complexity guarantee both in computation and communication, either for online or offline cases.

4 Experiments

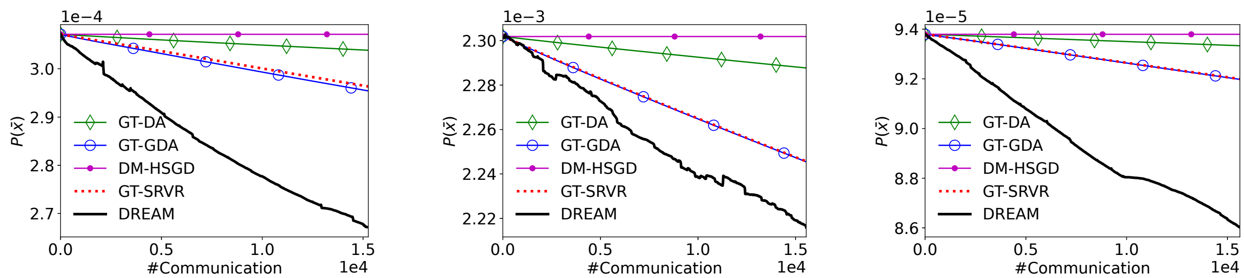
We conduct numerical experiments on the model of robust logistic regression (Luo et al., 2020; Yan et al., 2019). The source codes are available [1]. It considers training the binary classifier $x \in \mathbb{R}^d$ on dataset $\{(a_k, b_k)\}_{k=1}^N$, where $a_k \in \mathbb{R}^d$ is the feature of the k-th sample and $b_k \in \{1, -1\}$ is the corresponding label. The total N samples are equally distributed on m agents, with $n = N/m$ samples on each agent. The decentralized minimax optimization problem of

[1] https://github.com/TrueNobility303/DREAM



(a) a9a (b) w8a (c) ijcnn1

Figure 1: Comparison on the number of SFO calls against $P(\bar{x}) \triangleq \max_{y\in\Delta_N} f(\bar{x}, y)$.



(a) a9a (b) w8a (c) ijcnn1

Figure 2: Comparison on the number of communication rounds against $P(\bar{x}) \triangleq \max_{y\in\Delta_N} f(\bar{x}, y)$.

this model corresponds to formulation (1) with local functions

$$f_i(x) \triangleq \frac{1}{n}\sum_{i=1}^{n}\big(y_{ij}l_{ij}(x) - V(y) + g(x)\big),$$

where $l_{ij}(x) = \log(1 + \exp(b_{ij}a_{ij}^\top x))$,

$$V(y) = \frac{1}{2N^2}\|Ny - \mathbf{1}\|^2,\ g(x) = \theta\sum_{k=1}^{d}\frac{\nu x_k^2}{1+\nu x_k^2},\ \theta = 10^{-5},$$

$$\nu = 10 \ \text{ and } \ \Delta_N = \Big\{y \in \mathbb{R}^N : y_k \in [0,1], \sum_{k=1}^{N} y_k = 1\Big\}.$$

We set mixing matrix W as the π-lazy random walk matrix on a ring graph with 8 nodes ($m = 8$) and let $\tau = 0.999$, which leads to a low consensus rate since $\lambda_2(W) = \tau + (1-\tau)\cos(2\pi(m-1)/m)$.

We compare the performance of our proposed method `DREAM` with `DM-HSGD` (Xian et al., 2021), `GT-DA` (Tsaknakis et al., 2020), `GT-GDA` and `GT-SRVR` (Zhang et al., 2021b) on datasets "a9a", "w8a" and "ijcnn1" (Chang and Lin, 2011). All of the algorithms are implemented by MPI to simulate the distributed training scenario.

For `DREAM`, we tune b, b' from $\{64, 128, 256, 512\}$, tune p, q from $\{0.2, 0.5, 0.9\}$ and tune K_0, K, K' from $\{2, 5, 10\}$. For `DM-HSGD` , we set momentum parameter $\beta = 0.01$ by following Xian et al. (2021), and tune b, b_0 from $\{64, 128, 256, 512\}$. For `GT-SRVR`, we let the epoch length $Q = \lceil\sqrt{n}\,\rceil$ as suggested by the authors. For `GT-DA`, we let the number of inner loops $R = 4$ by following Tsaknakis et al. (2020). For each algorithm, we tune η and γ from $\{1, 0.1, 0.01, 0.001\}$ and $\{0.1, 0.01, 0.001, 0.0001\}$ respectively.

We present the empirical results for the comparisons of computation complexity and communication complexity in Figure 1 and 2. It is shown that the proposed `DREAM` performs apparently better than baselines.

5 Conclusion and Future Work

In this paper, we have proposed an efficient algorithm `DREAM` for decentralized stochastic nonconvex-strongly-concave minimax optimization. The theoretical analysis showed `DREAM` achieves the best-known SFO complexity of $\mathcal{O}(\min(\kappa^3\epsilon^{-3}, \kappa^2\sqrt{N}\epsilon^{-2}))$ and communication complexity of $\tilde{\mathcal{O}}(\kappa^2\epsilon^{-2})$. The numerical experiments on robust logistic regression show the empirical performance of `DREAM` is obviously better than state-of-the-art methods. We discuss possible future directions and some subsequent works in Appendix H.

Acknowledgements

Luo has been supported by National Natural Science Foundation of China (No. 62206058), Shanghai Sailing

Program (22YF1402900), and Shanghai Basic Research Program (23JC1401000). Ye has been supported by the National Natural Science Foundation of China (No. 12101491).

References

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology*, 2(3): 1–27, 2011. URL `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

Congliang Chen, Jiawei Zhang, Li Shen, Peilin Zhao, and Zhiquan Luo. Communication efficient primal-dual algorithm for nonconvex nonsmooth distributed optimization. In *AISTATS*, 2021.

Lesi Chen, Boyuan Yao, and Luo Luo. Faster stochastic algorithms for minimax optimization under Polyak-Łojasiewicz condition. In *NeurIPS*, 2022.

Ziyi Chen, Zhengyang Hu, Qunwei Li, Zhe Wang, and Yi Zhou. A cubic regularization approach for finding local minimax points in nonconvex minimax optimization. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *NeurIPS*, 2019.

Paolo Di Lorenzo and Gesualdo Scutari. NEXT: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *AISTATS*, 2021.

Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu. Learning with average top-k loss. In *NeurIPS*, 2017.

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

Farzan Farnia and Asuman Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *ICML*, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Zhishuai Guo, Yan Yan, Zhuoning Yuan, and Tianbao Yang. Fast objective & duality gap convergence for non-convex strongly-concave min-max problems with PL condition. *Journal of Machine Learning Research*, 24(148):1–63, 2023.

Hadrien Hendrikx, Francis Bach, and Laurent Massoulie. An optimal algorithm for decentralized finite-sum optimization. *SIAM Journal on Optimization*, 31 (4):2753–2783, 2021.

Feihu Huang and Songcan Chen. Near-optimal decentralized momentum method for nonconvex-PL minimax problems. *arXiv preprint arXiv:2304.10902*, 2023.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *ICML*, 2020.

Jikai Jin, Bohang Zhang, Haiyang Wang, and Liwei Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. In *NeurIPS*, 2021.

Yujia Jin and Aaron Sidford. Efficiently solving mdps with stochastic mirror descent. In *ICML*, 2020.

Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *ICML*, 2019.

Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. In *NeurIPS*, 2020.

Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *NeurIPS*, 2020.

Boyue Li, Shicong Cen, Yuxin Chen, and Yuejie Chi. Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. *Journal of Machine Learning Research*, 21(80): 1–51, 2020.

Boyue Li, Zhize Li, and Yuejie Chi. DESTRESS: Computation-optimal and communication-efficient decentralized nonconvex finite-sum optimization. *SIAM Journal on Mathematics of Data Science*, 4(3):1031–1051, 2022a.

Huan Li, Zhouchen Lin, and Yongchun Fang. Variance reduced extra and diging and their optimal acceleration for strongly convex decentralized optimization. *Journal of Machine Learning Research*, 23(222):1–41, 2022b.

Jiajin Li, Linglingzhi Zhu, and Anthony Man-Cho So. Nonsmooth composite nonconvex-concave minimax optimization. *arXiv preprint arXiv:2209.10825*, 2022c.

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

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

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

Ji Liu and A. Stephen Morse. Accelerated linear iterations for distributed averaging. *Annual Reviews in Control*, 35(2):160–165, 2011.

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

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *ICML*, 2021.

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

Luo Luo, Li Yujun, and Cheng Chen. Finding second-order stationary points in nonconvex-strongly-concave minimax optimization. In *NeurIPS*, 2022.

Gabriel Mancino-Ball and Yangyang Xu. Variance-reduced accelerated methods for decentralized stochastic double-regularized nonconvex strongly-concave minimax problems. *arXiv preprint arXiv:2307.07113*, 2023.

Vaishnavh Nagarajan and J. Zico Kolter. Gradient descent GAN optimization is locally stable. *NIPS*, 2017.

Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

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

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

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

Shuang Qiu, Zhuoran Yang, Xiaohan Wei, Jieping Ye, and Zhaoran Wang. Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear TD learning. *arXiv preprint arXiv:2008.10103*, 2020.

Guannan Qu and Na Li. Accelerated distributed Nesterov gradient descent. *IEEE Transactions on Automatic Control*, 65(6):2566–2581, 2019.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pages 1–35, 2021.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifying some distributional robustness with principled adversarial training. In *ICLR*, 2018.

Zhuoqing Song, Lei Shi, Shi Pu, and Ming Yan. Optimal gradient tracking for decentralized optimization. *Mathematical Programming*, pages 1–53, 2023.

Haoran Sun, Songtao Lu, and Mingyi Hong. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *ICML*, 2020.

Ioannis Tsaknakis, Mingyi Hong, and Sijia Liu. Decentralized min-max optimization: Formulations, algorithms and applications in network poisoning attack. In *ICASSP*, 2020.

César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. A dual approach for optimal algorithms in distributed optimization over networks. In *ITA*. IEEE, 2020.

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

Zhiguo Wang, Jiawei Zhang, Tsung-Hui Chang, Jian Li, and Zhi-Quan Luo. Distributed stochastic consensus optimization with momentum for nonconvex nonsmooth problems. *IEEE Transactions on Signal Processing*, 69:4486–4501, 2021.

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

Ran Xin, Usman A. Khan, and Soummya Kar. Fast decentralized nonconvex finite-sum optimization with recursive variance reduction. *SIAM Journal on Optimization*, 32(1):1–28, 2022.

Tengyu Xu, Zhe Wang, Yingbin Liang, and H. Vincent Poor. Gradient free minimax optimization: Variance reduction and faster convergence. *arXiv preprint arXiv:2006.09361*, 2020.

Yangyang Xu. Decentralized gradient descent maximization method for composite nonconvex strongly-concave minimax problems. *arXiv preprint arXiv:2304.02441*, 2023.

Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, pages 1–72, 2023.

Yan Yan, Yi Xu, Qihang Lin, Lijun Zhang, and Tianbao Yang. Stochastic primal-dual algorithms with faster convergence than $O(1/\sqrt{T})$ for problems without bilinear structure. *arXiv preprint arXiv:1904.10112*, 2019.

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

Haishan Ye, Luo Luo, Ziang Zhou, and Tong Zhang. Multi-consensus decentralized accelerated gradient descent. *arXiv preprint arXiv:2005.00797*, 2020a.

Haishan Ye, Ziang Zhou, Luo Luo, and Tong Zhang. Decentralized accelerated proximal gradient descent. In *NeurIPS*, 2020b.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

Zhuoning Yuan, Zhishuai Guo, Yi Xu, Yiming Ying, and Tianbao Yang. Federated deep AUC maximization for heterogeneous data with a constant communication complexity. *arXiv preprint arXiv:2102.04635*, 2021.

Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. In *NeurIPS*, 2020.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. In *NeurIPS*, 2019.

Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *UAI*, 2021a.

Xin Zhang, Zhuqing Liu, Jia Liu, Zhengyuan Zhu, and Songtao Lu. Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning. In *NeurIPS*, 2021b.

Xuan Zhang, Necdet Serhat Aybat, and Mert Gurbuzbalaban. SAPD+: An accelerated stochastic method for nonconvex-concave minimax problems. In *NeurIPS*, 2022.

Xuan Zhang, Gabriel Mancino-Ball, Necdet Serhat Aybat, and Yangyang Xu. Jointly improving the sample and communication complexities in decentralized stochastic minimax optimization. *arXiv preprint arXiv:2307.09421*, 2023.

Taoli Zheng, Linglingzhi Zhu, Anthony Man-Cho So, Jose Blanchet, and Jiajin Li. Doubly smoothed GDA: Global convergent algorithm for constrained nonconvex-nonconcave minimax optimization. *arXiv preprint arXiv:2212.12978*, 2022.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material: An Efficient Stochastic Algorithm for Decentralized Nonconvex-Strongly-Concave Minimax Optimization

A Some Useful Lemmas

We first provide some useful lemmas.

Lemma A.1. *For any $a_1, \dots, a_m \in \mathbb{R}^d$, we have*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} a_i \right\|^2 \leq \frac{1}{m} \sum_{i=1}^{m} \|a_i\|^2 .$$

Lemma A.2. *For any matrix $\mathbf{z} \in \mathbb{R}^{m \times d}$ and $\bar{z} = \frac{1}{m}\mathbf{1}^\top \mathbf{z}$, we have*

$$\|\mathbf{z} - \mathbf{1}\bar{z}\| \leq \|\mathbf{z}\| ,$$

Lemma A.3. *Under Assumption 2.2, we have*

$$\|\nabla \mathbf{f}(\mathbf{z}) - \nabla \mathbf{f}(\mathbf{z}')\| \leq L \|\mathbf{z} - \mathbf{z}'\|$$

for any $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^{m \times d}$.

Lemma A.4. *For Algorithm 2, we have $\bar{s}_t = \frac{1}{m}\mathbf{1}^\top \mathbf{g}_t = \bar{g}_t$.*

Proof. We prove this lemma by induction. For $t = 0$, we have

$$\bar{s}_0 = \frac{1}{m}\mathbf{1}^\top \mathbf{s}_0 = \frac{1}{m}\mathbf{1}^\top \mathbf{g}_0.$$

Suppose the statement holds for $t \leq k$. Then for $t = k+1$, the induction base means that

$$\begin{aligned}
&\bar{s}_{k+1} \\
=&\bar{s}_k + \bar{g}_{k+1} - \bar{g}_k \\
=&\frac{1}{m}\mathbf{1}^\top \mathbf{g}(\mathbf{z}_k) + \bar{g}_{k+1} - \frac{1}{m}\mathbf{1}^\top \mathbf{g}(\mathbf{z}_k) \\
=&\bar{g}_{k+1} \\
=&\frac{1}{m}\mathbf{1}^\top \mathbf{g}_{k+1},
\end{aligned}$$

which finished the proof. ∎

Lemma A.5. *Under Assumption 2.2, for Algorithm 2, it holds that*

$$\|\bar{s}_t - \nabla f(\bar{z}_t)\|^2 \leq 2 \left\| \frac{1}{m} \sum_{i=1}^{m} (\mathbf{g}_t(i) - \nabla f_i(\mathbf{z}_t(i))) \right\|^2 + \frac{2L^2}{m} \|\mathbf{z}_t - \mathbf{1}\bar{z}_t\|^2.$$

Proof. Using Young's inequality and Assumption 2.2, we have

$$
\begin{aligned}
&\|\bar{s}_t - \nabla f(\bar{z}_t)\|^2 \\
=&\left\|\frac{1}{m}\sum_{i=1}^{m}\big(\mathbf{g}_t(i) - \nabla f_i(\bar{z}_t)\big)\right\|^2 \\
\leq&2\left\|\frac{1}{m}\sum_{i=1}^{m}(\mathbf{g}_t(i) - \nabla f_i(\mathbf{z}_t(i)))\right\|^2 + 2\left\|\frac{1}{m}\sum_{i=1}^{m}(\nabla f_i(\mathbf{z}_t(i)) - \nabla f_i(\bar{z}_t))\right\|^2 \\
\leq&2\left\|\frac{1}{m}\sum_{i=1}^{m}(\mathbf{g}_t(i) - \nabla f_i(\mathbf{z}_t(i)))\right\|^2 + \frac{2}{m}\sum_{i=1}^{m}\|\nabla f_i(\mathbf{z}_t(i)) - \nabla f_i(\bar{z}_t)\|^2 \\
\leq&\ 2\ \left\|\frac{1}{m}\sum_{i=1}^{m}(\mathbf{g}_t(i) - \nabla f_i(\mathbf{z}_t(i)))\right\|^2 + \frac{2L^2}{m}\left\|\mathbf{z}_t - \mathbf{1}\bar{z}_t\right\|^2.
\end{aligned}
$$

□

Lemma A.6. *When $f(x,y)$ is L-smooth and μ-strongly-concave in y, it holds that*

$$\|\nabla P(x) - \nabla_x f(x,y)\| \leq 2\kappa\|G_\eta(x,y)\|,$$

where $G_\eta(x,y)$ denotes the constrained reduced gradient

$$G_\eta(x,y) = \frac{\Pi(y + \eta\nabla_y f(x,y)) - y}{\eta}$$

with any $\eta \leq 1/L$.

Proof. Using the Danskin's theorem, i.e. $\nabla P(x) = \nabla_x f(x, y^*(x))$ as well as the L-smoothness, we have

$$\|\nabla P(x) - \nabla_x f(x,y)\| \leq L\|y^*(x) - y\|. \tag{9}$$

The Corollary 1 of Luo et al. (2020, Appendix A) tells us

$$\mu\|y - y^*(x)\| \leq 2\|G_\eta(x,y)\|. \tag{10}$$

We prove this lemma by combing inequalities (9) and (10). □

B Discussions on DM-HSGD

Decentralized Minimax Hybrid Stochastic Gradient Descent (DM-HSGD) considers the minimax problem (1) in both unconstrained and constrained cases (Xian et al., 2021). However, the theoretical analysis of this method only works for unconstrained case, and it is problematic for the general constrained case.[2]

The analysis of DM-HSGD heavily relies on its Lemma 5 (Xian et al., 2021, Appendix A.2). However, the proof of this result is based on the following equation (Xian et al., 2021, Equation (29))[3]

$$2\eta_y\langle\bar{u}_t, \hat{y}_t - \bar{y}_t\rangle = \|\bar{y}_t - \hat{y}_t\|^2 + \|\bar{y}_{t+1} - \bar{y}_t\|^2 - \|\bar{y}_{t+1} - \hat{y}_t\|^2, \tag{11}$$

where $\hat{y}_t = \arg\max_{y\in\mathcal{Y}} f(\bar{x}_t, y)$ corresponds to $y^*(\bar{x}_t)$ in our notations. Note that Equation (11) does not hold for general convex and compact set $\mathcal{Y}$. We further illustrate this below.

We denote $\mathbf{y}_{t+1/2} = \mathbf{y}_t + \eta_y\mathbf{u}_t$ by following Xian et al. (2021)'s notation. The procedure of DM-HSGD guarantees that $\bar{y}_{t+1} = \overline{\Pi}(\mathbf{y}_{t+1/2})$, then we have

$$2\eta_y\langle\bar{u}_t, \hat{y}_t - \bar{y}_t\rangle = 2\langle\bar{y}_{t+1/2} - \bar{y}_t, \hat{y}_t - \bar{y}_t\rangle = \|\bar{y}_t - \hat{y}_t\|^2 + \|\bar{y}_{t+1/2} - \bar{y}_t\|^2 - \|\bar{y}_{t+1/2} - \hat{y}_t\|^2. \tag{12}$$

[2] The theoretical issue of DM-HSGD has also been mentioned by Zhang et al. (2023).

[3] The analysis of DM-HSGD uses η_y and $\bar{u}_t$ to present the stepsize and gradient estimator with respect to y, which play the similar roles of η and $\bar{v}_t$ in our notations.

The only difference between equations (11) and (12) is their last terms. In the unconstrained case that $\mathcal{Y} = \mathbb{R}^{d_y}$, it holds that

$$\bar{y}_{t+1} = \bar{\Pi}(\mathbf{y}_{t+1/2}) = \bar{y}_{t+1/2}, \tag{13}$$

which leads to

$$\|\bar{y}_{t+1} - \hat{y}_t\|^2 = \|\bar{y}_{t+1/2} - \hat{y}_t\|^2.$$

However, the equation (13) may not hold in the constrained case, since we can not guarantee $\bar{\Pi}(\mathbf{y}) = \bar{y}$ in general. For example, consider that

$$\mathcal{Y} = \{y \in \mathbb{R}^2 : \|y\|^2 = 2\}, \qquad \mathbf{y}(1) = (2,2) \qquad \text{and} \qquad \mathbf{y}(2) = (2,-2),$$

then we can have $\bar{y} = (2,0)$ while $\bar{\Pi}(\mathbf{y}) = \left(\sqrt{2}, 0\right)$.

For the analysis of projected first-order methods for constrained problems (even for minimization problems), we typically introduce the constrained reduced gradient (Nesterov, 2018, Definition 2.2.3) and apply the first-order optimality condition (such as Equation (16) in Appendix C) to establish the convergence results. However, such popular techniques are not included in the analysis of DM-HSGD.

In decentralized setting, the extension from unconstrained case to constrained case is non-trivial, since the projection step introduces additional consensus error. As a consequence, our analysis is essentially different from the Xian et al. (2021). We design a novel Lyapunov function and present the details in Lemma 3.1 to address this issue.

C The Proof of Lemma 3.1

Proof. Denote $\mathbf{y}'_{t+1} = \mathbf{\Pi}(\mathbf{y}_t + \eta \mathbf{v}_t)$. By Proposition 2.3, the update rules of $\mathbf{x}_t, \mathbf{y}_t$ means

$$\bar{x}_{t+1} = \bar{x}_t - \gamma\eta\bar{u}_t \qquad \text{and} \qquad \bar{y}_{t+1} = \bar{y}'_{t+1} = \bar{\Pi}(\mathbf{y}_t + \eta\mathbf{v}_t). \tag{14}$$

In the view of inexact gradient descent on $P(x)$, it yields

$$\begin{aligned}
&\mathbb{E}[P(\bar{x}_{t+1})] \\
&\le \mathbb{E}\left[P(\bar{x}_t) + \nabla P(\bar{x}_t)^\top(\bar{x}_{t+1} - \bar{x}_t) + \frac{L_P}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2\right] \\
&= \mathbb{E}\left[P(\bar{x}_t) - \gamma\eta\nabla P(\bar{x}_t)^\top\bar{u}_t + \frac{L_P\gamma^2\eta^2}{2}\|\bar{u}_t\|^2\right] \\
&= \mathbb{E}\left[P(\bar{x}_t) - \frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t)\|^2 - \left(\frac{\gamma\eta}{2} - \frac{L_P\gamma^2\eta^2}{2}\right)\|\bar{u}_t\|^2 + \frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t) - \bar{u}_t\|^2\right] \\
&\le \mathbb{E}\left[P(\bar{x}_t) - \frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t)\|^2 - \left(\frac{\gamma\eta}{2} - \frac{L_P\gamma^2\eta^2}{2}\right)\|\bar{u}_t\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{\gamma\eta}{m}\sum_{i=1}^m\|\nabla P(\bar{x}_t) - \nabla_x f(\bar{x}_t, \mathbf{y}_t(i))\|^2 + \frac{\gamma\eta}{m}\sum_{i=1}^m\|\bar{u}_t - \nabla_x f(\bar{x}_t, \mathbf{y}_t(i))\|^2\right] \\
&\le \mathbb{E}\left[P(\bar{x}_t) - \frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t)\|^2 - \left(\frac{\gamma\eta}{2} - \frac{L_P\gamma^2\eta^2}{2}\right)\|\bar{u}_t\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{4\kappa^2\gamma\eta}{m}\sum_{i=1}^m\|G_\eta(\bar{x}_t, \mathbf{y}_t(i))\|^2 + \frac{\gamma\eta}{m}\sum_{i=1}^m\|\bar{u}_t - \nabla_x f(\bar{x}_t, \mathbf{y}_t(i))\|^2\right] \\
&\le \mathbb{E}\left[P(\bar{x}_t) - \frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t)\|^2 - \left(\frac{\gamma\eta}{2} - \frac{L_P\gamma^2\eta^2}{2}\right)\|\bar{u}_t\|^2\right] \\
&\quad + \mathbb{E}\left[\frac{8\kappa^2\gamma\eta}{m}\sum_{i=1}^m\|\mathbf{v}'_t(i)\|^2 + \frac{8\kappa^2\gamma\eta}{m}\sum_{i=1}^m\|\mathbf{v}_t(i) - \nabla_y f(\bar{x}_t, \mathbf{y}_t(i))\|^2 + \frac{\gamma\eta}{m}\sum_{i=1}^m\|\bar{u}_t - \nabla_x f(\bar{x}_t, \mathbf{y}_t(i))\|^2\right],
\end{aligned} \tag{15}$$

where $\mathbf{v}_t'(i) = (\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i))/\eta$. Above, the first inequality follows from the smoothness of $P(x)$ by Proposition 2.1; the second one is due to the Young's inequality; the third inequality holds according to Lemma A.6; in the last one we use the Young's inequality along with

$$\begin{aligned}
&\|\mathbf{v}_t'(i) - G_\eta(\bar{x}_t, \mathbf{y}_t(i))\| \\
=& \frac{1}{\eta}\|\Pi(\mathbf{y}_t + \eta\mathbf{v}_t(i)) - \Pi(\mathbf{y}_t + \eta\nabla_y f(\bar{x}_t, \mathbf{y}_t(i)))\| \\
\leq& \|\mathbf{v}_t(i) - \nabla_y f(\bar{x}_t, \mathbf{y}_t(i))\|.
\end{aligned}$$

Recall that $\mathbf{y}_{t+1}' = \mathbf{\Pi}(\mathbf{y}_t + \eta\mathbf{v}_t)$ and the first-order optimality of the projection, we have

$$(\mathbf{y}_t(i) + \eta\mathbf{v}_t(i) - \mathbf{y}_{t+1}'(i))^\top (y - \mathbf{y}_{t+1}'(i)) \leq 0 \tag{16}$$

for any $y \in \mathcal{Y}$. Taking $y = \bar{y}_t$ in above inequality, we get

$$\begin{aligned}
&\mathbf{v}_t(i)^\top(\bar{y}_t - \mathbf{y}_{t+1}'(i)) \\
\leq& \frac{1}{\eta}(\bar{y}_t - \mathbf{y}_{t+1}'(i))^\top(\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)) \\
=& \frac{1}{2\eta}\|\bar{y}_t - \mathbf{y}_t(i)\|^2 - \frac{1}{2\eta}\|\bar{y}_t - \mathbf{y}_{t+1}'(i)\|^2 - \frac{1}{2\eta}\|\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)\|^2.
\end{aligned} \tag{17}$$

In the view of inexact gradient descent ascent on $f(x, y)$, it yields

$$\begin{aligned}
&\mathbb{E}[-f(\bar{x}_{t+1}, \mathbf{y}_{t+1}'(i))] \\
\leq& \mathbb{E}\left[-f(\bar{x}_{t+1}, \mathbf{y}_t(i)) - \nabla_y f(\bar{x}_{t+1}, \mathbf{y}_t(i))^\top(\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)) + \frac{L}{2}\|\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)\|^2\right] \\
\leq& \mathbb{E}\left[-f(\bar{x}_t, \mathbf{y}_t(i)) - \nabla_x f(\bar{x}_t, \mathbf{y}_t(i))^\top(\bar{x}_{t+1} - \bar{x}_t) + \frac{L}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2\right] \\
&+ \mathbb{E}\left[-\nabla_y f(\bar{x}_{t+1}, \mathbf{y}_t(i))^\top(\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)) + \frac{L}{2}\|\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)\|^2\right] \\
\leq& \mathbb{E}\left[-f(\bar{x}_t, \bar{y}_t) + \nabla_y f(\bar{x}_t, \mathbf{y}_t(i))^\top(\bar{y}_t - \mathbf{y}_t(i)) - \nabla_x f(\bar{x}_t, \mathbf{y}_t(i))^\top(\bar{x}_{t+1} - \bar{x}_t) + \frac{L}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2\right] \\
&+ \mathbb{E}\left[-\nabla_y f(\bar{x}_{t+1}, \mathbf{y}_t(i))^\top(\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)) + \frac{L}{2}\|\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)\|^2\right] \\
=& \mathbb{E}\left[-f(\bar{x}_t, \bar{y}_t) + \nabla_y f(\bar{x}_t, \mathbf{y}_t(i))^\top(\bar{y}_t - \mathbf{y}_{t+1}'(i)) - \nabla_x f(\bar{x}_t, \mathbf{y}_t(i))^\top(\bar{x}_{t+1} - \bar{x}_t) + \frac{L}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2\right] \\
&+ \mathbb{E}\left[(\nabla_y f(\bar{x}_t, \mathbf{y}_t(i)) - \nabla_y f(\bar{x}_{t+1}, \mathbf{y}_t(i)))^\top(\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)) + \frac{L}{2}\|\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)\|^2\right] \\
=& \mathbb{E}\left[-f(\bar{x}_t, \bar{y}_t) + \mathbf{v}_t(i)^\top(\bar{y}_t - \mathbf{y}_{t+1}'(i)) - \nabla_x f(\bar{x}_t, \mathbf{y}_t(i))^\top(\bar{x}_{t+1} - \bar{x}_t) + \frac{L}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2\right] \\
&+ \mathbb{E}\left[(\nabla_y f(\bar{x}_t, \mathbf{y}_t(i)) - \nabla_y f(\bar{x}_{t+1}, \mathbf{y}_t(i)))^\top(\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)) + \frac{L}{2}\|\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)\|^2\right] \\
&+ \mathbb{E}\left[(\nabla_y f(\bar{x}_t, \mathbf{y}_t(i)) - \mathbf{v}_t(i))^\top(\bar{y}_t - \mathbf{y}_{t+1}'(i))\right] \\
\leq& \mathbb{E}\left[-f(\bar{x}_t, \bar{y}_t) - \frac{1}{2\eta}\|\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)\|^2 - \nabla_x f(\bar{x}_t, \mathbf{y}_t(i))^\top(\bar{x}_{t+1} - \bar{x}_t) + \frac{L}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2\right] \\
&+ \mathbb{E}\left[(\nabla_y f(\bar{x}_t, \mathbf{y}_t(i)) - \nabla_y f(\bar{x}_{t+1}, \mathbf{y}_t(i)))^\top(\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)) + \frac{L}{2}\|\mathbf{y}_{t+1}'(i) - \mathbf{y}_t(i)\|^2\right] \\
&+ \mathbb{E}\left[\frac{\eta}{2}\|\nabla_y f(\bar{x}_t, \mathbf{y}_t(i)) - \mathbf{v}_t(i))\|^2 + \frac{1}{2\eta}\|\bar{y}_t - \mathbf{y}_t(i)\|^2\right] \\
\leq& \mathbb{E}\left[-f(\bar{x}_t, \bar{y}_t) - \left(\frac{\eta}{4} - \frac{\eta^2 L}{2}\right)\|\mathbf{v}_t'(i)\|^2 + \left(\frac{\gamma^2\eta^2 L}{2} + \gamma^2\eta^3 L^2 + \frac{3\gamma^2\eta}{2}\right)\|\bar{u}_t\|^2\right] \\
&+ \mathbb{E}\left[\frac{\eta}{2}\|\nabla_x f(\bar{x}_t, \mathbf{y}_t(i)) - \bar{u}_t\|^2 + \frac{\eta}{2}\|\nabla_y f(\bar{x}_t, \mathbf{y}_t(i)) - \mathbf{v}_t(i))\|^2 + \frac{1}{2\eta}\|\bar{y}_t - \mathbf{y}_t(i)\|^2\right],
\end{aligned}$$

where $\mathbf{v}'_t(i) = (\mathbf{y}'_{t+1}(i) - \mathbf{y}_t(i))/\eta$. Above, the equations rely on rearranging the terms; the first two inequalities are based on the L-smoothness of the objective $f(x, y)$. The third one follows from the concavity of in the direction of y; the second last inequality is a use of (17) and the Young's inequality; the last inequality follows from the Young's inequality and the L-smoothness that leads to

$$
\begin{aligned}
& (\nabla_y f(\bar{x}_t, \mathbf{y}_t(i)) - \nabla_y f(\bar{x}_{t+1}, \mathbf{y}_t(i)))^\top (\mathbf{y}'_{t+1}(i) - \mathbf{y}_t(i)) \\
\le\ & \frac{1}{4\eta} \|\mathbf{y}'_{t+1}(i) - \mathbf{y}_t(i)\|^2 + \eta \|\nabla_y f(\bar{x}_{t+1}, \mathbf{y}_t(i)) - \nabla_y f(\bar{x}_t, \mathbf{y}_t(i))\| \\
\le\ & \frac{1}{4\eta} \|\mathbf{y}'_{t+1}(i) - \mathbf{y}_t(i)\|^2 + \eta L^2 \|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
=\ & \frac{\eta}{4} \|\mathbf{v}'_t(i)\|^2 + \gamma^2 \eta^3 L^2 \|\bar{u}_t\|^2.
\end{aligned}
$$

Using the concavity of $f(x, y)$ in variable y as well as the Jensen's inequality, we have

$$
\begin{aligned}
& \mathbb{E}[-f(\bar{x}_{t+1}, \bar{y}'_{t+1})] \\
\le\ & \mathbb{E}\left[-f(\bar{x}_t, \bar{y}_t) - \left(\frac{\eta}{4} - \frac{\eta^2 L}{2} \right) \frac{1}{m} \sum_{i=1}^m \|\mathbf{v}'_t(i)\|^2 + \left(\frac{\gamma^2 \eta^2 L}{2} + \gamma^2 \eta^3 L^2 + \frac{3\gamma^2 \eta}{2} \right) \|\bar{u}_t\|^2 \right] \\
& + \frac{1}{m} \sum_{i=1}^m \mathbb{E}\left[\frac{\eta}{2} \|\nabla_x f(\bar{x}_t, \mathbf{y}_t(i)) - \bar{u}_t\|^2 + \frac{\eta}{2} \|\nabla_y f(\bar{x}_t, \mathbf{y}_t(i)) - \mathbf{v}_t(i))\|^2 + \frac{1}{2\eta} \|\bar{y}_t - \mathbf{y}_t(i)\|^2 \right].
\end{aligned}
\tag{18}
$$

Note that $\bar{y}'_{t+1} = \bar{y}_{t+1}$. Adding (15) multiplying $(1+\alpha)$ along with (18) multiplying α, we get

$$
\begin{aligned}
& \mathbb{E}[P_{t+1}] \\
\le & P_t + (1+\alpha) \mathbb{E}\left[-\frac{\gamma\eta}{2} \|\nabla P(\bar{x}_t)\|^2 - \left(\frac{\gamma\eta}{2} - \frac{L_P \gamma^2 \eta^2}{2} \right) \|\bar{u}_t\|^2 \right] + (1+\alpha) \times \\
& \mathbb{E}\left[\frac{8\kappa^2 \gamma\eta}{m} \sum_{i=1}^m \|\mathbf{v}'_t(i)\|^2 + \frac{8\kappa^2 \gamma\eta}{m} \sum_{i=1}^m \|\mathbf{v}_t(i) - \nabla_y f(\bar{x}_t, \mathbf{y}_t(i))\|^2 + \frac{\gamma\eta}{m} \sum_{i=1}^m \|\bar{u}_t - \nabla_x f(\bar{x}_t, \mathbf{y}_t(i))\|^2 \right] \\
& + \alpha \left[-\left(\frac{\eta}{4} - \frac{\eta^2 L}{2} \right) \frac{1}{m} \sum_{i=1}^m \|\mathbf{v}'_t(i)\|^2 + \left(\frac{\gamma^2 \eta^2 L}{2} + \gamma^2 \eta^3 L^2 + \frac{3\gamma^2 \eta}{2} \right) \|\bar{u}_t\|^2 \right] \\
& + \frac{\alpha}{m} \sum_{i=1}^m \mathbb{E}\left[\frac{\eta}{2} \|\nabla_x f(\bar{x}_t, \mathbf{y}_t(i)) - \bar{u}_t\|^2 + \frac{\eta}{2} \|\nabla_y f(\bar{x}_t, \mathbf{y}_t(i)) - \mathbf{v}_t(i))\|^2 + \frac{1}{2\eta} \|\bar{y}_t - \mathbf{y}_t(i)\|^2 \right].
\end{aligned}
$$

Rearranging the above result leads to

$$
\begin{aligned}
\mathbb{E}[\Psi_{t+1}] \le\ & \mathbb{E}\left[\Psi_t - \frac{(1+\alpha)\gamma\eta}{2} \|\nabla P(\bar{x}_t)\|^2 \right] \\
& - \left((1+\alpha) \left(\frac{\gamma\eta}{2} - \frac{L_P \gamma^2 \eta^2}{2} \right) - \alpha \left(\frac{L\gamma^2 \eta^2}{2} + \gamma^2 \eta^3 L + \frac{3\gamma^2 \eta}{2} \right) \right) \mathbb{E}[\|\bar{u}_t\|^2] \\
& - \left(\alpha \left(\frac{\eta}{4} - \frac{L\eta^2}{2} \right) - (1+\alpha) 8\kappa^2 \gamma\eta \right) \frac{1}{m} \sum_{i=1}^m \mathbb{E}[\|\mathbf{v}'_t(i)\|^2] \\
& + \left((1+\alpha)\gamma + \frac{\alpha}{2} \right) \frac{\eta}{m} \sum_{i=1}^m \mathbb{E}[\|\bar{u}_t - \nabla_x f(\bar{x}_t, \mathbf{y}_t(i))\|^2] \\
& + \left((1+\alpha) 8\kappa^2 \gamma + \frac{\alpha}{2} \right) \frac{\eta}{m} \sum_{i=1}^m \mathbb{E}[\|\mathbf{v}_t(i) - \nabla_y f(\bar{x}_t, \mathbf{y}_t(i))\|^2] \\
& + \frac{\alpha}{2\eta m} \sum_{i=1}^m \mathbb{E}[\|\bar{y}_t - \mathbf{y}_t(i)\|^2].
\end{aligned}
$$

Since $\alpha \in (0,1]$, taking $\eta \le 1/(4L)$ and the definition of γ in (5) mean

$$\alpha\left(\frac{\eta}{4}-\frac{L\eta^2}{2}\right)-(1+\alpha)8\kappa^2\gamma\eta \ge \frac{\alpha\eta}{16}$$

as well as

$$(1+\alpha)\left(\frac{\gamma\eta}{2}-\frac{L_P\gamma^2\eta^2}{2}\right)-\alpha\left(\frac{L\gamma^2\eta^2}{2}+\gamma^2\eta^3L+\frac{3\gamma^2\eta}{2}\right)\ge\frac{\gamma\eta}{4}.$$

Also, the fact $\alpha \in (0,1]$ and our choice of γ means

$$8(1+\alpha)\kappa^2\gamma+\frac{\alpha}{2}\le\alpha \qquad \text{and} \qquad (1+\alpha)8\kappa^2\gamma+\frac{\alpha}{2}\le\alpha.$$

Therefore, we obtain the optimization bound as

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}] \le& \mathbb{E}\left[\Psi_t-\frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t)\|^2-\frac{\gamma\eta}{4}\|\bar{u}_t\|^2-\frac{\alpha\eta}{16m}\sum_{i=1}^m\|\mathbf{v}_t'(i)\|^2\right]\\ &+\mathbb{E}\left[\frac{\alpha\eta}{m}\sum_{i=1}^m\|\bar{u}_t-\nabla_x f(\bar{x}_t,\mathbf{y}_t(i)\|^2+\frac{\alpha\eta}{m}\sum_{i=1}^m\|\mathbf{v}_t(i)-\nabla_y f(\bar{x}_t,\mathbf{y}_t(i))\|^2+\frac{\alpha}{2\eta m}\sum_{i=1}^m\|\bar{y}_t-\mathbf{y}_t(i)\|^2\right].\end{aligned} \tag{19}$$

Note that it holds that

$$\sum_{i=1}^m\|\mathbf{v}_t(i)-\nabla_y f(\bar{x}_t,\mathbf{y}_t(i))\|^2\le 3m\|\bar{v}_t-\nabla_y f(\bar{x}_t,\bar{y}_t)\|^2+3\|\mathbf{v}_t-\mathbf{1}\bar{v}_t\|^2+3L^2\|\mathbf{y}_t-\mathbf{1}\bar{y}_t\|^2$$

and

$$\sum_{i=1}^m\|\bar{u}_k-\nabla_x f(\bar{x}_t,\mathbf{y}_t(i))\|^2\le 2m\|\bar{u}_t-\nabla_x f(\bar{x}_t,\bar{y}_t)\|^2+2L^2\|\mathbf{y}_t-\mathbf{1}\bar{y}_t\|^2,$$

where we use the L-smoothness and Young's inequality. Then we combine Lemma A.5 to get

$$\begin{aligned}&\frac{\alpha\eta}{m}\sum_{i=1}^m\|\bar{u}_t-\nabla_x f(\bar{x}_t,\mathbf{y}_t(i)\|^2+\frac{\alpha\eta}{m}\sum_{i=1}^m\|\mathbf{v}_t(i)-\nabla_y f(\bar{x}_t,\mathbf{y}_t(i))\|^2\\ \le&\frac{6\alpha\eta L^2}{m}\|\mathbf{y}_t-\mathbf{1}\bar{y}_t\|^2+\frac{3\alpha\eta}{m}\|\bar{s}_t-\nabla f(\bar{x}_t,\bar{y}_t)\|^2+\frac{3\alpha\eta}{m}\|\mathbf{v}_t-\mathbf{1}\bar{v}_t\|^2\\ \le&\frac{\alpha}{2\eta m}\|\mathbf{y}_t-\mathbf{1}\bar{y}_t\|^2+3\alpha\eta\|\bar{s}_t-\nabla f(\bar{x}_t,\bar{y}_t)\|^2+\frac{3\alpha\eta}{m}\|\mathbf{v}_t-\mathbf{1}\bar{v}_t\|^2,\end{aligned} \tag{20}$$

where we use $\eta \le 1/(4L)$. Plugging (20) into (19) and then use the inequality

$$\frac{1}{m}\sum_{i=1}^m\|\mathbf{v}_t'(i)\|^2\ge\|\bar{v}_t'\|^2.$$

Hence, we have

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}]\le&\ \mathbb{E}\left[\Psi_t-\frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t)\|^2-\frac{\gamma\eta}{4}\|\bar{u}_t\|^2-\frac{\alpha\eta}{16}\|\bar{v}_t'\|^2\right]\\ &+\mathbb{E}\left[3\alpha\eta\|\bar{s}_t-\nabla_x f(\bar{x}_t,\bar{y}_t)\|^2+\frac{3\alpha\eta}{m}\|\mathbf{v}_t-\mathbf{1}\bar{v}_t\|^2+\frac{\alpha}{\eta m}\|\mathbf{y}_t-\bar{y}_t\|^2\right].\end{aligned}$$

Combining with Lemma A.5 and using $\|\mathbf{y}_t-\mathbf{1}\bar{y}_t\|\le\|\mathbf{z}_t-\mathbf{1}\bar{z}_t\|$ and $\eta\le 1/(4L)$, we obtain

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}]\le&\ \mathbb{E}\left[\Psi_t-\frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t)\|^2-\frac{\gamma\eta}{4}\|\bar{u}_t\|^2-\frac{\alpha\eta}{16}\|\bar{v}_t'\|^2\right]\\ &+\mathbb{E}\left[\frac{3\alpha\eta}{m}\|\mathbf{v}_t-\mathbf{1}\bar{v}_t\|^2+6\alpha\eta\left\|\frac{1}{m}\sum_{i=1}^m(\mathbf{g}_t(i)-\nabla f_i(\mathbf{z}_t(i)))\right\|^2+\frac{2\alpha}{\eta m}\|\mathbf{z}_t-\bar{z}_t\|^2\right].\end{aligned}$$

Note that

$$\bar{u}_t = \frac{\bar{x}_t - \bar{x}_{t+1}}{\gamma\eta}, \quad \bar{v}_t' = \frac{\bar{y}_{t+1} - \bar{y}_t}{\eta} \quad \text{and} \quad \frac{1}{8\gamma\eta} \geq \frac{\alpha}{16\eta},$$

then we have

$$\begin{aligned}
\mathbb{E}[\Psi_{t+1}] \leq \mathbb{E}\left[\Psi_t - \frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t)\|^2 - \frac{1}{8\gamma\eta}\|\bar{x}_{t+1} - \bar{x}_t\|^2 - \frac{\alpha}{16\eta}\|\bar{z}_{t+1} - \bar{z}_t\|^2\right] \\
+ \mathbb{E}\left[\frac{3\alpha\eta}{m}\|\mathbf{v}_t - \mathbf{1}\bar{v}_t\|^2 + 6\alpha\eta\left\|\frac{1}{m}\sum_{i=1}^{m}(\mathbf{g}_t(i) - \nabla f_i(\mathbf{z}_t(i)))\right\|^2 + \frac{2\alpha}{\eta m}\|\mathbf{z}_t - \mathbf{1}\bar{z}_t\|^2\right].
\end{aligned}$$

Recalling the definition of C_t and U_t, we obtain the result of Lemma 3.1. □

D The Proof of Lemma 3.2

Proof. The relation of (14) means

$$\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{1}\bar{x}_{t+1}\| \\
=&\rho\|\mathbf{x}_t - \gamma\eta\mathbf{u}_t - \mathbf{1}(\bar{x}_t - \eta\bar{u}_t)\| \\
\leq&\rho\left(\|\mathbf{x}_t - \mathbf{1}\bar{x}_t\| + \gamma\eta\|\mathbf{u}_t - \mathbf{1}\bar{u}_t\|\right),
\end{aligned}$$

where the last step is due to triangle inequality. Similarly, we define the notation $\bar{\Pi}(\cdot) = \frac{1}{m}\mathbf{1}\mathbf{1}^\top(\cdot)$ for convenience. Then for variable y, we can verify that

$$\begin{aligned}
&\|\mathbf{y}_{t+1} - \mathbf{1}\bar{y}_{t+1}\| \\
\leq&\rho\left\|\Pi(\mathbf{y}_t + \eta\mathbf{v}_t) - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\Pi(\mathbf{y}_t + \eta\mathbf{v}_t)\right\| \\
\leq&\rho\|\Pi(\mathbf{y}_t + \eta\mathbf{v}_t) - \Pi(\mathbf{1}\bar{y}_t + \eta\mathbf{1}\bar{v}_t)\| + \rho\left\|\Pi(\mathbf{1}\bar{y}_t + \eta\mathbf{1}\bar{v}_t) - \mathbf{1}\bar{\Pi}(\mathbf{y}_t + \eta\mathbf{v}_t)\right\| \\
\leq&2\rho\|\mathbf{y}_t + \eta\mathbf{v}_t - \mathbf{1}(\bar{y}_t + \eta\bar{v}_t)\| \\
\leq&2\rho\left(\|\mathbf{y}_t - \mathbf{1}\bar{y}_t\| + \eta\|\mathbf{v}_t - \mathbf{1}\bar{v}_t\|\right),
\end{aligned}$$

where in the third inequality we use the non-expansiveness of projection and Lemma 11 in Ye et al. (2020b), i.e.

$$\left\|\mathbf{1}\bar{\Pi}(\mathbf{x}) - \Pi(\mathbf{1}\bar{x})\right\| \leq \|\mathbf{x} - \mathbf{1}\bar{x}\|.$$

Consequently, we use Young's inequality together with $\gamma \in (0, 1]$ to obtain

$$\begin{aligned}
&\|\mathbf{z}_{t+1} - \mathbf{1}\bar{z}_{t+1}\|^2 \\
=& \|\mathbf{x}_{t+1} - \mathbf{1}\bar{x}_{t+1}\|^2 + \|\mathbf{y}_{t+1} - \mathbf{1}\bar{y}_{t+1}\|^2 \\
\leq& 8\rho^2\|\mathbf{y}_t - \mathbf{1}\bar{y}_t\|^2 + 8\rho^2\eta^2\|\mathbf{v}_t - \mathbf{1}\bar{v}_t\|^2 + 2\rho^2\|\mathbf{x}_t - \mathbf{1}\bar{x}_t\|^2 + 2\rho^2\eta^2\|\mathbf{y}_t - \mathbf{1}\bar{y}_t\|^2 \\
\leq& 8\rho^2\|\mathbf{z}_t - \mathbf{1}\bar{z}_t\|^2 + 8\rho^2\eta^2\|\mathbf{s}_t - \mathbf{1}\bar{s}_t\|^2.
\end{aligned} \tag{21}$$

Furthermore, if $24\rho^2 \leq 1$, we have

$$\begin{aligned}
&\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 \\
\leq& 3\|\mathbf{z}_{t+1} - \mathbf{1}\bar{z}_{t+1}\|^2 + 3\|\mathbf{z}_t - \mathbf{1}\bar{z}_t\|^2 + 3\|\mathbf{1}\bar{z}_t - \mathbf{1}\bar{z}_{t+1}\|^2 \\
\leq& (24\rho^2 + 3)\|\mathbf{z}_t - \mathbf{1}\bar{z}_k\|^2 + 24\rho^2\eta^2\|\mathbf{s}_t - \mathbf{1}\bar{s}_t\|^2 + 3\|\mathbf{1}\bar{z}_t - \mathbf{1}\bar{z}_{t+1}\|^2 \\
\leq& 4\|\mathbf{z}_t - \mathbf{1}\bar{z}_t\|^2 + \eta^2\|\mathbf{s}_t - \mathbf{1}\bar{s}_t\|^2 + 3m\|\bar{z}_t - \bar{z}_{t+1}\|^2.
\end{aligned} \tag{22}$$

We let $\rho_t = \rho'$ for $\zeta_t = 1$ and $\rho_t = \rho$ otherwise. The update of $\mathbf{g}_{t+1}(i)$ means

$$
\begin{aligned}
&\mathbb{E}\big[\rho_t^2 \left\|\mathbf{g}_{t+1}(i) - \mathbf{g}_t(i)\right\|^2\big] \\
=&\rho'^2 p\mathbb{E}\left\|\frac{1}{b'}\sum_{\xi_{i,j}\in\mathcal{S}'_t(i)} \nabla F_i(\mathbf{z}_{t+1}(i);\xi_{i,j}) - \mathbf{g}_t(i)\right\|^2 + \frac{\rho^2(1-p)}{bq}\mathbb{E}\left\|\nabla F_i(\mathbf{z}_{t+1}(i);\xi_{i,1}) - \nabla F_i(\mathbf{z}_t(i);\xi_{i,1})\right\|^2 \\
\le& 3\rho'^2 p\mathbb{E}\left\|\frac{1}{b'}\sum_{\xi_{i,j}\in\mathcal{S}'_t(i)} \nabla F_i(\mathbf{z}_{t+1}(i);\xi_{i,j}) - \nabla f_i(\mathbf{z}_{t+1}(i))\right\|^2 + 3\rho'^2 p\mathbb{E}\left\|\nabla f_i(\mathbf{z}_{t+1}(i)) - \nabla f_i(\mathbf{z}_t(i))\right\|^2 \\
&+ 3\rho'^2 p\mathbb{E}\left\|\nabla f_i(\mathbf{z}_t(i)) - \mathbf{g}_t(i)\right\|^2 + \frac{\rho^2(1-p)L^2}{bq}\mathbb{E}\left\|\mathbf{z}_{t+1}(i) - \mathbf{z}_t(i)\right\|^2 \\
\le& \frac{3\rho'^2 p\sigma^2}{b'}\mathbb{I}[b'<n] + 3\rho'^2 pL^2\mathbb{E}\left\|\mathbf{z}_{t+1}(i) - \mathbf{z}_t(i)\right\|^2 \\
&+ 3\rho'^2 p\mathbb{E}\left\|\nabla f_i(\mathbf{z}_t(i)) - \mathbf{g}_t(i)\right\|^2 + \frac{\rho^2(1-p)L^2}{bq}\mathbb{E}\left\|\mathbf{z}_{t+1}(i) - \mathbf{z}_t(i)\right\|^2 \\
\le& \frac{3\rho'^2 p\sigma^2}{b'}\mathbb{I}[b'<n] + 3\rho'^2 p\mathbb{E}\left\|\nabla f_i(\mathbf{z}_t(i)) - \mathbf{g}_t(i)\right\|^2 + \left(\frac{1-p}{bq} + 3p\right)\rho^2 L^2\mathbb{E}\left\|\mathbf{z}_{t+1}(i) - \mathbf{z}_t(i)\right\|^2,
\end{aligned}
$$

where the first inequality is based on update rules and Assumption 2.4; the second inequality is based on triangle inequality and the last inequality is due to Assumption 2.2. Summing over above result over $i = 1, \ldots, m$, obtain

$$
\mathbb{E}\big[\rho_t^2\left\|\mathbf{g}_{t+1} - \mathbf{g}_t\right\|^2\big] \le \frac{3\rho'^2 mp\sigma^2}{b'}\mathbb{I}[b'<n] + 3\rho'^2 p\mathbb{E}\left\|\nabla\mathbf{f}(\mathbf{z}_t) - \mathbf{g}_t\right\|^2 + \left(\frac{1-p}{bq} + 3p\right)\rho^2 L^2\mathbb{E}\left\|\mathbf{z}_{t+1} - \mathbf{z}_t\right\|^2.
$$

Let $c = \max\{1/(bq), 1\}$ and plug it into (22), then

$$
\begin{aligned}
&\mathbb{E}[\rho_t^2\|\mathbf{g}_{t+1} - \mathbf{g}_t\|^2] \\
\le& \frac{3\rho'^2 m\sigma^2}{b'}\mathbb{I}[b'<n] + 3\rho'^2\mathbb{E}[\|\nabla\mathbf{f}(\mathbf{z}_t) - \mathbf{g}_t\|^2] + c\rho^2 L^2\mathbb{E}[\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2] \\
\le& \frac{3\rho'^2 m\sigma^2}{b'}\mathbb{I}[b'<n] + 3\rho'^2\mathbb{E}[\|\nabla\mathbf{f}(\mathbf{z}_t) - \mathbf{g}_t\|^2] \\
&+ 16c\rho^2 L^2\mathbb{E}[\|\mathbf{z}_t - \mathbf{1}\bar{z}_t\|^2] + 4c\rho^2 L^2\eta^2\mathbb{E}[\|\mathbf{s}_t - \mathbf{1}\bar{s}_t\|^2] + 12c\rho^2 mL^2\mathbb{E}[\|\bar{z}_{t+1} - \bar{z}_t\|^2].
\end{aligned}
\tag{23}
$$

Furthermore, we have

$$
\begin{aligned}
&\|\mathbf{s}_{t+1} - \mathbf{1}\bar{s}_{t+1}\| \\
\le& \rho_t\left\|\mathbf{s}_t + \mathbf{g}_{t+1} - \mathbf{g}_t - \frac{1}{m}\mathbf{1}\mathbf{1}^\top(\mathbf{s}_t + \mathbf{g}_{t+1} - \mathbf{g}_t)\right\| \\
\le& \rho_t\left\|\mathbf{s}_t - \mathbf{1}\bar{s}_t\right\| + \rho_t\left\|\mathbf{g}_{t+1} - \mathbf{g}_t - \frac{1}{m}\mathbf{1}\mathbf{1}^\top(\mathbf{g}_{t+1} - \mathbf{g}_t)\right\| \\
\le& \rho_t\left\|\mathbf{s}_t - \mathbf{1}\bar{s}_t\right\| + \rho_t\left\|\mathbf{g}_{t+1} - \mathbf{g}_t\right\|,
\end{aligned}
\tag{24}
$$

where the second inequality is based on triangle inequality and the last step uses Lemma A.2. Combining the results of (23) and (24) and using $\eta \le 1/(4L)$, we have

$$
\begin{aligned}
&\eta^2\mathbb{E}\left\|\mathbf{s}_{t+1} - \mathbf{1}\bar{s}_{t+1}\right\|^2 \\
\le& 2\rho^2\eta^2\left\|\mathbf{s}_t - \mathbf{1}\bar{s}_t\right\|^2 + 2\eta^2\mathbb{E}[\rho_t^2\left\|\mathbf{g}_{t+1} - \mathbf{g}_t\right\|^2] \\
\le& 4c\rho^2\eta^2\mathbb{E}[\|\mathbf{s}_t - \mathbf{1}\bar{s}_t\|^2] + 4c\rho^2\mathbb{E}[\|\mathbf{z}_t - \mathbf{1}\bar{z}_t\|^2] \\
&+ 6\rho'^2\eta^2\mathbb{E}[\|\nabla\mathbf{f}(\mathbf{z}_t) - \mathbf{g}_t\|^2] + 2c\rho^2 m\mathbb{E}[\|\bar{z}_{t+1} - \bar{z}_t\|^2] + \frac{6\rho'^2 m\eta^2\sigma^2}{b'}\mathbb{I}[b'<n].
\end{aligned}
\tag{25}
$$

Combining (25) and (21), we obtain

$$
\begin{aligned}
&\mathbb{E}\big[\|\mathbf{z}_{t+1} - \mathbf{1}\bar{z}_{t+1}\|^2 + \eta^2\|\mathbf{s}_{t+1} - \mathbf{1}\bar{s}_{t+1}\|^2\big] \\
\le& 12c\rho^2\mathbb{E}[\|\mathbf{z}_t - \mathbf{1}\bar{z}_t\|^2 + \eta^2\|\mathbf{s}_t - \mathbf{1}\bar{s}_t\|^2] + 6\rho'^2\eta^2\mathbb{E}[\|\nabla\mathbf{f}(\mathbf{z}_t) - \mathbf{g}_t\|^2] + 2c\rho^2 m\mathbb{E}[\|\bar{z}_{t+1} - \bar{z}_t\|^2] + \frac{6\rho'^2 m\eta^2\sigma^2}{b'}\mathbb{I}[b'<n],
\end{aligned}
$$

which finishes our proof. □

E The Proof of Lemma 3.3

Proof. The update of $\mathbf{g}_{t+1}(i)$ means

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{g}_{t+1}(i)-\nabla f_i(\mathbf{z}_{t+1}(i))\right\|^2 \\
=&p\mathbb{E}\left\|\frac{1}{b'}\sum_{\xi_{i,j}\in\mathcal{S}'_t(i)}\nabla F_i(\mathbf{z}_{t+1}(i);\xi_{i,j})-\nabla f_i(\mathbf{z}_{t+1}(i))\right\|^2 \\
&+(1-p)\mathbb{E}\left\|\mathbf{g}_t(i)+\frac{\omega_t(i)}{bq}\sum_{\xi_{i,j}\in\mathcal{S}_t(i)}\big(\nabla F_i(\mathbf{z}_{t+1}(i);\xi_{i,j})-\nabla F_i(\mathbf{z}_t(i);\xi_{i,j})\big)-\nabla f_i(\mathbf{z}_{t+1}(i))\right\|^2 \\
\leq&\frac{p\sigma^2}{b'}\mathbb{I}[b'<n]+(1-p)\mathbb{E}\Big\|\mathbf{g}_t(i)-\nabla f_i(\mathbf{z}_t(i)) \\
&\quad+\frac{\omega_t(i)}{bq}\sum_{\xi_{i,j}\in\mathcal{S}_t(i)}\big(\nabla F_i(\mathbf{z}_{t+1}(i);\xi_{i,j})-\nabla F_i(\mathbf{z}_t(i);\xi_{i,j})-\nabla f_i(\mathbf{z}_{t+1}(i))+\nabla f_i(\mathbf{z}_t(i))\big)\Big\|^2 \\
=&\frac{p\sigma^2}{b'}\mathbb{I}[b'<n]+(1-p)\mathbb{E}\left\|\mathbf{g}_t(i)-\nabla f_i(\mathbf{z}_t(i))\right\|^2 \\
&+(1-p)\mathbb{E}\left\|\frac{\omega_t(i)}{bq}\sum_{\xi_{i,j}\in\mathcal{S}_t(i)}\big(\nabla F_i(\mathbf{z}_{t+1}(i);\xi_{i,j})-\nabla F_i(\mathbf{z}_t(i);\xi_{i,j})-\nabla f_i(\mathbf{z}_{t+1}(i))+\nabla f_i(\mathbf{z}_t(i))\big)\right\|^2 \\
\leq&\frac{p\sigma^2}{b'}\mathbb{I}[b'<n]+(1-p)\mathbb{E}\left\|\mathbf{g}_t(i)-\nabla f_i(\mathbf{z}_t(i))\right\|^2+\frac{1-p}{bq}\mathbb{E}\left\|\nabla F_i(\mathbf{z}_{t+1}(i);\xi_{i,j})-\nabla F_i(\mathbf{z}_t(i);\xi_{i,j})\right\|^2 \\
\leq&\frac{p\sigma^2}{b'}\mathbb{I}[b'<n]+(1-p)\mathbb{E}\left\|\mathbf{g}_t(i)-\nabla f_i(\mathbf{z}_t(i))\right\|^2+\frac{(1-p)L^2}{bq}\mathbb{E}\left\|\mathbf{z}_{t+1}(i)-\mathbf{z}_t(i)\right\|^2,
\end{aligned}
$$

where the first inequality is based on Assumption 2.4; the second inequality use the property of variance; the last inequality is based on Assumption 2.2; the second equality uses the property of martingale according to Proposition 1 in (Fang et al., 2018). Taking the average over $i=1,\dots,m$ for above result and using (22), we obtain

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{g}_{t+1}-\nabla\mathbf{f}(\mathbf{z}_{t+1})\right\|^2 \\
\leq&\frac{mp\sigma^2}{b'}\mathbb{I}[b'<n]+(1-p)\mathbb{E}\left\|\mathbf{g}_t-\nabla\mathbf{f}(\mathbf{z}_t)\right\|^2+\frac{(1-p)L^2}{bq}\mathbb{E}\left\|\mathbf{z}_{t+1}-\mathbf{z}_t\right\|^2 \\
\leq&\frac{mp\sigma^2}{b'}\mathbb{I}[b'<n]+(1-p)\mathbb{E}\left\|\mathbf{g}_t-\nabla\mathbf{f}(\mathbf{z}_t)\right\|^2 \\
&+\frac{(1-p)L^2}{bq}\mathbb{E}[4\|\mathbf{z}_t-\mathbf{1}\bar{z}_t\|^2+\eta^2\|\mathbf{s}_t-\mathbf{1}\bar{s}_t\|^2+3m\|\bar{z}_{t+1}-\bar{z}_t\|^2],
\end{aligned}
\tag{26}
$$

which is the variance bound as claimed by the definition of V_t. □

F The Proof of Lemma 3.4

We omit the detailed proof since it is almost identical to the proof of Lemma 3.3. We leave this as an exercise to the reader. Compared with Lemma 3.3, the quantities in Lemma 3.4 can be scaled with an additional factor of $1/m$ by using the fact

$$
\mathbb{E}\Big\|\sum_{i=1}^m a_i\Big\|^2=\sum_{i=1}^m\mathbb{E}\|a_i\|^2,
$$

where each $a_1,\dots,a_m$ are independent with zero mean.

G The Proof of Theorem 3.1

Proof. Combing Lemma 3.1, 3.2, 3.3 and 3.4 together, we obtain

$$\begin{aligned}
&\mathbb{E}[\Phi_{t+1}] \\
\leq&\mathbb{E}\left[\Phi_t - \frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t)\|^2 - \frac{1}{8\gamma\eta}\|\bar{x}_{t+1}-\bar{x}_t\|^2 - \frac{\alpha}{16\eta}\|\bar{z}_{t+1}-\bar{z}_t\|^2 + 6\alpha\eta U_t + \frac{3\alpha}{m\eta}C_t\right] \\
&+ \frac{\eta}{mp}\mathbb{E}\left[-pV_t + \frac{4(1-p)L^2}{mbq}C_t + \frac{3(1-p)L^2}{bq}\|\bar{z}_{t+1}-\bar{z}_t\|^2 + \frac{p\sigma^2}{b'}\mathbb{I}[b'<n]\right] \\
&+ \frac{\eta}{p}\mathbb{E}\left[-pU_t + \frac{4(1-p)L^2}{m^2bq}C_t + \frac{3(1-p)L^2}{mbq}\|\bar{z}_{t+1}-\bar{z}_t\|^2 + \frac{p\sigma^2}{mb'}\mathbb{I}[b'<n]\right] \\
&+ \frac{1}{\eta m}\mathbb{E}\left[-(1-12c\rho^2)C_t + 6m\rho'^2\eta^2 V_t + 2c\rho^2 m\mathbb{E}[\|\bar{z}_{t+1}-\bar{z}_t\|^2] + \frac{6\rho'^2 m\eta^2\sigma^2}{b'}\mathbb{I}[b'<n]\right] \\
=&\Phi_t + \mathbb{E}\left[-\frac{\gamma\eta}{2}\|\nabla P(\bar{x}_t)\|^2 - \frac{1}{8\gamma\eta}\|\bar{x}_{t+1}-\bar{x}_t\|^2 - \left(\frac{\alpha}{16\eta} - \frac{2c\rho^2}{\eta} - \frac{6\eta(1-p)L^2}{mbpq}\right)\|\bar{z}_{t+1}-\bar{z}_t\|^2\right] \\
&- (1-6\alpha)\eta U_t - \frac{(1-6m\rho'^2)\eta}{m}V_t - \left(\frac{1-12c\rho^2-3\alpha}{\eta m} - \frac{8(1-p)L^2}{m^2bpq}\right)C_t \\
&+ \left(6\rho'^2\frac{2}{m}\right)\frac{\eta\sigma^2}{b'}\mathbb{I}[b'<n].
\end{aligned}$$

Plugging in our setting of parameters in (7), it can be seen that

$$\mathbb{E}[\Phi_{t+1}] \leq \Phi_t - \frac{\gamma\eta}{2}\mathbb{E}[\|\nabla P(\bar{x}_t)\|^2] - \frac{4\alpha}{\eta m}C_t + \frac{3\eta\sigma^2}{mb'}\mathbb{I}[b'<n].$$

Telescoping for $t = 0, 1\cdots, T-1$, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla P(\bar{x}_t)\|^2] \leq \frac{2}{\gamma\eta T}\Phi_0 - \frac{8\alpha}{\gamma\eta^2 mT}\sum_{t=0}^{T-1}C_t + \frac{6\sigma^2}{\gamma mb'}\mathbb{I}[b'<n]. \tag{27}$$

Note that achieving $\bar{x}_t$ is not simple, so the output x_{out} is sampled from $\{\mathbf{x}_t(i)\}$ where $t = 0, \ldots, T-1$ and $i = 1, \ldots, m$. We also has the following bound:

$$\begin{aligned}
\mathbb{E}\left\|\nabla P(x_{\text{out}})\right\|^2 =& \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=0}^{T-1}\left\|\nabla P(\mathbf{x}_t(i))\right\|^2 \\
\leq& \frac{2}{mT}\sum_{i=1}^{m}\sum_{t=0}^{T-1}\left(\left\|\nabla P(\bar{x}_t)\right\|^2 + \left\|\nabla P(\mathbf{x}_t(i)) - \nabla P(\bar{x}_t)\right\|^2\right) \\
\leq& \frac{2}{mT}\sum_{i=1}^{m}\sum_{t=0}^{T-1}\left(\left\|\nabla P(\bar{x}_t)\right\|^2 + L^2\left\|\mathbf{x}_t(i) - \bar{x}_t\right\|^2\right) \\
=& \frac{2}{T}\sum_{t=0}^{T-1}\left\|\nabla P(\bar{x}_t)\right\|^2 + \frac{2L^2}{mT}\sum_{t=0}^{T-1}\left\|\mathbf{x}_t - \mathbf{1}\bar{x}_t\right\|^2 \\
\leq& \frac{2}{T}\sum_{t=0}^{T-1}\left\|\nabla P(\bar{x}_t)\right\|^2 + \frac{2L^2}{mT}\sum_{t=0}^{T-1}C_t,
\end{aligned}$$

where the first step use Young's inequality; the second inequality is due to Assumption 2.2. Now we plug in (27) to get the following bound as

$$\mathbb{E}\|\nabla P(x_{\text{out}})\|^2 \leq \frac{4}{\gamma\eta T}\Phi_0 - \left(\frac{16\alpha}{\gamma\eta^2} - 2L^2\right)\frac{1}{mT}\sum_{t=0}^{T-1}C_t + \frac{12\sigma^2}{\gamma mb'}\mathbb{I}[b'<n].$$

Recall the choice of γ in (5); $\eta \leq 1/(4L)$ and $pT \geq 2$. Hence, we have

$$\begin{aligned}
\mathbb{E}\|\nabla P(x_{\mathrm{out}})\|^2 &\leq \frac{4}{\gamma\eta T}\Phi_0 + \frac{12\sigma^2}{\gamma m b'}\mathbb{I}[b' < n] \\
&= \frac{4}{\gamma\eta T}\left(\Psi_0 + \frac{\eta}{mp}V_0 + \frac{\eta}{p}U_0 + \frac{\eta}{m}C_0\right) + \frac{12\sigma^2}{\gamma m b'}\mathbb{I}[b' < n] \\
&= \frac{4}{\gamma\eta T}\left(\Psi_0 + \frac{2\eta\sigma^2}{mb'p}\mathbb{I}[b' < n] + \frac{\eta}{m}\|\mathbf{s}_0 - \mathbf{1}\bar{s}_0\|^2\right) + \frac{12\sigma^2}{\gamma m b'}\mathbb{I}[b' < n] \\
&\leq \frac{4}{\gamma\eta T}\left(\Psi_0 + \frac{2\eta\sigma^2}{mb'p}\mathbb{I}[b' < n] + \frac{\eta\rho_0^2}{m}\|\mathbf{g}_0 - \mathbf{1}\bar{g}_0\|^2\right) + \frac{12\sigma^2}{\gamma m b'}\mathbb{I}[b' < n] \\
&\leq \frac{8}{\gamma\eta T}\Psi_0 + \frac{16\sigma^2}{\gamma m b'}\mathbb{I}[b' < n] + \frac{4\rho_0^2\|\mathbf{g}_0 - \mathbf{1}\bar{g}_0\|^2}{\gamma m}.
\end{aligned}$$

Therefore the parameters in (7) and (8) guarantee that $\mathbb{E}\|\nabla P(x_{\mathrm{out}})\|^2 \leq \epsilon^2$. The Jensen's inequality further implies that the output is a nearly stationary point satisfying $\mathbb{E}\|\nabla P(x_{\mathrm{out}})\| \leq \epsilon$.

Recall our choice of γ in (5), we know that $1/\gamma = \Theta(\kappa^2)$. Then the total SFO complexity for all agents in expectation is

$$mb' + mT(b'p + bq(1-p)) = mb' + \frac{2mTb'bq}{b' + bq} \leq mb' + 2mTbq.$$

Plug in the choice of b, b', q yields the SFO complexity as claimed. Next, recalling the definition of ρ, ρ', ρ_0 in (6), we know the total number of communication rounds is

$$K_0 + T(pK' + (1-p)K) = \begin{cases} \mathcal{O}\left(\kappa^2 L\epsilon^{-2}/\sqrt{\delta}\right), & b' \geq m; \\ \mathcal{O}\left(\kappa^2 L\epsilon^{-2}\log(m/b')/\sqrt{\delta}\right), & b' < m. \end{cases}$$

□

H Future Directions and Subsequent Works

A future direction is to establish decentralized stochastic algorithms with better dependency on the condition number κ by devising multiple-looped algorithms such as the single-machine setting (Zhang et al., 2022). It is also interesting to consider decentralized minimax optimization with the nonconvex-non-strongly-concave objectives (Lin et al., 2020a,b; Xu et al., 2023; Zhang et al., 2020), or some classes of nonconvex-nonconcave objectives (Chen et al., 2022; Diakonikolas et al., 2021; Guo et al., 2023; Jin et al., 2020; Li et al., 2022c; Yang et al., 2020; Zheng et al., 2022).

After we posted our work on arXiv, we also noticed that some subsequent works studying similar problem setups (Huang and Chen, 2023; Mancino-Ball and Xu, 2023; Xu, 2023). Some of the results in these works can apply to more general setups than our problem, but the convergence rates in these works do not surpass the results of our paper. (a) Although Xu (2023) considers more general regularizers (we only consider the case that regularizer in y is the indicator function of set $\mathcal{Y}$), Xu (2023) only studies the offline case and does not study stochastic optimization. Following our notations, the method proposed by Xu (2023) requires the computation complexity of $\mathcal{O}(N\kappa^{2.5}\epsilon^{-2})$, which is worse than our result of $\mathcal{O}(N + \sqrt{N}\kappa^2\epsilon^{-2})$. (b) Compared with Xu (2023), Mancino-Ball and Xu (2023) additionally considers the stochastic optimization, but only for the offline case. It requires the computation complexity of $\mathcal{O}(N + \sqrt{mN}\kappa^2\epsilon^{-2})$, which is still worse than our $\mathcal{O}(N + \sqrt{N}\kappa^2\epsilon^{-2})$. (c) Huang and Chen (2023) considers the unconstrained online case under Polyak–Łojasiewicz (PL) condition. In fact, all of the analyses in our paper still hold if we relax the strong concavity assumption to the PL condition in the unconstrained case. Furthermore, although Huang and Chen (2023)'s computation complexity $\mathcal{O}(\kappa^3\epsilon^{-3})$ matches our result, their communication complexity $\mathcal{O}(\kappa^3\epsilon^{-3})$ is worse than our $\tilde{\mathcal{O}}(\kappa^2\epsilon^{-2})$.